**UNIVERSITY OF SÃO PAULO
SÃO CARLOS SCHOOL OF ENGINEERING**

Nádia Junqueira Martarelli Fróes

**Contributions to the mixed data clustering problem:
From a conceptual codification and classification proposal
to the usage of optimization methods**

São Carlos

2021

**UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS**

Nádia Junqueira Martarelli Fróes

**Contribuições para o problema de agrupamento de dados
mistos: de uma proposta conceitual de codificação e
classificação ao uso de métodos de otimização**

São Carlos

2021

Nádia Junqueira Martarelli Fróes

**Contributions to the mixed data clustering problem:
From a conceptual codification and classification proposal
to the usage of optimization methods**

Corrected Version

Doctoral thesis presented to the Graduate Program in Production Engineering from the São Carlos School of Engineering, University of São Paulo, to obtain the degree of Doctor in Science

Área de concentração: Process and Operations Management

Advisor: Prof. Dr. Marcelo Seido Nagano

São Carlos

2021

I AUTHORIZE TOTAL OR PARTIAL REPRODUCTION OF THIS WORK BY ANY CONVENTIONAL OR ELECTRONIC MEANS, FOR RESEARCH PURPOSES, SO LONG AS THE SOURCE IS CITED.

Index card prepared by User Service at "Prof. Dr. Sergio Rodrigues Fontes Library" at EESC/USP

F925c Fróes, Nádia Junqueira Martarelli
Contributions to the mixed data clustering problem:
from a conceptual codification and classification
proposal to the usage of optimization methods / Nádia
Junqueira Martarelli Fróes; advisor Marcelo Seido Nagano.
-- São Carlos, 2021.

Doctoral (Dissertation) - Graduate Program in
Production Engineering and Concentration area in Process
and Operations Management. -- São Carlos School of
Engineering of University of São Paulo, 2021.

1. Mixed data clustering. 2. Standard representation.
3. Systematic literature review. 4. Feature balancing. 5.
Feature weighting. I. Title.

FOLHA DE JULGAMENTO

Candidata: Engenheira **NÁDIA JUNQUEIRA MARTARELLI FRÓES**.

Título da tese: "Contribuições para o problema de agrupamento de dados mistos: de uma proposta conceitual de codificação e classificação ao uso de métodos de otimização".

Data da defesa: 16/07/2021

Comissão Julgadora

Resultado

Prof. Dr. **Marcelo Seido Nagano** (Orientador)

Aprovada

(Escola de Engenharia de São Carlos/EESC-USP)

Prof. Associado **Mário de Castro Andrade Filho**

Aprovada

(Instituto de Ciências Matemáticas e de Computação/ICMC-USP)

Prof. Dr. **Roberto Fernandes Tavares Neto**

Aprovada

(Universidade Federal de São Carlos/UFSCar)

Prof. Associado **Evandro Eduardo Seron Ruiz**

Aprovada

(Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto-FFLCRP/USP)

Prof. Dr. **Bruno de Athayde Prata**

Aprovada

(Universidade Federal do Ceará/UFC)

Coordenadora do Programa de Pós-Graduação em Engenharia de Produção:

Profa. Dra. **Janaina Mascarenhas Hornos da Costa**

Presidente da Comissão de Pós-Graduação:

Prof. Titular **Murilo Araujo Romero**

This thesis is dedicated to my parents, Andrea and Antonio, my sister Thais, my brother Rafael, my husband Thiago. Each line of this document shows the support I received from all of you and is a way of expressing the love I feel for each one of you.

ACKNOWLEDGEMENTS

I would like to thank God for the many opportunities I had that allowed me not only to do this thesis but living it intensively as I thought it should be. Many of the achievements I reach throughout these years are not written or shown in this document because they are intangible, however, they were just as important as the tangible results I materialized so far. I thank God for this and because I could feel his presence in the difficult and glorious moments of this journey.

In addition, I would thank my parents, sister, brother, and my husband for all the support and patience they have given me not only in these last years but throughout my life. Thank you for understanding the time it took me on many weekends, nights, holidays to read, write, and program things that are and are not in this document, but that were part of the path that brought me here. You are all my foundations. I could not fail to thank my dog Snow, who joined us at the beginning of the pandemic, in the last moments of development of this thesis. His presence allowed me to live moments of lightness and joy that only a canine buddy can provide.

I would also express my sincere gratitude to my advisor, Professor Marcelo Seido Nagano for the support, professionalism, and friendship over the years we have worked together. His patience, motivation, wisdom, and guidance make my Ph.D. study possible, pleasant, and purposeful.

My thanks are also to the friends of the Laboratory of Applied Operations Research (LAOR), who have always been present, and to the employees of the Production Engineering Department, who have been with me since my undergraduate study.

Finally, I would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), which enabled the development of this thesis through financial support.

*“Todo jardim começa com um sonho de amor.
Antes que qualquer árvore seja plantada
ou qualquer lago seja construído,
é preciso que as árvores e os lagos
tenham nascido dentro da alma.
Quem não tem jardins por dentro,
não planta jardins por fora
e nem passeia por eles...”*
Rubem Alves

ABSTRACT

FRÓES, N. J. M. **Contributions to the mixed data clustering problem: From a conceptual codification and classification proposal to the usage of optimization methods.** 2021. 282p. Thesis (Dissertation) - São Carlos School of Engineering, University of São Paulo, São Carlos, 2021.

Data mining techniques have gained prominence in recent years due to their wide range of applications. Among such techniques is data clustering, which identifies groups in unlabeled datasets. In view of the need to describe the characteristics of a phenomenon in more detail through numerical and categorical attributes, the development of the data clustering technique started to include the study of the mixed datasets in the process of clustering. Although promising, this new branch of study is still recent in the literature. Thus, this thesis aims to contribute to the advancement of the mixed data clustering problem, through four objectives, which are: to propose a standard representation for documents published in the knowledge-discovery field, to carry out a systematic review of the literature that provides a comprehensive view of this thematic and also a detailed comprehension of the selected documents; to perform the modeling of the meta-heuristic Biased Random-Key Genetic Algorithm (BRKGA) to propose a solution to the feature balancing problem in distance-based mixed data clustering algorithms; and to perform the modeling and hybridization of the following meta-heuristics: Evolutionary Clustering Search, Iterated Local Search, and BRKGA, to propose a solution to the feature weighting problem in a model-based mixed data clustering algorithm. As a result, one proposed for an initial idea for a standard representation, and a systematic review of the literature was obtained with a bibliometric and individual analysis of 160 documents resulted from the selection step of the designed methodological procedure. In addition, the proposed meta-heuristics obtained the best performances in most of the 476 simulated datasets, which contemplated several characteristics, such as data generated through the normal and lognormal distribution, balancing and unbalancing numerical and categorical attributes in relation to the amount of each type of feature in the dataset, different levels of overlapping of attributes, among others. Therefore, one concludes that this thesis reached its objectives, contributing to the advancement of the mixed data clustering technique.

Keywords: Mixed Data Clustering. Standard Representation. Systematic Literature Review. Feature Balancing. Feature Weighting.

RESUMO

FRÓES, N. J. M. **Contribuições para o problema de agrupamento de dados mistos: De uma proposta conceitual de codificação e classificação ao uso de métodos de otimização.** 2021. 282p. Thesis (Dissertation) - São Carlos School of Engineering, University of São Paulo, São Carlos, 2021.

As técnicas de mineração de dados têm ganhado destaque nos últimos anos devido sua ampla gama de aplicações. Dentre tais técnicas, está o agrupamento de dados, o qual identifica grupos em conjuntos de dados não rotulados. Diante da necessidade de descrever as características do fenômeno estudado com mais detalhes por meio do uso de atributos numéricos e categóricos, o desenvolvimento da técnica de agrupamento de dados passou a contemplar o estudo do agrupamento de dados mistos. Embora promissor, este novo ramo de estudo ainda é recente na literatura. Dessa forma, esta tese visa contribuir com o avanço da técnica de agrupamento de dados mistos, por meio de quatro objetivos, que são: propor uma representação padrão para documentos publicados no campo de descoberta do conhecimento, realizar uma revisão sistemática da literatura que forneça uma visão abrangente desta temática e também detalhada dos documentos selecionados; realizar a modelagem da meta-heurística Biased Random-Key Genetic Algorithm (BRKGA) para propor uma solução para o problema do balanceamento de atributos em algoritmos de agrupamento de dados mistos baseados em distância; e realizar a modelagem e hibridização das meta-heurísticas: Evolutionary Clustering Search, Iterated Local Search e BRKGA, para propor uma solução para o problema de ponderação de atributos em um algoritmo de agrupamento de dados mistos baseado em modelo. Como resultados, obteve-se uma proposta de ideia inicial para a representação padrão e uma revisão sistemática da literatura com uma análise bibliométrica e individual de 160 documentos resultantes da seleção feita para o estudo. Além disso, as meta-heurísticas propostas obtiveram os melhores desempenhos na maioria dos 476 conjuntos de dados simulados, os quais contemplaram diversas características, como dados gerados por meio de distribuição normal e lognormal, balanceamento e desbalanceamento de atributos numéricos e categóricos com relação a quantidade de cada tipo de atributo no conjunto de dados, sobreposição de atributos em diferentes níveis, entre outras. Portanto, conclui-se que esta tese alcançou seus objetivos, contribuindo para o avanço da técnica de agrupamento de dados mistos.

Palavras-chave: Agrupamento de Dados Mistos. Representação Padrão. Revisão Sistemática da Literatura. Balanceamento de Atributos. Ponderação de Atributos.

LIST OF FIGURES

Figure 1 – The KDD process (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a)	36
Figure 2 – The KDD process with the three main steps (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a)	54
Figure 3 – The standard representation proposal, the five-fields notation	55
Figure 4 – Branches of mixed data clustering	67
Figure 5 – Some examples of what is addressed in the branches <i>others</i> and IAPRA	67
Figure 6 – Steps of the protocol to conduct the systematic literature review	69
Figure 7 – Main deliveries of the systematic review: bibliometric and individual analysis	70
Figure 8 – Document collection - Annual scientific production	75
Figure 9 – Sources analysis - Most relevant sources	76
Figure 10 – Sources analysis - Most cited sources	77
Figure 11 – Sources analysis - Source growth	77
Figure 12 – Authors analysis - Most relevant authors	78
Figure 13 – Authors analysis - H-index of the authors	79
Figure 14 – Authors analysis - Production of the authors over the time	79
Figure 15 – Authors analysis - Productivity of the authors through Lotka's law	80
Figure 16 – Authors analysis - Most cited countries	80
Figure 17 – Authors analysis - Corresponding author's country, where red means multiple countries publications (MCP) and blue is regard to single country publication (SCP)	81
Figure 18 – Authors analysis - Scientific production by country	81
Figure 19 – Documents analysis - Most global cited documents	82
Figure 20 – Documents analysis - Most local cited documents	83
Figure 21 – Documents analysis - Most cited references	83
Figure 22 – Documents analysis - Reference publication year spectroscopy, where the red line means the deviation from the 5-year median and the blue line the number of cited reference per year	84
Figure 23 – Documents analysis - Most frequent author's keywords	84
Figure 24 – Documents analysis - Most frequent keywords plus - Word cloud	85
Figure 25 – Conceptual analysis - Co-occurrences network of the authors' keywords	86
Figure 26 – Conceptual analysis - Thematic map regarding the authors' keywords	86
Figure 27 – Intellectual analysis - Co-citation network	87
Figure 28 – Intellectual analysis - Historiography mapping	87
Figure 29 – Social analysis - Co-authorship network	89
Figure 30 – Social analysis - Co-country network	90

Figure 31 – Individual analysis, Part 2 - Most addressed datasets	115
Figure 32 – Individual analysis, Part 2 - Most addressed preprocessing tasks	115
Figure 33 – Individual analysis, Part 2 - Most addressed processing tasks	116
Figure 34 – Individual analysis, Part 2 - Most addressed post-processing tasks	116
Figure 35 – Individual analysis, Part 2 - Auxiliary approaches	117
Figure 36 – Individual analysis, Part 3 - Simulated datasets per branch	121
Figure 37 – Individual analysis, Part 3 - Real-world datasets per branch	122
Figure 38 – Individual analysis, Part 3 - Most used real-world datasets	122
Figure 39 – Individual analysis, Part 3 - Internal performance measures	123
Figure 40 – Individual analysis, Part 3 - External performance measures	123
Figure 41 – Individual analysis, Part 3 - Relative performance measures	124
Figure 42 – Individual analysis, Part 3 - Programming languages per branch	124
Figure 43 – Individual analysis, Part 3 - Availability of the code	125
Figure 44 – Individual analysis, Part 3 - Most used programming languages	125
Figure 45 – Decoding process of an individual	141
Figure 46 – Execution time in different numbers of instances	152
Figure 47 – Execution time in different numbers of attributes	153
Figure 48 – Experiment A - Normalized average values of the objective function of the Kamila algorithm for all dataset	167
Figure 49 – Experiment B - Normalized average values of the objective function of the Kamila algorithm for all dataset	168
Figure 50 – Group 1	273
Figure 51 – Group 2	274
Figure 52 – Group 3	275
Figure 53 – Group 4	276
Figure 54 – Execution time in different numbers of instances	277
Figure 55 – Execution time in different numbers of attributes	278

LIST OF TABLES

Table 1 – Attribute-value format of a dataset D with N observations e M attributes	37
Table 2 – A toy example of a mixed dataset - Raw data	37
Table 3 – A toy example of a mixed dataset - Closer to a target data	38
Table 4 – A toy example of a mixed dataset - Transformed data (discretization method)	40
Table 5 – A toy example of a mixed dataset - Transformed data (Dummy coding method)	41
Table 6 – Acronyms for characteristics of the datasets	56
Table 7 – Acronyms for the activities from the preprocessing step	57
Table 8 – Acronyms for the activities from the post-processing step	57
Table 9 – Acronyms for the auxiliary approaches	58
Table 10 – Acronyms for the activities from the processing step (data clustering) .	59
Table 11 – Standard representation for some real documents in data clustering theme	60
Table 12 – Questions - Bibliometric analysis	70
Table 13 – Questions - Individual analysis, Part 1 - Basic description	71
Table 14 – Questions - Individual analysis, Part 2 - Standard representation	71
Table 15 – Questions - Individual analysis, Part 3 - Dataset, validation measures and programming language	72
Table 16 – Number of documents throughout the selection step	73
Table 17 – Number of documents in each branch	74
Table 18 – Co-occurrence of documents in the branches	74
Table 19 – Document collection - Description	75
Table 20 – Sources analysis - Sources in zone 1 of the Bradford's law	78
Table 21 – Authors analysis - Scientific production by country	82
Table 22 – Details of the experiments	144
Table 23 – Details of the experiments	145
Table 24 – Groups of methods for comparisons	145
Table 25 – Results of experiment A1 - Normal distribution, 250 instances and two levels in each categorical attribute	147
Table 26 – Results of experiment A2 - LogNormal distribution, 250 instances and two levels in each categorical attribute	148
Table 27 – Results of experiment B1 - Normal distribution, 250 instances and two levels in each categorical attribute	149
Table 28 – Results of experiment B2 - LogNormal distribution, 250 instances and two levels in each categorical attribute	149

Table 29 – Results of experiment C1 - Normal distribution, 250 instances and ten levels in each categorical attribute	150
Table 30 – Results of experiment C2 - LogNormal distribution, 250 instances and ten levels in each categorical attribute	151
Table 31 – Results of experiment D1 - Normal distribution, two levels in each categorical attribute	151
Table 32 – Results of experiment D2 - LogNormal distribution, two levels in each categorical attribute	152
Table 33 – Details of the experiments	162
Table 34 – Results of experiment A - Average and standard deviation of the value of the objective function of the Kamila algorithm (the greater the better). Bold values indicate the best method with respect to performance (solution quality).	165
Table 35 – Results of experiment B - Average and standard deviation of the value of the objective function of the Kamila algorithm (the greater the better). Bold values indicate the best method with respect to performance (solution quality).	166
Table 36 – Results of experiment C - Average and standard deviation of the value of the objective function of the Kamila algorithm (the greater the better). Bold values indicate the best method with respect to performance (solution quality).	167
Table 37 – Individual analysis, Part 1 - Partitional branch	192
Table 38 – Individual analysis, Part 1 - Hierarchical branch	194
Table 39 – Individual analysis, Part 1 - Model-based branch	194
Table 40 – Individual analysis, Part 1 - Neural network-based branch	195
Table 41 – Individual analysis, Part 1 - Others branch	196
Table 42 – Individual analysis, Part 1 - IAPRA branch	198
Table 43 – Individual analysis, Part 2 - Partitional branch	206
Table 44 – Individual analysis, Part 2 - Hierarchical branch	208
Table 45 – Individual analysis, Part 2 - Model-based branch	208
Table 46 – Individual analysis, Part 2 - Neural network-based branch	209
Table 47 – Individual analysis, Part 2 - Others branch	210
Table 48 – Individual analysis, Part 2 - IAPRA branch	211
Table 49 – Acronyms - Datasets' characteristics and preprocessing tasks	213
Table 50 – Acronyms - Processing tasks	214
Table 51 – Acronyms - Post-processing tasks and auxiliary approaches	215
Table 52 – Individual analysis, Part 3 - Partitional branch, Datasets	218
Table 53 – Individual analysis, Part 3 - Partitional branch, Performance measures	220

Table 54 – Individual analysis, Part 3 - Partitional branch, Programming languages and code availability	223
Table 55 – Individual analysis, Part 3 - Hierarchical branch, Datasets	226
Table 56 – Individual analysis, Part 3 - Hierarchical branch, Performance measures	227
Table 57 – Individual analysis, Part 3 - Hierarchical branch, Programming languages and code availability	227
Table 58 – Individual analysis, Part 3 - Model-based branch, Datasets	228
Table 59 – Individual analysis, Part 3 - Model-based branch, Performance measures	229
Table 60 – Individual analysis, Part 3 - Model-based branch, Programming languages and code availability	231
Table 61 – Individual analysis, Part 3 - Neural network-based branch, Datasets . .	232
Table 62 – Individual analysis, Part 3 - Neural network-based branch, Performance measures	233
Table 63 – Individual analysis, Part 3 - Neural network-based branch, Programming languages and code availability	234
Table 64 – Individual analysis, Part 3 - Others branch, Datasets	234
Table 65 – Others - Performance measures	236
Table 66 – Others - Programming language and code	237
Table 67 – Individual analysis, Part 3 - IAPRA branch, Datasets	238
Table 68 – IAPRA - Performance measures	239
Table 69 – IAPRA - Programming language and code	240
Table 70 – Description of the datasets - Acronyms, address and composition	243
Table 71 – Acronyms for validation measures	248
Table 72 – Parameters' values defined by the irace method	252
Table 73 – Results of experiment A1 - ARPD values considering the ARI values of each repetition	253
Table 74 – Results of experiment A1 - Average and standard deviation of numerical attributes weights	254
Table 75 – Results of experiment A1 - Average and standard deviation of categorical attributes weights	255
Table 76 – Results of experiment A1 - Average and standard deviation of execution time	256
Table 77 – Results of experiment A2 - ARPD values considering the ARI values of each repetition	257
Table 78 – Results of experiment A2 - Average and standard deviation of numerical attributes weights	258
Table 79 – Results of experiment A2 - Average and standard deviation of categorical attributes weights	259

Table 80 – Results of experiment A2 - Average and standard deviation of execution time	260
Table 81 – Results of experiment B1 - ARPD values considering the ARI values of each repetition	260
Table 82 – Results of experiment B1 - Average and standard deviation of numerical attributes weights	261
Table 83 – Results of experiment B1 - Average and standard deviation of categorical attributes weights	261
Table 84 – Results of experiment B1 - Average and standard deviation of execution time	262
Table 85 – Results of experiment B2 - ARPD values considering the ARI values of each repetition	262
Table 86 – Results of experiment B2 - Average and standard deviation of numerical attributes weights	263
Table 87 – Results of experiment B2 - Average and standard deviation of categorical attributes weights	263
Table 88 – Results of experiment B2 - Average and standard deviation of execution time	264
Table 89 – Results of experiment C1 - ARPD values considering the ARI values of each repetition	264
Table 90 – Results of experiment C1 - Average and standard deviation of numerical attributes weights	265
Table 91 – Results of experiment C1 - Average and standard deviation of categorical attributes weights	266
Table 92 – Results of experiment C1 - Average and standard deviation of execution time	267
Table 93 – Results of experiment C2 - ARPD values considering the ARI values of each repetition	267
Table 94 – Results of experiment C2 - Average and standard deviation of numerical attributes weights	268
Table 95 – Results of experiment C2 - Average and standard deviation of categorical attributes weights	269
Table 96 – Results of experiment C2 - Average and standard deviation of execution time	270
Table 97 – Results of experiment D1 - Average and standard deviation of ARI values	270
Table 98 – Results of experiment D1 - ARPD values considering the ARI values of each repetition	270
Table 99 – Results of experiment D1 - Average and standard deviation of numerical attributes weights	271

Table 100–Results of experiment D1 - Average and standard deviation of categorical attributes weights	271
Table 101–Results of experiment D1 - Average and standard deviation of execution time	271
Table 102–Results of experiment D2 - Average and standard deviation of ARI values	271
Table 103–Results of experiment D2 - ARPD values considering the ARI values of each repetition	271
Table 104–Results of experiment D2 - Average and standard deviation of numerical attributes weights	271
Table 105–Results of experiment D2 - Average and standard deviation of categorical attributes weights	272
Table 106–Results of experiment D2 - Average and standard deviation of execution time	272
Table 107–Baseline methods - Parameters' values defined by irace method	279
Table 108–Proposed methods, Part 1 - Parameters' values defined by irace method	279
Table 109–Proposed methods, Part 2 - Parameters' values defined by irace method	279
Table 110–Results of experiment A - Average and standard deviation of the execution time.	280
Table 111–Results of experiment B - Average and standard deviation of the execution time.	281
Table 112–Results of experiment C - Average and standard deviation of the execution time.	282

CONTENTS

I	WHERE ARE WE? AND WHERE DO WE GO TO?	29
1	INTRODUCTION	31
1.1	Objectives and outline of the thesis	31
2	MAIN CONCEPTS	35
2.1	KDD process	35
2.2	Attribute-value format	36
2.3	Mixed dataset	37
2.3.1	What is not a mixed data in this thesis?	38
2.4	Strategies to cope with mixed data	39
2.4.1	Transformation methods	39
2.4.2	Compound assignment functions	41
2.4.2.1	Dissimilarity functions	41
2.4.2.2	Statistical models	44
2.5	Data clustering technique	45
2.5.1	Classification	45
2.5.2	Performance measures	46
II	ORGANIZING TO GO BEYOND	49
3	STANDARD REPRESENTATION: FIVE-FIELDS NOTATION . . .	51
3.1	Literature review	52
3.2	The standard representation	53
3.2.1	DTS - Datasets	56
3.2.2	PRE - Preprocessing tasks	56
3.2.3	PRO - Processing tasks	57
3.2.4	POS - Post-processing tasks	57
3.2.5	AUX - Auxiliary approaches	58
3.3	An example for data clustering technique	58
3.4	Discussion	60
3.5	Final remarks	61
III	MANY VIEWS OF THE PAST	63
4	SYSTEMATIC LITERATURE REVIEW	65

4.1	Branches of data clustering algorithms and adjacent approaches . .	67
4.2	Methodology	68
4.2.1	Steps 1 and 2	69
4.2.2	Step 4	69
4.2.3	Step 3	69
4.2.4	Steps 5 and 6	72
4.2.5	Steps 7, 8 and 9	72
4.3	Search results	73
4.3.1	Document collection	75
4.4	Bibliometric analysis	76
4.4.1	Sources	76
4.4.2	Authors	77
4.4.3	Document	81
4.4.4	Conceptual structure	84
4.4.5	Intellectual structure	85
4.4.6	Social structure	88
4.4.7	Answers of the questions defined for the bibliometric analysis	88
4.5	Individual analysis - Part 1	93
4.5.1	Answers of the questions defined for the individual analysis - Part 1	93
4.6	Individual analysis - Part 2	95
4.6.1	Partitional	95
4.6.1.1	Discussion	101
4.6.2	Hierarchical	101
4.6.2.1	Discussion	103
4.6.3	Model-based	103
4.6.3.1	Discussion	106
4.6.4	Neural network-based	106
4.6.4.1	Discussion	108
4.6.5	Others	108
4.6.5.1	Discussion	111
4.6.6	IAPRA	112
4.6.6.1	Discussion	113
4.6.7	Answers of the questions defined for the individual analysis - Part 2	114
4.7	Individual analysis - Part 3	116
4.7.1	Answers of the questions defined for the individual analysis - Part 3	120
4.8	Challenges and paths for future research	121
4.8.1	Bibliometric analysis and individual analysis, part 1	122
4.8.2	Individual analysis, part 2	126
4.8.3	Individual analysis, part 3	127

4.9	Final remarks	128
IV	PATTERNS BEHIND OPTIMIZATION	131
5	FEATURE BALANCING	133
5.1	Motivation	133
5.2	Main concepts	134
5.2.1	Differences among balancing, weighting and selection	134
5.2.2	Feature balancing relies on the numerical data normalization method	135
5.2.3	Why does one improve distance-based algorithms?	136
5.3	Related works	137
5.4	Proposed optimizer	139
5.4.0.1	Initial population, coding, decoding scheme, and fitness function	140
5.4.0.2	Selection, crossover, mutation, and stopping rule	141
5.5	Proposed algorithms	142
5.6	Methodology	143
5.6.1	Experiments	143
5.6.2	Methods	144
5.6.3	Parameters definition	144
5.6.4	Performance measure and statistical test	145
5.6.5	Computational setting	146
5.7	Results	146
5.8	Discussion	146
5.8.1	Experiment A	146
5.8.2	Experiment B	153
5.8.3	Experiment C	154
5.8.4	Experiment D	155
5.9	Final remarks	156
6	FEATURE WEIGHTING	157
6.1	Motivation	157
6.2	Related works	157
6.3	Proposed algorithms	159
6.3.1	Evolutionary Clustering Search	159
6.3.2	Iterated Local Search	160
6.3.3	Hybridized algorithm	161
6.4	Methodology	161
6.4.1	Experiments	162
6.4.2	Methods	163

6.4.3	Parameters definition	163
6.4.4	Performance measure and statistical test	164
6.4.5	Computational setting	164
6.5	Results and discussion	164
6.6	Final remarks	166
V	ACHIEVEMENTS AND PATHS TO ADVANCE	169
7	CONCLUSION, RESULTS AND FUTURE RESEARCH LINES . . .	171
7.1	Conclusion	171
7.2	Results	173
7.3	Future research lines	173
7.4	Acknowledgment	174
	REFERENCES	175
A	APPENDIX A	191
B	APPENDIX B	205
C	APPENDIX C	217
C.1	Meaning of the acronyms	242
D	APPENDIX D	251
D.1	Parameters	251
D.2	Results - Tables	251
D.2.1	Experiment A	251
D.2.1.1	A1 - Normal distribution	251
D.2.1.2	A2 - LogNormal distribution	251
D.2.2	Experiment B	251
D.2.2.1	B1 - Normal distribution	252
D.2.2.2	B2 - LogNormal distribution	252
D.2.3	Experiment C	252
D.2.3.1	C1 - Normal distribution	253
D.2.3.2	C2 - LogNormal distribution	253
D.2.4	Experiment D	253
D.2.4.1	D1 - Increasing the number of instances	253
D.2.4.2	D2 - Increasing the number of attributes	253
D.3	Results - Figures	253

E	APPENDIX E	279
E.1	Parameters	279
E.2	Results	279
E.2.1	Experiment A	279
E.2.2	Experiment B	279
E.2.3	Experiment C	279

Part I

Where are we? and where do we go to?

1 INTRODUCTION

“Science does not know its debt to imagination.” - Ralph Waldo Emerson

“Imagination is everything. It is the preview of life’s coming attractions.” - Albert Einstein

Over the past century, we have witnessed the increasing role of digital data generation and storage due to advances in computing and electronic devices. This scenario proved to be conducive to the application and development of statistical, mathematical, and computational techniques that recognize patterns in datasets, including those resulted from data clustering methods. The data clustering technique is considered one of the most challenging among mining tasks because it searches for similarities (or relationships) in unlabeled data (JAIN; DUBES, 1988; TAN; STEINBACH; KUMAR, 2005; BOCK, 2008). Such algorithms are required by several fields of knowledge to give an answer to many real-world issues, such as customer segmentation, disease diagnosis, classification of stars and planets, analysis of market trends, definition of taxonomies, identification of genes and proteins, biometric recognition, and behavior pattern analysis (XU; WUNSCH, 2009). Although most of the datasets resulting from the aforementioned problems are composed of mixed attributes, i.e. numerical and categorical features, the data clustering technique has not been so explored to deal with mixed datasets as to deal with a single-type set of data. This fact was highlighted by five recent review articles on mixed data clustering (see in Balaji and Lavanya (2018), Miyamoto, Huynh and Fujiwara (2018), Velden, D’Enza and Markos (2019), Foss, Markatou and Ray (2019), Ahmad and Khan (2019)), which drew, with different deepness, the development scenario of the mixed data clustering technique. Due to the incipience of this theme and some specific improvements in related topics, we believe that there are opportunities concerning:

1. devising a representation for the main choices the authors made in a knowledge-discovery study;
2. deepening understanding of the mixed data clustering technique by multi-views of its past;
3. addressing mixed data clustering problems using optimization methods.

1.1 Objectives and outline of the thesis

As mentioned in the previous section, this thesis aims to contribute to the advancement of the mixed data clustering technique and related fields by proposing a new way to characterize knowledge-discovery documents, understanding this theme deeper

through a multi-views of its past, and addressing optimization methods to solve mixed data clustering problems. To make such a contribution, the following general research objectives are defined.

- Objective 1. To create a standard representation string that characterizes a scientific document in the knowledge-discovery field by the description of the main choices the authors made in their studies;
- Objective 2. To provide a multi-view of the past of the mixed data clustering technique throughout a wide and in-depth systematic literature review, and bibliometric analysis;
- Objective 3. To propose the modeling of an optimization method for the feature balancing problem, which requires the definition of the weight for each type of features, and provide a wide and in-depth analysis of the behavior of the considered methods facing different characteristics of datasets;
- Objective 4. To propose a set of optimization methods for the feature weighting problem, which requires the definition of the weights of each attribute belonging to the mixed datasets, and provide a wide and in-depth analysis of the behavior of the considered methods facing different characteristics of datasets.

To achieve this objectives, this document is organized in five parts, as follows.

- Part I comprehends two chapters. In Chapter 1, we introduce the problem and the objectives of this thesis. In Chapter 2 we present the main concepts the remaining chapters will based to build their contributions;
- Part II is composed of Chapter 3, which brings the standard representation string proposal. Such a proposal is a first attempt to provide a way for the community to quickly recognize the main characteristics of a document in the knowledge-discovery field. The context to propose such an objective was that until that moment there was not any standard representation to this field, though such string already exists in other areas to characterize queuing systems and scheduling problem, for instance;
- Part III brings Chapter 4 in which we present the systematic literature review on mixed data clustering. The context to proposed such an objective was the fact that in 2018 and 2019, five traditional review articles were proposed of which four provided an analysis restricted to some types of data transformation methods. Only the review carried out by Ahmad and Khan (2019) brought a more extensive and complete study on the mixed data clustering, discussing a wide range of works over the different classes of data clustering algorithms, in addition to conducting a wide

discussion on the challenges and open research questions on this topic. Although the five aforementioned works had revised mixed data clustering methods, there were still some gaps to be filled. Motivated by the mentioned facts, we established the hypothesis that the mixed data clustering technique could be mapped more deeply, through a systematic process of searching, selecting, and discussing scientific papers, in which all classes of algorithms and approaches related to this topic could be considered. Thus, in Chapter 4 we conducted a wide and in-depth systematic review of the literature on mixed data clustering, including the activities that precede and follow such pattern recognition technique in the Knowledge Discovery in Databases (KDD) process;

- Part IV is composed of Chapters 5 and 6 in which we present the feature balancing and weighing problems in mixed data clustering technique. Feature balancing in mixed data clustering is an essential task to equilibrate the contribution of numerical and categorical attributes in the clustering process. Although some types of algorithms already have automatic balancing of the variables due to their intrinsic characteristics, distance-based mixed data clustering algorithms do not benefit from this advantage. In this case, it is necessary to define weights that compensate for the imbalance of the dissimilarities of each type of feature. The feature weighting provides emphasis on the contribution of relevant attributes to obtain a better performance of the clustering methods. Although this activity is already widely explored in single-type data clustering algorithms, few approaches have been proposed for mixed datasets. As these weights are highly dependent on the characteristics of the datasets, our motivation was using optimization methods to solve these issues;
- Part V brings Chapter 7 in which we conclude this thesis, present the papers we published, and bring the paths for future research.

2 MAIN CONCEPTS

"I seem to have been only like a boy playing on the seashore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me."
Isaac Newton

The objective of this chapter is to describe the main concepts that will be used in this thesis. More specific definitions will be presented in the appropriate chapter. Therefore, here are defined some notations, terminologies and premises of the KDD process, dataset representation, the mixed data formalization, the strategies available to cope with this kind of data, the classification of data clustering algorithms, and some measurements that indicate the quality of a clustering.

2.1 KDD process

According to Fayyad, Piatetsky-Shapiro and Smyth (1996a), the term *Knowledge Discovery in Databases* was coined in 1989 at the first KDD workshop (PIATETSKY-SHAPIRO, 1991) in order to make clear that the knowledge is a result of a process guided by data. The authors also established a specific meaning for the terms data and pattern, defining that data are a set of observations, cases, or facts, and a pattern is an expression of some description of a subset of data, and/or a model applicable to the subset.

For Fayyad, Piatetsky-Shapiro and Smyth (1996b, p. 83), the KDD process is defined as follows:

"Knowledge Discovery in Databases is the nontrivial process of identifying valid, novel, potentially, useful, and ultimately understandable patterns in data".

From this definition, Figure 1 shows the activities of the KDD process, their precedence and relations. The process begins with the target dataset, which is obtained by refining the raw data (selection). Next, the preprocessing task is carried out with the following activities, if appropriate: deleting or filling the missing values, and/or deleting or manipulating of the noisy data and outliers. Afterward, the preprocessed dataset might need to pass by a transformation, which can be included the feature scaling, and/or the dimensionality reduction, which can be done through feature selection or transformation methods (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b). Then, a data mining technique needs to be chosen, followed by the choice of an appropriate pattern

recognition algorithm. Thus, the dataset is imputed in the elected algorithm so that it can be executed (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b). The result from the pattern recognition algorithm needs to be validated and made it more understandable, which is done through performance measures, visualization and interpretation methods, respectively. Thus, these last tasks of the KDD process support the expert in understanding the analysis and discovering a possible new knowledge (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b).

Figure 1 – The KDD process (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a)

In Chapter 3, the KDD process is used as inspiration for the first proposal of this thesis, the standard representation string, therefore each step and its relations are explained more in-depth in that chapter. The following sections present the concepts as they appear in the KDD process, i.e. the data, transformation methods, the pattern recognition algorithm, which in this case is the data clustering technique, and the performance measures.

2.2 Attribute-value format

The selection and transformation of raw data into target data comprise, among other activities, their organization in a convenient representation to be processed by pattern recognition algorithms. Although there are several types of representations, one of them is widely used in this area, called the attribute-value format. This format establishes the data is structured into a matrix, where each row represents an observation and each column represents a characteristic of these observations (BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2015). Table 1 presents the attribute-value format representation of a dataset D with N observations and M attributes.

For the sake of clarity, it should be noted that this document considers the following words as synonyms of *observation*: record, object, tuple, instance, and case. In addition, the words characteristic, variable, and feature are considered as synonyms for *attribute*.

Table 1 – Attribute-value format of a dataset D with N observations e M attributes

Observations	Attributes				
	p_1	p_2	p_3	$\dots$	p_M
r_1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$\dots$	$x_{1,M}$
r_2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$\dots$	$x_{2,M}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$
r_N	$x_{N,1}$	$x_{N,2}$	$x_{N,3}$	$\dots$	$x_{N,M}$

It should also be noted that it is not the scope of this thesis the usage of unstructured datasets, as well as semi-structured ones.

2.3 Mixed dataset

In this thesis, a mixed dataset is a set of data described by numerical and categorical features, where numerical attributes, also known as quantitative variables, comprehend those whose values are on the metric scale (interval or ratio), i.e. those that can be measured, such as height, temperature and water flow rate; and categorical attributes, also known as qualitative attributes, which are values that represent a class, a name, and are expressed on a nominal or ordinal scale. Nominal categorical attributes are those that do not follow an order, such as color, gender and blood type. The attributes expressed in the ordinal scale are composed of classes that follow an order, e.g. water level (low, medium and high), education level (undergraduate, master, doctorate) and index satisfaction (low, neutral, high). The values of numerical attributes can be discrete (belonging to the set of integers), binary (a particular case of discrete values, where only two numbers are possible, usually zero and one), and continuous (belonging to the set of real numbers). With regard to categorical attributes, it is common for classes expressed in names to be replaced by discrete numbers. This assignment of discrete values to classes is agreed to start with number one and continue with the addition of one unit to each new class of the attribute.

Table 2 shows an example of a mixed dataset in the medical field to highlight what one considers to be a mixed dataset in this document.

Table 2 – A toy example of a mixed dataset - Raw data

ID	Age (years)	Height (meters)	Fever	Site of the disease	Level of pain
1	15	1.55	yes	neck	high
2	39	1.83	no	chest	low
3	28	1.76	yes	head	medium
4	54	1.68	yes	feet	high

In Table 2, it is possible to observe that there are numerical and categorical attributes, specifically two numerical (age and height), two nominal (site of the disease and fever) and one ordinal (level of pain). The first attribute, called ID, is not selected as the target data for the analysis, as it only identifies the record and has no meaning for the problem in question. In order to organize the notations and terminologies, it is defined the set of variables that represent the numerical and categorical attributes are pn e pc , respectively, so that pn_{v_n} and pc_{v_c} indicate a specific attribute, where $v_n = 1, 2, \dots, m_n$, $v_c = 1, 2, \dots, m_c$, m_n and m_c are the number of numerical and categorical features, respectively, resulting in $m_n + m_c = M$.

Table 3 shows the dataset presented in Table 2 with the changes described in the previous paragraphs, i.e. with the elimination of the ID attribute, and the replacement of the names presented in the categorical features by discrete values, making this set of data closer to a target dataset to be handled by a pattern recognition algorithm.

Table 3 – A toy example of a mixed dataset - Closer to a target data

Age (years)	Height (meters)	Fever	Site of the disease	Level of pain
13	1.50	1	1	3
48	1.83	0	2	1
24	1.76	1	3	2
74	1.60	1	4	3

Some other procedures are necessary to prepare a dataset to be analyzed by a pattern recognition algorithm, among them transformation methods might be applied and features may be selected, weighted, or balanced. Some of these activities are explained in the following sections, however, firstly one highlights what is not a mixed dataset in this document in order to make clear the scope of this thesis.

2.3.1 What is not a mixed data in this thesis?

For the sake of clarity, it is important to highlight that no type of attributes other than numerical and categorical data (restricted to the definition given in this chapter) are part of the scope of this thesis. One highlights some examples of types of attributes that are not the focus of this work, such as the symbolic ones, which characterize an interval, e.g. those present in the fuzzy logic; and the coded ones, which represent a coding structure, such as IP addresses, ZIP codes, structures that represent DNA, chemical structures, among others. In addition, it is not the scope of this thesis datasets increased by zero, nor the single-type datasets, i.e. those that present only one type of attribute, numerical or categorical.

2.4 Strategies to cope with mixed data

The literature offers two ways to deal with mixed data, specifically when the pattern recognition algorithm is a data clustering algorithm. The first strategy addresses a transformation method to transform a mixed dataset into a single-type set of data. The second strategy, especially used in distance and model-based data clustering algorithms, addresses a specific function for each type of feature, summing them in a unique objective function (FOSS; MARKATOU; RAY, 2019). The following subtopics explore the advantages and drawbacks of each strategy from the perspective of data clustering algorithms.

2.4.1 Transformation methods

A common advantage of using transformation methods is the fact that the analysis will be benefited from a higher number of single-type data clustering algorithms as well as their scientific advancements, because the research on mixed data clustering algorithms is very incipient and the variety of this kind of approach is low (FOSS *et al.*, 2016; FOSS; MARKATOU; RAY, 2019). For the sake of exemplification, the discrepancy in the development of single and mixed data clustering can be measure in the time taking into account two seminal partitional data clustering algorithms (see more about the classification of the data clustering algorithms in Section 2.5), the k-means and k-prototypes algorithms. The difference, regarding date of publication, between both proposals is approximately thirty years, since the former was proposed in 1967 (MACQUEEN, 1967), and the latest was proposed in 1998 (HUANG, 1998). The state-of-the-art of mixed data clustering technique can be seen in Chapter 4.

Concerning the drawbacks of such transformation methods, one can point that transforming the dataset might distort the comprehension of the data, making them less intelligible, i.e. less understandable from the semantic point of view since the transformed value has no direct meaning in the real problem domain. In addition, some of these approaches make the data clustering algorithm intractable, because they add too many new attributes, thereafter the algorithm suffers from the curse of dimensionality, specially when the categorical attributes are transformed into numerical ones. As can be seen from the last statement, some concerns with respect to transformation methods are directly depended to the direction of the transformation, i.e. mixed data to numerical dataset or mixed data to categorical set of data. Thus, in the following paragraphs one briefly describes the key differences and implications of both directions.

Transforming a mixed dataset into a categorical set of data consists of applying discretization methods in the numerical attributes, which means that the continuous values need to be grouped in k_B parts so that each part is characterized as one class of the new categorical feature. Table 4 shows the numerical attributes (age and height) from Table 3 transformed by a discretization method that considers $k_B = 3$ for age and $k_B = 2$ for

height.

Table 4 – A toy example of a mixed dataset - Transformed data (discretization method)

Age (years)	Height (meters)	Fever	Site of the disease	Pain level
1	1	1	1	3
2	2	0	2	1
2	2	1	3	2
3	1	1	4	3

As can be seen, the new attribute that expresses the age of the patient is now a categorical feature with three classes, where the first represents young people, the second the adults, and the third the elderly. One of the consequences in applying such transformation methods is the possibility of losing relevant information for the analysis, since the data are aggregated and soft phenomenon might be hidden. The pattern recognition algorithm may also be harmed because the pattern might need the original data to be revealed (HARCH *et al.*, 1999). Finally, one challenge that discretization methods impose is the choice of intervals the continuous data will be grouped (k_B). Such number has a directly influence in the pattern recognition performance, thereafter in the results of the analysis (FOSS; MARKATOU; RAY, 2019).

With respect to transforming a mixed dataset into a numerical set of data, one can apply methods that make each class of categorical attributes be transformed into a new binary feature, as the dummy coding method does. In this case, the dataset increases in the number of attributes, which might result in an intractable problem, especially when the mixed dataset has many categorical features since the transformed dataset will be characterized as a sparse and high-dimensional set of data, making the pattern probably less defined. Concerning the dummy coding method, this approach transforms categorical attributes into numerical ones by adding new columns with 0 and c values, where each new column shows whether a class of a categorical feature occurs in a specific register (indicating this with a constant c), or not (with the number zero). It is very common to define the c value as the number one, although this choice being totally arbitrary (HENNIG; LIAO, 2013; FOSS *et al.*, 2016; FOSS; MARKATOU, 2018; FOSS; MARKATOU; RAY, 2019). Another typical method is applying the z-score normalization to unit variance in each dummy coding attribute, resulting in $c \approx \frac{1}{p_l(1-p_l)}$, where p_j is the frequency of registers that have the l -th label of the categorical feature dummy coded (FOSS; MARKATOU, 2018, p. 3).

Taking $c = 1$ into account, Table 4 brings the transformed dataset after applying the dummy coding method into the dataset from Table 4.

As can be seen, the new dataset contains continuous and binary features, and, as all

Table 5 – A toy example of a mixed dataset - Transformed data (Dummy coding method)

Years	Height	Fever	Neck	Chest	Head	Feet	High	Medium	Low
15	1.55	1	1	0	0	0	1	0	0
39	1.83	0	0	1	0	0	0	0	1
28	1.76	1	0	0	1	0	0	1	0
54	1.68	1	0	0	0	1	1	0	0

the attributes represent numbers, they can be handled by numerical dissimilarity function, such as Euclidean distance, which will be presented in the following section. In this scenario, one usually considers that all the features contribute in the same way to the distance function. However, it is not true. The values of the distance regarding binary attributes will always be zero *or* one, whereas the values of the distance between two normalized numerical attributes will always have more values than only zero *or* one. This fact imposes an unbalanced contribution of both types of features in the dissimilarity function (MODHA; SPANGLER, 2003) (see theoretical calculations that show this affirmative in Foss *et al.* (2016, Sec. 2.2)). Therefore, the question is how to define the appropriate value for the constant c in order to provide the balancing of the contribution of both types of attributes. Chapter 5 explores more in-depth this problem, and provides a proposal to solve this challenge.

As could be seen, using transformation methods is one possibility to cope with mixed datasets. Although these approaches are an interesting alternative for many real-world cases, not all of them can be handled with them. For this reason, the literature brings another strategy to cope with mixed datasets, the composite assignment functions, which attribute two different treatments for each type of feature, the numerical and categorical ones. In the next section, one describes the most usual approaches applied in distance and model-based data clustering algorithms.

2.4.2 Compound assignment functions

Compound assignment functions are understood as mathematical models that assign records to the groups in which they are most similar. These models can be, for example, a composition of dissimilarity functions, correlation, or statistical models. In the following sections, one makes the usage of the composite assignment functions more explicit in the mixed data clustering algorithms.

2.4.2.1 Dissimilarity functions

A dissimilarity function $d(x, y)$ measures how dissimilar two observations x e y are, while a similarity function $s(x, y)$ quantifies the similarity between x e y . In many cases,

dissimilarity functions are a distance measure, and because of this, they need to respect three mathematical axioms, which are

- i Symetry. $d(x, y) = d(y, x)$.
- ii Positivity. $d(x, y) > 0$, if $x \neq y$.
- iii Reflectivity. $d(x, y) = 0$, if and only if $x = y$.

It is also defined that a distance is a metric if this, in addition to respecting the above axioms, also respects a fourth axiom, the triangular inequality, which is expressed as follows.

- iv Triangular inequality. $d(x, y) + d(x, z) \geq d(y, z)$, for all $x, y, z \in X$.

Euclidean distance is a metric widely used by data clustering algorithms, especially by the partitional and hierarchical ones (see more about the classification of data clustering algorithms in Section 2.5), to quantify the dissimilarity between observations in numerical datasets. Such distance is defined by the square root of the sum of the squared differences between two points (x, y) , which is expressed by:

$$d_E(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} \quad (2.1)$$

For categorical data, the dissimilarity function called simple matching (KAUFMAN; ROUSSEEUW, 1990) is commonly used to measure the difference between qualitative attributes. Such a function is expressed by:

$$d_{sm}(x, y) = \sum_{i=1}^n \delta(x_i, y_i), \quad (2.2)$$

where

$$\delta(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i; \\ 1 & \text{if } x_i \neq y_i \end{cases} \quad (2.3)$$

Considering these two described dissimilarity functions, the composite assignment function resulted from them is

$$d(x, y) = d_E + \omega d_{sm}, \quad (2.4)$$

where ω is a weight attributed to the sum of the dissimilarities of the categorical features to compensate for the difference between these dissimilarity values and the sum of the dissimilarities of the numerical attributes.

The systematic review on mixed data clustering presented in Chapter 3 shows that many of the works developed nowadays employ efforts to define the weight ω so that the dissimilarities values of both types of attributes can be expressed in the same scale. In that same chapter, it is also possible to verify that the Equation 2.4 is the composite assignment function of the k-prototypes algorithm (HUANG, 1998), which was the first partitional data clustering algorithm proposed for mixed datasets, whose strategy was to join the dissimilarities of both types of attributes in a single function.

Another composite assignment function widely used in mixed data clustering algorithms is the Gower distance, which was proposed by Gower (1971) to make dissimilarity measure between mixed vectors possible. This distance applies different treatments for each type of feature, considered by this distance, which are: numerical, nominal, and ordinal ones. The Gower distance (D_{ij}) is calculated as being the complement of the Gower coefficient (S_{ij}), which is a similarity measure that ranges between 0 and 1, where 1 represents the highest value of similarity, and 0 the opposite. Then, $D_{ij} = 1 - S_{ij}$. The weighting Gower coefficient is detailed ahead (GOWER, 1971, p. 861, Equation 5).

$$S_{ij} = \frac{\sum_{k=1}^M s_{ijk} \omega_k(x_{ik}, x_{jk})}{\sum_{k=1}^M \delta_{ijk} \omega_k} \quad (2.5)$$

where $\delta_{ijk}(x_i, y_j)$ indicates whether the comparison between x_i and x_j is possible in the attribute k , which is set equal to zero, if such comparison is not possible, and set to one otherwise. When $\sum_{k=1}^M \delta_{ijk}(x_i, y_j) = 0$, then S_{ij} is undefined. The Gower coefficient also allows assigning different weights (ω_k) for each attribute k . Besides, s_{ijk} represents a specific function for each of the three types of attributes considered by this measure (GOWER, 1971, p. 859). For quantitative features, one has that s_{ijk} is defined as

$$s_{ijk} = 1 - \frac{|x_i - x_j|}{R_k}, \quad (2.6)$$

where R_k is the range of the attribute k . For qualitative attributes, s_{ijk} is defined as:

$$s_{ijk} = \begin{cases} 1 & \text{se, } x_i = x_j; \\ 0 & \text{se, } x_i \neq x_j \end{cases} \quad (2.7)$$

and, for dichotomous attributes, one has that s_{ijk} is 1 when both observations have the presence of the character, and 0 in the other cases; and $\delta_{ijk}(x_i, y_j)$ is 0 when both characters are absents, and 1 in the other cases.

The Gower distance also presents the same challenge as the function presented in Equation 2.4, i.e. it is need to balance the contribution of the dissimilarities of the different type of attributes. Some data clustering algorithms accept a dissimilarity matrix as input. This matrix contains the dissimilarity values of all the observations, as shown as follows:

$$D = \begin{bmatrix} d(x_1, x_1) & d(x_1, x_2) & d(x_1, x_3) & \dots & d(x_1, x_n) \\ d(x_2, x_1) & d(x_2, x_2) & d(x_2, x_3) & \dots & d(x_2, x_n) \\ \vdots & \vdots & \vdots & \dots & \vdots \\ d(x_n, x_1) & d(x_n, x_2) & d(x_n, x_3) & \dots & d(x_n, x_n) \end{bmatrix}$$

In this subsection, it was possible to note that there are some possibilities to measure the dissimilarity between two observations that are described by numerical and categorical attributes. In the next subsection, one presents the same idea with respect to assigning a specific approach to each type of feature, however, now using probability distribution functions instead of dissimilarity functions.

2.4.2.2 Statistical models

The finite mixture models are widely used to handle mixed data because they assume the data are generated by finite probability distribution functions. Mathematically, such models are described as follows.

Let C be a set of k groups ($C = C_1, C_2, C_3, \dots, C_k$) with probability density functions $f_1, f_2, f_3, \dots, f_k$, and probabilities $\omega_1, \omega_2, \omega_3, \dots, \omega_k$, respectively. The probability of an observation r has been generated by the group C_j is

$$P(r, C_j) = P(C_j)P(r|C_j) = \omega_j f_j(r) \quad (2.8)$$

Thus, the probability of an observation r has been generated by the set of the groups C is

$$P(r|C) = \sum_{j=1}^k \omega_j f_j(o) \quad (2.9)$$

Assuming that the data are generated in the same independent way, thereafter, for a set of data $D = r_1, r_2, r_3, \dots, r_N$, one has that the probability of the set of data D given the set of the groups C is

$$P(D|C) = \prod_{i=1}^N P(r_i|C) = \prod_{i=1}^N \sum_{j=1}^k \omega_j f_j(r_i) \quad (2.10)$$

In this way, the data clustering problem handled by finite mixture models consists of finding a set C of k probabilistic groups so that the conditional probability $P(D|C)$ is

maximized. As maximizing the $P(D|C)$ is generally an intractable problem by the fact that the probability density function of a group might become a complicated arbitrary form, one assumes (with a compromised of the result) that the probability density functions are parameterized distributions. It is very common observing a finite Gaussian mixture in the model-based numerical data clustering algorithms, while for categorical data the usual choice is considering finite multinomial mixtures.

In this section, one exposes the main strategies to cope with mixed datasets, especially when data clustering technique is the pattern recognition algorithm. In the following section, one briefly describes the main concepts that surround the data clustering algorithms, showing a definition for this technique, a common classification for its algorithms and some performance measures, which are also known as clustering quality index, validation measure, validation index. This set of words are used in this thesis as synonyms.

2.5 Data clustering technique

Data clustering is a data mining task that searches for *natural* groups in unlabeled datasets (JAIN; DUBES, 1988). By working with unraveling patterns, this kind of technique is considered one of the most challenging among data mining tasks (SAXENA *et al.*, 2017). Clustering algorithms are highly demanded by several knowledge fields for a variety of purposes as already pointed out in the first chapter of this thesis. It is also known that it is common for the groups resulted from a data clustering algorithm to be exclusives e only considered the cluster of observations; however Hennig (2015) complements this idea affirming the data to be clustered by these kinds of algorithms can be "variables, dissimilarities, or weighted edges in a graph", and the resulted groups might be overlapped or non-exhaustive, i.e. they can be fuzzy, not only crisp. Besides this technical data clustering definition, Hennig (2015) provides a philosophical and comprehensive discussion about the meaning of clustering in which the author reflects about the *ground truth* in cluster analysis. Such a discussion is presented in Chapter 5.

2.5.1 Classification

Over the years, data clustering algorithms have been proposing to solve several kinds of problems, which demands specific strategies to handle the characteristics of the datasets, constraints imposed by the problem, and technical issues, such as improving the time consuming process taken by the algorithms in order to make them feasible to solve real-world problems. In this scenario, the literature has attempted to classify most part of these algorithms to organize the theoretical outline in this field. Thus, many classifications have been proposed over the years considering different common characteristics of the algorithms. The classification adopted in this thesis appears in many articles and books, which proposes

five classes of data clustering algorithms (XU; WUNSCH, 2009), which are: i) partitional algorithms: those that separate the objects into k groups, considering a dissimilarity function; ii) hierarchical algorithms: those that separate the objects hierarchically into all possible groups, considering a dissimilarity measure; iii) model-based algorithms: those that separate the data into k groups, considering statistical models for that; iv) neural network-based algorithms: those that separate the objects following neural networks rules, and v) *others* those that are specific approaches in many emergent themes, and do not fit within the previously classes. To obtain a more detailed explanation about each class, see in Tan, Steinbach and Kumar (2005), Xu and Wunsch (2009). Basing on the five mentioned classes, Chapter 4 proposes a sixth one and takes all of them into consideration to organize the results of the systematic literature review on mixed data clustering.

Following the KDD process structure, this chapter has already described the data, transformation methods, the pattern recognition technique of this thesis (data clustering). Thus, one finishes the main concepts with some common performance measures the scientific community adopts in data clustering works. A performance measure quantifies the quality of the clustering resulted from the data clustering algorithm. Although it is a consensus that there is not the best performance measure regardless of the algorithm and problem's characteristics and the objective of the analysis, one can choose it among three main strategies, which are also described in the following section.

2.5.2 Performance measures

The first strategy is the external performance measures, which consist of comparing the resulted clusters from the data clustering algorithm to the *ground truth* clusters based on the supposition that the dataset has already had the clusters defined for each observation. This strategy in real-world problems is not effective because when a specialist considers the data clustering technique to analyze the data, he or she does not even know whether there are groups in that dataset or not. Hence, this strategy is mostly used to compare algorithms in theoretical experiments, since simulated datasets can be built and served as *ground truth*. There are many concerns regarding works that take datasets with labels from classification tasks and assume that these labels are the clusters. Although these concerns are expressed by many researchers for the scientific community, there are still new papers that have been publishing data clustering works considering classification datasets and applying external performance measures to quantify the quality of the clusters an algorithm obtained. Some examples of external performance measures are Rand index (RAND, 1971), adjusted Rand index (ARI) (HUBERT; LEVIN, 1975), and Fowlkes and Mallows index (FOWLKES; MALLOWS, 1983).

The second strategy consists of applying internal performance measures, which mean searching for some desired characteristics on the resulted clustering, such as homogeneity

and cohesion. There are many internal performance measures, however there is not one that is the best because each analysis will impose more attention in specific and different characteristics. It is very common using such strategy to define the number of the clusters a dataset might have, although it can be also done using the third strategy, the relative performance measures. Some examples of internal performance measures are silhouette width (ROUSSEEUW, 1987; KAUFMAN; ROUSSEEUW, 1990), Dunn index (DUNN, 1974), Davies-Bouldin index (DAVIES; BOULDIN, 1979), also known as *DB index*, PBM index (PAKHIRA; BANDYOPADHYAY; MAULIK, 2004), and c-index (HUBERT; LEVIN, 1975). To obtain more information about internal performance measures, see in Vendramin, Campello and Hruschka (2010) and Xiong and Li (2014).

The third strategy quantifies the quality in a relative way, which means assuming the comparison of the same validation criteria in two, or more, different situations, or validating the resulted clusters by checking a statistic hypothesis. Generally, these hypotheses are (i) random position hypothesis, in which is assumed the observations from the datasets are “independent samples from a d -dimensional distribution, such as uniform over a hypercube or hypersphere or unimodal Gaussian” (DUBES; JAIN, 1979, p. 240); (ii) random graph hypothesis, in which is assumed the values from the dissimilarity matrix are equally likely (JAIN; DUBES, 1988; GORDON, 1994), and (iii) random label hypothesis, in which is assumed that all possible partitions of observations in k groups are equally likely (DUBES; JAIN, 1979; JAIN; DUBES, 1988). To obtain the distribution of the baseline under the null hypothesis, statistical sampling techniques are used, such as Monte Carlo analysis (JAIN; DUBES, 1988). Some examples of internal performance measures are Bayesian information criterion (SCHWARZ, 1978) and Akaike information criterion (AKAIKE, 1974).

In Chapter 4, it is pointed out all the performance measures used by the authors in the 160 works that were considered as the result of the systematic literature review on mixed dataset. In the following paragraph, one explains the performance measure used in this thesis (ARI), which belongs to the first strategy, i.e. it is an external performance measure. The choice of an external measure was done because the practical results of this thesis are theoretical, and the experiment results were done from simulated datasets.

The ARI results by a comparison of the resulted clusters from the data clustering algorithm and the *ground truth* clusters. To provide a better explanation, let consider a set of group C resulted from a data clustering algorithm, with a independent and predefined partition P from the dataset D . Let x_i and x_j be two observations from D . The mentioned comparison is realized following the cases ahead

Case 1 x_i and x_j belong to the same cluster of C and the same category of P . The number of objects in this case is represented by the letter a .

Case 2 x_i and x_j belong to the same cluster of C , but to different categories of P . The number of objects in this case is represented by the letter b .

Case 3 x_i and x_j belong to different clusters of C , but the same category of P . The number of objects in this case is represented by the letter c .

Case 4 x_i and x_j belong to different clusters of C , and to different categories of P . The number of objects in this case is represented by the letter d .

Considering $S = a + b + c + d$, $s_1 = a + b$ e $s_2 = a + c$, one has the Rand index is

$$Ind = \frac{a + d}{S} \quad (2.11)$$

Thus, the adjusted Rand index is

$$ARI = \frac{Ind - E(Ind)}{max(Ind) - E(Ind)}, \quad (2.12)$$

where Ind is the Rand index, $E(Ind)$ is the expectation of Ind , and $max(Ind)$ is the maximum of Ind . Therefore, the closer the ARI index values are to 1, the higher the quality of the clustering, as opposed to a zero ARI value, which represents poor clustering quality.

In this chapter, one presented the main concepts that the remaining part of this thesis will use to delivery the results of this research. Eventually, specific definitions and theoretical discussions will be made in appropriate chapter. In the next chapter, it is presented the first contribution of this work, a proposal for defining a standard representation string to characterize scientific document in knowledge-discovery field.

Part II

Organizing to go beyond

3 STANDARD REPRESENTATION: FIVE-FIELDS NOTATION

“We build too many walls and not enough bridges” - Isaac Newton
“Simplicity is the ultimate sophistication” - Leonardo da Vinci

In the late 1990s, Fayyad, Piatetsky-Shapiro and Smyth (1996a), Fayyad, Piatetsky-Shapiro and Smyth (1996b) and Brachman and Anand (1996) observing the countless methods emerging in the literature on pattern recognition techniques and related proposals, proposed the first discussions on *Knowledge Discovery in Databases* (KDD), to organize the concepts and practices of that area. From these first works and workshops, the authors defined an interactive and iterative process, called the KDD process, which consists of a set of steps that prepares a raw dataset so that it is able to properly identify useful information (FAYYAD *et al.*, 1996).

The KDD process consists of selecting the target dataset from a set of raw data, considering the purpose of the analysis. Then, the target data are cleaned and transformed, if appropriate. At this point, a data mining task is chosen, followed by the election and execution of a specific pattern recognition algorithm. Last of all, the result of the algorithm is validated, interpreted, and made visible to the decision-maker.

Undoubtedly, those first contributions were unique to structure the knowledge-discovery-in-databases field, which supported its growth consistently. Since it was proposed, the KDD process has been fundamental for organizing practically and theoretically the different methods and concepts in this thematic, which helped many scientists and practitioners in the understanding of how to manage the data to extract useful information from them.

Although the authors established many new concepts at that time, they mentioned some concerns with respect to the development of this field, thereby they pointed out some topics that could be researched and developed. One of them refers to the development of knowledge representations. To be more specific, the original transcription of a work dated of 1996 from Fayyad, Piatetsky-Shapiro and Smyth (1996a, p. 51) follows, showing their explanation on how it would be interesting the creation of “...ontologies, new concepts for representing, storing, and accessing knowledge. Also included are schemes for representing knowledge and allowing the use of prior human knowledge about the underlying process by the KDD system”

More than two decades have already passed and, to the best of our knowledge, there is not any proposal of schemes that represents knowledge in the knowledge-discovery-in-databases field, though these schemes (strings) are a reality in other areas, such as in the scheduling problems, and queuing systems (see more in Section 3.1).

Hence, taking the importance of the KDD process and the potential contribution quoted above into consideration, one proposes a standard representation, based on the KDD process, to characterize the main contributions that scientific documents make in the knowledge-discovery-in-databases field. In addition, the proposed representation also characterizes problem scenario, i.e. the problem complexity with respect to the characteristics of the datasets handled by the scientific documents.

This proposal consists of a first attempt to build a standard representation, and we call it as five-fields notation, which is expressed through a string composed of a set of acronyms that characterizes: i) the scenario, with respect to the characteristics of the datasets, the article considered when it proposed its objectives; ii) the main contributions the described paper did in each one of the main steps of the KDD process, and iii) the auxiliary approaches the authors used to develop their contributions. The justification for this solution is on the fact of the growing interest in this thematic has generated, year by year, more specific publications in the different activities of the KDD process. Besides, such researches contemplate scenarios with different levels of difficulty (regarding the characteristics of the datasets), and the usage of different auxiliary approaches to achieve the proposed objectives. All of these publications are not been characterized by a simple, wide and detailed representation, which might create an obstacle to the reader in quickly identifying its general and specific objectives.

The outline of this chapter is organized as follows. In Section 3.1, one does a literature review to describe other standard representations, proposed in other areas. Sections 3.2 presents the standard representation proposal. Section 3.3 exemplifies the proposal for the data clustering technique. Section 3.4 discusses the advantages of adopting the standard representation by the scientific community. Finally, Section 3.5 concludes this chapter.

3.1 Literature review

Many other fields of knowledge use knowledge representations. The scheduling area gives a traditional example. In the late 1970s, Graham *et al.* (1979) proposed a string to describe scheduling problems. The authors observed that any kind of scheduling problem had been composed of the same three essential information: the machine environment, job characteristics, and the optimal criteria.

As a result, they proposed a three-part string $\alpha/\beta/\gamma$ to characterize scheduling problems, where α informs the machine environment, β the job characteristics, and γ the optimality criteria. Since Graham's three-part string was proposed, the scheduling-field scientific community has been using this string to announce its objectives (the characteristics of the problem to be solved) in its publications.

Thus, a researcher in the scheduling area, by seen Graham's string in a scientific document, is able to quickly identify the type of problem to be solved by it, the problem complexity, and the environment that it is considered. An example of the scheduling problem representation is $1/prec/L_{max}$, which represents a problem of a single machine, subjected to general precedence constraints, with the objective to minimize the maximum lateness.

Another example is the representation of queuing problems. At the beginning of the 1950s, Kendall (1953) identified that a queuing system could be represented with three main information: i) the input, which represents the distribution of the inter-arrival times; ii) the queuing-discipline, which represents the way the elements of the queue are attended by the services, and iii) the service mechanism, which represents the number of servers. For this reason, the author proposed a string composed of three main parts to describe the queuing systems. The Kendall notation, as it is currently known, consists of a string composed of three main information: the *inter-arrival times distribution/the service-time distribution/the number of parallel servers*. The three-part string is a contracted version of Kendall notation. It is used when some information has its value in the default configuration, therefore they are omitted. An example of the Kendall contracted label is D/G/3, which means a queuing system with regular arrivals, no special assumption about the service-time distribution, and with three servers (KENDALL, 1953).

An extended string is composed of five parts A/S/c/K/Q, where A refers to the inter-arrival time distribution, S is the service-time distribution, c is the number of servers, K is the maximum allowable system size (if omitted, it is assumed to be infinite), and Q represents the queuing discipline, which has the default configuration *first-come, first-served*.

3.2 The standard representation

The KDD process inspired the standard representation proposal along with the main challenges the researchers are facing nowadays, specifically with respect to the characteristics of the target dataset as such characteristics might add extra complexity in the problem to be solved. Thus, to build the string of the standard representation, one joined some tasks from Figure 1 into three main steps, which one calls herein of preprocessing, processing, and post-processing steps. Figure 2 shows that tasks from the KDD process belong to each one of the three steps.

The preprocessing step comprehends since the first activities regarding the raw data until, and inclusive, the transformation task. The second step consists of the choice of the data mining technique, appropriate pattern recognition algorithm, and its execution. Finally, the last step comprehends the evaluation of the algorithm result, visualization and/or interpretation of it, if appropriate. Thus, the knowledge is achieved if what was

Figure 2 – The KDD process with the three main steps (FAYYAD; PIATETSKY-SHAPIO; SMYTH, 1996a)

discovered is valid, novel, and potentially useful.

Considering the three main steps defined, it is relevant to define the importance of informing the scenario the problem is placed, related to the characteristics of the datasets. It is common the scientific works consider the trivial scenario in its firsts proposals concerning any problematic. One understands that the trivial scenario is characterized by a dataset that has only numerical attributes, static data (i.e., it does not change over the time), which is not high-dimensional, which is centralized (i.e., it is not distributed over many files, computers, and/or servers), and that is not too large neither too small regarding the number of observations. Beyond the trivial scenario, the datasets can impose additional challenges to solve the problems. Because one thinks the nontrivial characteristics of the datasets are very important to distinguish the complexity of the problems, one adds this information on the string that represents the standard representation.

Hence, this proposal consists of a string composed of a set of acronyms that (i) represents the nontrivial characteristics of the datasets that were considered by the described document; (ii) informs all the activities belonged to the preprocessing step that the described document contributed; (iii) informs all the activities belonged to the processing step that the described document contributed; (iv) informs all the activities belonged to the post-processing step that the described document contributed, and (v) informs all the auxiliary approaches that the authors from the described document used to achieve the solution, proposed in this work.

Thus, the five-fields notation is represented as DTS/PRE/PRO/POS/AUX, where DTS represents the acronym of dataset, PRE is an acronym for preprocessing tasks, PRO

is an acronym for processing tasks, POS is an acronym for post-processing tasks, and AUX is an acronym for auxiliary approaches.

For the sake of illustration, Figure 3 shows the standard representation proposal.

Figure 3 – The standard representation proposal, the five-fields notation

In the following subtopics, one details each element of the string, providing a suggestion of acronyms for some characteristics of the datasets, and for some tasks belonging to the preprocessing, processing, and post-processing steps, as well as acronyms for the auxiliary approaches. It is important to note that one does not have the intention of exhausting the possibilities of characteristics, tasks, and methods in this chapter.

With respect to the acronyms, one suggests they are built using the first letter of the respective word following by the two consecutive consonants. If the characteristic of the dataset, task, or method is expressed by two words, then one suggests the authors use the first letter of the first word following by the first next consonant of this word separated by a hyphen of the acronym of the second word, which is built of the same way when there is only one word. If there are more than two words, one suggests to build the acronym of the same way of the second scenario described here, except to the second letter of the first block (before the hyphen), which must be the first letter of the penultimate word. Propositions and articles do not compose the acronyms, and in the absence of consonants, take the vowels. Note that the acronyms and the way they are built are only a suggestion. One details the limitations of this suggestion and other concerns in Section 3.4.

3.2.1 DTS - Datasets

As already mentioned, the dataset element represents nontrivial characteristics of the target dataset the described document used to propose solutions for the problem addressed. Nontrivial characteristics usually add more challenges to solve problems in data-driven issues. Some of these characteristics are listed in Table 6 along with their acronyms.

Table 6 – Acronyms for characteristics of the datasets

Acronym	Meaning
CTG	Categorical
DST	Distributed
DYN	Dynamic
HG-DMN	High-dimensional
LRG	Large
MXD	Mixed

Although the reader might already know the meaning of many characteristics listed in Table 6, one gives a brief explanation of each one to make sure that the acronym is used appropriately on the string of the standard representation. One gave the name *distributed* for those datasets that have their observations and/or attributes spread throughout many files, computers, and/or servers. A dynamic dataset is a set of data that is constantly updated with new data, i.e., there is a dynamic environment over time with respect to the arrival of data to the pattern algorithm manages. High-dimensional datasets are those that the number of attributes is extremely high, becoming the algorithm executions very difficult. Large datasets are those that have a large number of observations. Categorical datasets are composed of categorical attributes, i.e. nominal and ordinal features. Finally, mixed datasets are composed of numerical (continuous and discrete) and categorical (nominal and ordinal) attributes.

3.2.2 PRE - Preprocessing tasks

The second component of the string refers to tasks from the preprocessing step that the described document might have contributed. Some of these tasks are listed in Table 7 along with their acronyms.

Dimensionality reduction means reducing the number of attributes, which can be done by selecting relevant features and deleting the other ones, or transforming the original features' space into another one that can be a combination of the original attributes. Feature scaling is the operation that changes the scale of the attributes, such as normalization or standardization. Handling missing and noisy data, as well as outliers, means defining a strategy to cope with these kinds of values.

Table 7 – Acronyms for the activities from the preprocessing step

Acronym	Meaning
DM-RDC	Dimensionality reduction
FT-SCL	Feature scaling
MD-HND	Missing data handling
ND-HND	Noisy data handling
OT-HND	Outlier handling

3.2.3 PRO - Processing tasks

The processing component refers to tasks from the processing step that the described document might have proposed a contribution. As this step refers to a specific data mining technique, one does not suggest any acronym here. One hopes the scientists specialized in each data mining technique can propose appropriate acronyms for it. In Section 3.4, one exposes some ideas to make feasible the definition, validation, and searching of the acronyms to compose the string of the standard representation. Because one works with data clustering algorithms, it is presented in Section 3.3 a table regarding the acronyms for this data mining task. One also shows some standard representation strings for a few scientific articles that were selected for this purpose.

3.2.4 POS - Post-processing tasks

The fourth component of the string refers to tasks from the post-processing step that the described document might have contributed. Some of these tasks are listed in Table 8 along with their acronyms.

Table 8 – Acronyms for the activities from the post-processing step

Acronym	Meaning
IN-PTT	Interpretation of the patterns
VL-MSR	Validation measure
VSL	Visualization

The described document will have acronyms in this component if it proposed new methods, or improvements in already published methods, which help the decision-maker do an interpretation of the patterns and/or visualize them. The document in question will also receive an acronym if it proposed new validation measures, or improvements in already published validation measures, to validate the results from the pattern recognition algorithm.

3.2.5 AUX - Auxiliary approaches

Finally, the last component of the string reports the auxiliary approaches that might have been used by the described document. Some of these approaches are listed in Table 9 along with their acronyms.

Table 9 – Acronyms for the auxiliary approaches

Acronym	Meaning
CD-AVL	Code available
CMP	Comparison
EN-THR	Entropy theory
FRM	Framework
MT-PRV	Mathematical proof
OPT	Optimization
PRL	Parallelism
PR-APP	Pure application
RVW	Review
SP-KNW	Specialist knowledge

One understands as auxiliary approaches those that support the authors to achieve their objectives. It is in this component that review papers might match. This is because one represents in this component the documents, e.g., which are written exclusively to announce the availability of a code of an already proposed method, and/or to make a comparison among methods that have already been proposed in the literature, and/or to demonstrate a mathematical proof, a pure application of an already proposed method in a real-world problem, and/or to make a review about a specific topic regarding the knowledge-discovery-in-databases field. One also considers as auxiliary approaches when the authors uses the framework idea, entropy theory, a specialist knowledge, parallelism techniques, or optimization methods, in at least one of the three main steps of the KDD process to support them in their particular goals.

With this fifth component, the standard representation contemplates all the possibilities of contributions a scientific document can deliver to the community in this field of knowledge.

3.3 An example for data clustering technique

For the sake of exemplification, one brings in Table 8 some tasks, along with their acronyms, in which scientific documents usually make contributions in the data clustering technique. In addition, it is given a brief explanation of each of these tasks.

In data clustering algorithms, it is common addressing the feature weighting to emphasize the contributions of the attributes in the clustering results. Another common task is determining strategies to define the centroids because this definition generally

Table 10 – Acronyms for the activities from the processing step (data clustering)

Acronym	Meaning
AT-WGH	Attributes weighting
CN-DFN	Centroids definition
DS-FNC	Dissimilarity function
FZ-STF	Fuzzy sets
NM-GRP	Number of groups
ST-MDL	Statistical model
SBS	Subspace

influences in a direct way the algorithm's performance. As many clustering algorithms are based on distance, defining new dissimilarities functions, or improving those that already exist is another usually contribution seen in the documents that address this data mining technique. The usage of fuzzy logic through fuzzy sets is widely seen in data clustering problems, especially when the overlap of the groups is desired.

The next task is the determination of the number of groups, which might be the most explored and challenge task in this data mining task. One knows that any data clustering algorithm is going to result in a set of clusters, independently whether those groups are true, i.e. whether those groups are real or whether they are only a forced attempt to make the data match into n number of groups. Then, discovering the number of groups is a hard activity, equally hard of the task to validate the algorithm's result. In many works, the definition of the number of groups and the usage of a validation measure, especially the relative ones, is directly related. Here, the standard representation receives the acronym from the task of the number of groups if the authors addressed this challenge, even if they just use an already proposed validation measure.

Another common task is to define the statistical model that guides the discovering of the clusters because many data clustering algorithms are based on statistical models. Finally, the last task Table 10 shows is the subspace that represents when the algorithm considers searching clusters in subspaces.

To provide an example of the standard representation, let us consider a hypothetical scientific document that proposes a new approach to determine the number of the groups automatically in a data clustering algorithm, using the genetic algorithm and a validation measure that is widely known in this area, called silhouette width. Let us also consider that this document affirms its proposal considers large datasets in a dynamic environment, and it proposes a strategy to handle missing data. In this case, the standard representation for this hypothetical work should be DYN+LRG/MD-HND /NM-GRP/- /OPT. Thus, with this sequence, it is possible to quickly recognize in which scenario the work is inserted and where it proposes improvements.

For the sake of exemplification, one selects some real articles in the data clustering theme to show how one defines the standard representation for each one. Table 11 brings the mentioned documents and their standard representations.

Table 11 – Standard representation for some real documents in data clustering theme

Document	Standard representation string
Huang (1998)	CTG+MXD/-/DS-FNC/-/-
Hunt and Jorgensen (2003)	MXD/MD-HND/ST-MDL/-/-
Blomstedt <i>et al.</i> (2015)	MXD/-/ST-MDL/-/OPT
Yao <i>et al.</i> (2019)	MXD/-/CN-DFN+NM-GRP/-/EN-THR
Velden, Enza and Markos (2019)	MXD/-/-/-/RVW
Bragato, Ganis and Feoli (2019)	MXD/-/-/-/CMP+PR-APP

A brief explanation about each document given in Table 11 follows. Huang (1998) proposed two modifications in the dissimilarity function of the k-means algorithm so that it could cope with categorical and mixed datasets. Hunt and Jorgensen (2003) extended the MULTIMIX algorithm in 2003, which was proposed on their own previously (see in Harch *et al.* (1999)), addressing missing data and changing the statistical model to make the algorithm able to cope with mixed attributes. Blomstedt *et al.* (2015) addressed the mixed data clustering problem, proposing a Bayesian predictive framework with an optimizer, called greedy stochastic search, to search optimal partitions. Yao *et al.* (2019) proposed a modified k-prototypes algorithm, inspired on the work of Liang *et al.* (2012), to determine the number of groups automatically using the entropy theory and define the centroids. Velden, Enza and Markos (2019) presented a review, bringing some distance-based clustering algorithms that handle mixed datasets and some dissimilarity measures used in this type of dataset. Bragato, Ganis and Feoli (2019) applied a complete linkage hierarchical clustering algorithm and compared two dissimilarities (Burnaby and Goodall) in the same clustering method to analyze the influence of each one in the clustering results.

3.4 Discussion

In this section, one presents some advantages in using the standard representation. One also brings opportunities to improve the viability of this incipient proposal, especially with regard to the definition of new acronyms and the searching of those which have already been proposed.

- Recognizing the main information: the possibility to quickly recognize the main information about a scientific document is the center of the idea of this standard representation proposal. In this case, the string represents not only the problem, but the main contributions a document did in this knowledge field. This kind of representation for this area is unique in the literature, to the best of our knowledge;

- Research databases and computer programs: the proposed standard representation allows the research databases and computer programs to automatically recognize the main information of a scientific document because the acronyms are going to be unique for each characteristic of the datasets, tasks, and auxiliary methods. Thus, finding relevant documents will be a more precise and quick activity, supporting the researchers in their researches and making the advance of the area faster;
- Review articles: with this standard representation, the authors of review articles are going to be able to make a table with summarized information about those documents they considered in their review, using the five main information of the proposed standard representation along with appropriate acronyms. As a result, the documents that will bring a state-of-the-art of an area can be done considering this new possibility (standard representation), providing for their reader and community a widely and summarized view of any topic they wish;
- Standardization of the representation: as the name expresses, the idea of standardizing the presentation of main information permits it to go beyond to its first objectives because the standardization opens doors for automation processes, one of which was pointed out in the first topic. By knowing this fact, this concern was addressed in this proposal.

As this proposal is the first attempt toward a standard representation in the knowledge-discovery field, there are some points to be improved and developed, especially with respect to the acronyms. This is because the standard representation only will work if there is a unique acronym for each characteristic of the dataset, tasks, and approaches. In addition, to guarantee the correct usage of the standard representation, the authors must have a unique and centralized platform to search the acronyms and to propose new ones. The proposals of new acronyms must be approved by the community, and then they must be inserted into that platform. With these cautions, the standard representation can be widely used in this knowledge field, as many other standard representations already are used in other areas.

3.5 Final remarks

In this chapter, one brought the first attempt of proposing a standard representation to characterize scientific documents in the knowledge-discovery-in-databases field, which we called five-fields notation. To the best of our knowledge, this kind of representation has not ever been proposed in the literature.

Its main objective is representing the five main information a scientific document has in this field, with the following components: i) characteristics of the datasets that were handled by the described document in order to show the complexity of the addressed

problem; ii) the tasks from the preprocessing step from KDD process the described document contributed; iii) the tasks from the possessing step the described document contributed; iv) the tasks from the post-possessing step the described document contributed, and v) the auxiliary methods the described document used to achieve its objective. All this information is expressed in a string composed of five parts separated with backslashes. Each part is built with a set of acronyms of the characteristics, tasks, and approaches the described document was inserted, which it contributed, and used, respectively. One showed examples of standard representations for the data clustering technique, though the standard representation idea to addresses all the data mining tasks, such as summarization, classification, and regression.

Although this proposal must be improved with the development of a centralized platform with the acronyms and with a defined process to search the acronyms that have already been proposed and propose new ones, its usage provides many advantages for the area and community. Among these benefits, there are i) the possibility to quickly recognize the main information about a scientific document; ii) the proposed standard representation allows the research databases and computer programs to automatically recognize the main information of a scientific document, and iii) with this standard representation, the authors of review articles are going to be able to make a table with summarized information about those documents they considered in their review, using the five main information of the proposed standard representation along with appropriate acronyms.

In this chapter, one presented the first contribution of this thesis, the standard representation string to characterize scientific documents on knowledge-discovery field. In the next chapter, it is presented the systematic literature review on mixed data clustering, in which the standard representation string is used.

Part III

Many views of the past

4 SYSTEMATIC LITERATURE REVIEW

“Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less”
Marie Curie

By the fact the mixed data clustering theme is not so explored as the single-type data clustering, one noted that, at the moment of carrying out such a review, there was no any systematic study on this thematic. Hence, the motivation of this chapter is to provide a state-of-the-art overview on mixed data clustering, in a widely and individually perspective, following a systematic protocol to select, analyze and describe the scientific documents. Taking this into consideration, the following paragraphs bring a brief description concerning the traditional reviews that have already been published on this topic, as well as the objectives of this chapter.

In 2018, two review articles were published about mixed data clustering. The work of Balaji and Lavanya (2018) aimed to perform a critical analysis of most effective algorithms regarding the mixed data clustering. Although the authors have described many scientific documents, they have only cited a few algorithms of each class of clustering. Miyamoto, Huynh and Fujiwara (2018) provided a brief presentation of some dissimilarity measures used for categorical data and briefly described a few algorithms, providing superficial details that did not explore deeply the theme.

In 2019, three more traditional reviews were published. Velden, D’Enza and Markos (2019) proposed a review on distance-based data clustering algorithms, specifically the partitional and hierarchical ones, addressing how dimensionality reduction is performed on mixed datasets and highlighting dissimilarity functions most commonly used. The authors also compared five mixed data clustering algorithms using three datasets. Although the article proposed a profitable discussion, the authors explored superficially the topics, not mentioning important documents in this area and merely presenting the main approaches, without conducting a consistent discussion and summarization. Foss, Markatou and Ray (2019) presented a review on distance measurements and clustering methods for mixed data, which was provided a theoretically and empirically analysis of some scientific contributions, covering the partitional and model-based algorithms. Ahmad and Khan (2019) brought an extensive and complete review of mixed data clustering, citing a wide range of documents on several classes of clustering algorithm. In addition, the paper also made a wide discussion about the challenges and the research issues in the mixed data clustering problem.

Although the five mentioned studies have revised the mixed data clustering technique, from the best of our knowledge, any systematic literature review on this topic

has not yet been published. A systematic review is concerned with answering specific questions about a subject, benefiting from a well-established method for finding and selecting documents that are aligned with the objective of the analysis, which contributes to the transparency of the research.

According to Kitchenham (2004), the advantage of a systematic review is to summarize the existing information about some phenomena in a complete and impartial way and explore the data using meta-analysis techniques, increasing the likelihood of detecting real effects that individual studies might be unable to detect.

In this chapter, one performs a systematic literature review on mixed data clustering to obtain an overview of this theme, as well as an in-depth analysis of the development of the algorithms, and point out the state-of-the-art of this area in its different classes of algorithms, discussing some research opportunities. Therefore, the main contributions of this chapter are:

- A systematic literature review on mixed data clustering linked with a rigorous and systematic protocol of selection, analysis, and discussion, with the following differentials:
 - A wide bibliometric analysis, addressing the sources, authors and documents metrics, in addition to the information regarding the conceptual, intellectual and social structure of the knowledge;
 - An in-depth individual analysis, which consists of detailing of the main characteristics of each document in a brief text and organized tables, which summarizes the following information:
 - * The origins, strategy, and objective of each document;
 - * The standard representation string;
 - * The datasets used in the experiments of the selected documents, the performance measures and the programming language the authors used to implement their algorithms, as well as whether they made the code of their proposals available for the community.
- A summarization of research challenges and opportunities.

This chapter is organized as follows. Section 2 describes a proposal to obtain a better organization of the results of the review, i.e. the branches of the data clustering algorithms. Section 3 presents the methodology used in this systematic literature review. Section 4 describes the results from the search in the databases. Section 5 and 6 exhibit the bibliometric analysis and the answers for a part of the questions of this systematic review, respectively. Section 7 and 8 bring the individual analysis and the answers for

the second part of the questions of this systematic review, respectively. Section 9 shows the research challenges and opportunities for this theme. And, Section 10 brings the final remarks.

4.1 Branches of data clustering algorithms and adjacent approaches

To have a better organization in the individual analysis section, one defines the branches of the mixed data clustering. Here, one considers these branches are composed of the most common five classes that have already been mentioned, in Chapter 2 and one more, which we call independent approaches to the pattern recognition algorithm (IAPRA). The IAPRA branch consists of quantitative or theoretical contributions that surround the main classes of data clustering algorithms, which many of them are regarding the preprocessing and post-processing activities from the KDD process (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a). We believe this sixth branch is important for the analysis of the results because, although it is not directly related to the mixed data clustering algorithms, it might influence their performance and/or the comprehension of the thematic direction. Figure 4 brings what one considers as the branches of the mixed data clustering.

Figure 4 – Branches of mixed data clustering

To provide a better explanation regarding the last two branches (others and IAPRA), Figure 5 shows some examples of approaches that are part of each one. In the following, one briefly explains each method given in Figure 5, which were chosen inspired by the work of Ahmad and Khan (2019).

Figure 5 – Some examples of what is addressed in the branches *others* and IAPRA

Spectral clustering algorithms (NG; JORDAN; WEISS, 2001) use eigenvectors of matrices derived from the data to cluster a dataset. From an affinity matrix previously calculated, it is found the k the largest eigenvectors, which is transformed into a new normalized dataset. This last dataset is used by a data clustering algorithm to find the final clusters. Clustering ensemble methods (ALQURASHI; WANG, 2019), also called consensus clustering, combine several clustering algorithms to provide a better solution than would be delivered by only one data clustering algorithm. Conceptual clustering algorithms (MICHALSKI; STEPP, 1983) provide a description, a concept, of the groups found by the clustering algorithm to facilitate the interpretation of the clusters by the user. Semi-supervised algorithms (CHAPELLE; SCHLKOPF; ZIEN, 2010) combines labeled and unlabeled data, in different proportions, to achieve better performance without the cost of labeling all observations. Density-based clustering methods (ESTER *et al.*, 1996) consider a cluster is a density region in the data space. Co-clustering methods (GOVAERT; NADIF, 2013); also known as bi-clustering or block clustering; simultaneously carry out the clustering of the objects and attributes of a dataset. Graph-based clustering algorithms (FOGGIA *et al.*, 2007) use the graph theory and consider the objects are nodes of a graph and the distance between two objects is the weights of the edges. These algorithms find a cluster when identifying a set of connected components.

Regarding IAPRA approaches, most of them are preprocessing and post-processing activities, which are really independent of the pattern recognition algorithm. In this chapter, one considers that a scientific work belongs to the IAPRA branch when the main proposal is indeed an independent approach of a pattern recognition algorithm, and not those documents that, although mentioned and proposed something with respect to preprocessing or post-processing activities, also made some proposal in one of the remaining five branches of the data clustering algorithms.

4.2 Methodology

One of the main differences between systematic reviews and traditional ones is that the former formally establishes a protocol, i.e. a plan to conduct the search, selection, and discussion of the scientific documents. Thus, one defined the protocol with nine steps based on the steps suggested by Kitchenham (2004). Figure 6 shows them all sequentially, from the left to right.

The first step consists of defining of the scientific databases in which the search of the mixed data clustering documents is conducted. The second activity refers to conducting an earlier (non-systematic) study of the topic to obtain familiarity with the notations and terminologies. In the third step, it is defined the main keywords and the way they need to be connected in the search string. The fourth task is related to the definition of the questions of this review. Note that the second, third, and fourth tasks occur simultaneously

Figure 6 – Steps of the protocol to conduct the systematic literature review

until the search string is fully defined as well as the questions. Next, the selection criteria and the strategy to explore the documents are determined. With these in hand, in the seventh stage, it is defined how the synthesis of the documents is done. Finally, this protocol includes the answers to the questions defined in the third step and a discussion of all considered documents.

4.2.1 Steps 1 and 2

Regarding the scientific databases, one elected the Scopus and Web of Science, because both contain the most important journals in computer science, statistics, engineering, and mathematics, which are the main areas the data clustering technique comes from. With respect to the non-systematic study, one analyzed some well-cited documents that returned in both mentioned databases. From these documents, one could obtain more knowledge about other recognized documents, and thereafter establishing the main keywords as well as the search string.

4.2.2 Step 4

As a result, one defined the following search string *((cluster* AND ("mixed data*" OR "mixed type" OR "mixed" OR "multiple attribute*" OR "mixed feature*" OR "mixed variable*" OR "numeric* and categoric*" OR "categoric* and numeric*" OR "numeric* and non-numeric*" OR "non-numeric* and numeric*" OR "internal and categoric*" OR "categoric* and internal"))*). The search, which was carried out in August 2019, considered the title, keywords, and abstract of documents that were indexed in both databases.

4.2.3 Step 3

The step 3 consists of defining the questions the systematic review will answer. These questions are defined into blocks of information, following the objectives announced in the beginning of this chapter, i.e. the bibliographic analysis and the individual analysis, as shown in Figure 7.

Figure 7 – Main deliveries of the systematic review: bibliometric and individual analysis

Concerning the bibliometric analysis (B), one follows the structure cited in Aria and Cuccurullo (2017), which is composed of information metrics about sources (ST), authors (AT) and documents (DC), besides the structures of knowledge, which are the conceptual (CN), intellectual (IN) and social (SC). Taking this into consideration, Table 12 brings the questions with respect to the bibliometric analysis.

Table 12 – Questions - Bibliometric analysis

Code	Question
B-SR-1	What are the most relevant sources?
B-SR-2	How is the growth of the sources over the years?
B-AT-1	Who are the most relevant authors?
B-AT-2	What are the most relevant countries?
B-AT-3	What is the proportion of documents produced in international partnerships in each country?
B-AT-4	How many documents on average does an author publish?
B-AT-5	What is the frequency of publications by the most relevant authors?
B-DC-1	What are the most cited documents?
B-DC-2	What are the most cited references?
B-DC-3	How is the behavior of the size of the references' list belonging to the analyzed documents over the years?
B-DC-4	How are the documents related over the years?
B-DC-5	What are the most frequent words?
B-CN-1	What do concepts appear together most often?
B-CN-2	What are the highly developed/isolated, motor, emerging/decline and basic themes?
B-IN-1	How are the references cited in the documents related?
B-SC-1	Which do authors and countries work together?

With respect to the individual analysis (I), the questions follow three blocks of information, which are: (i) basic description, (ii) standard representation, and (iii) datasets, validation measures and programming languages.

The first block is composed of the following information: i) the approaches that inspired (IN) the described documents to propose their contributions; ii) the strategy (ST)

the described documents chose to cope with mixed datasets; iii) the type of objective (OB) the described documents declared; iv) the publication quantity (QN); v) the publication constancy (CS), and vi) the diversity (DV) of authors. With respect to the objectives, one considers the following scenarios: i) the described document proposes a pure method on some theoretical issue about mixed data clustering algorithms or related themes (pure proposal - PP); ii) the described document proposes an improvement based on, inspired by, and/or restricted to a solution of a specific problem (proposal inspired on application - PIA), and iii) the described document proposes a pure application of some existed method (pure application - PA). If the described document had as objective the last scenario, then one also informs the theme of the application/problem (application theme - AT), such as health, education, and finances. Table 13 shows the questions with respect to this block.

Table 13 – Questions - Individual analysis, Part 1 - Basic description

Code	Question
I-IN-1	On what algorithms, methods, and/or concepts do the documents base its proposals or applications?
I-ST-1	What is the most used type of strategy to deal with mixed data?
I-OB-1	What is the proportion of pure-proposal documents?
I-OB-2	What is the proportion of pure-application documents?
I-OB-3	What is the proportion of proposal-inspired-on-application documents?
I-OB-4	What are the most addressed themes by pure-application and proposal inspired on application documents?
I-QN-1	Which branches do have the most and fewer documents?
I-CS-1	How is the publication constancy in each branch over time?
I-DV-1	Are there diversity among authors who contributed to each branch?

With respect to the second block of information, the standard representation (SR), the five components are filled, when appropriate, for each of document resulted from the systematic literature review. Table 14 shows the questions with respect to this block.

Table 14 – Questions - Individual analysis, Part 2 - Standard representation

Code	Question
I-SR-1	Which are the main characteristics of datasets considered in the documents?
I-SR-2	Which are the preprocessing tasks that received the most contributions?
I-SR-3	Which are the processing tasks that received the most contributions?
I-SR-4	Which are the post-processing tasks that received the most contributions?
I-SR-5	Which are the auxiliary approaches most used?

Concerning the third block of information from the individual analysis, one considers the following topics i) Datasets (DTS): simulated and real-world datasets, ii) Validation

measures (VM): internal, external and relative measures, and iii) programming language and code availability (PLC): the programming language and code availability. Table 15 shows the questions with respect to this block.

Table 15 – Questions - Individual analysis, Part 3 - Dataset, validation measures and programming language

Code	Question
I-DTS-1	Do the documents more use simulated or real-world datasets?
I-DTS-2	What are the sources of the simulated and real-world datasets?
I-DTS-3	What are the most used simulated and real-world datasets?
I-VM-1	Do the documents more use internal, external, or relative validation measures?
I-VM-2	What are the most used validation measures?
I-PLC-1	Is it common the documents inform the programming language that was used to carry out the experiments?
I-PLC-2	What is the most used programming language in this area?
I-PLC-3	Is it common for the documents to make the proposal code available to the community?
I-PLC-4	Where are most of these codes available?

4.2.4 Steps 5 and 6

With respect to the selection step, one established the following criteria: i) it was considered only documents written in English, published in high-impact scientific vehicles, specifically with journal citation report (JCR) greater than one, or with at least two indicators greater than one, considering the following metrics, the JCR, citescore, SCImago Journal Rank, and Source Normalized Impact per Paper. All the impact metrics were extracted from the report of 2018^{1,2}.

With respect to the sixth step, strategy for exploring the documents, one decided a document is analyzed in two reading depths. In the first, one reads only the title, abstract, and keywords. If, with this information in hand, it is not possible to decide whether the document is aligned with this review, then the second type of reading is performed, which consists of reading the entire article, seeking for more information to support the decision to include or not the article in the review.

4.2.5 Steps 7, 8 and 9

In step 7, one defined that the bibliometric and individual analyses will be explored considering the blocks shown in Figure 7. Regarding, the bibliometric analysis, the metrics block focuses on providing information about the article meta-data, such as the most relevant sources and authors and their dynamics throughout the time, H-index, annual

¹ <https://www.scopus.com/sources.uri>

² <https://jcr.clarivate.com/JCRJournalHomeAction.action>

production per author, and the most cited documents. The knowledge block provided a synthesis regarding the scientific knowledge throughout the years, attempting to draw the main themes and trends (conceptual structure), the influence of some documents on the scientific community (intellectual structure), and the interaction of the scientific communities agents, such as authors, institutions and countries (social structure) (ARIA; CUCCURULLO, 2017). Concerning the individual analysis, one discussed the topics from the first block, each component of the standard representation string, and following information:

1. *Datasets*: it reported whether the authors have used simulated and/or real-world datasets in their computational experiments; if yes, one provided the address of the datasets;
2. *Performance measures*: it provided the name of each internal, external and relative measures that the authors used to evaluate the algorithms;
3. *Programming language and code availability*: it informed the programming language the authors chosen to implement their algorithms and whether they made the code available; if yes, one provides the programming language, the name of the code, and where it is.

After analyzing each document, one accomplished a wide discussion and summarization (step 9) about the provided information and answer the questions of the systematic review (step 8). To finalize, one pointed out some open research topics on mixed data clustering algorithms (step 9).

After analyzing the blocks, one accomplishes a wide discussion and summarization (step 9) about the provided information answering of the questions previously defined (step 8).

4.3 Search results

Table 16 summarizes the number of documents that returned from the databases and remained throughout the review after applying the selection criteria and reading filters.

Table 16 – Number of documents throughout the selection step

Databases	Total	in English	Except CR	Duplicated	Exclusive	Impact metrics	Reading filter	Total
Scopus	1,104	1,010	979	-	979	458	161	157
Web of Science	731	723	723	633	90	27	3	3
Total	1,835	1,733	1,702	633	1,069	485	164	160

As can be seen in Table 16, a total of 1,835 documents returned as a result of the research carried out in the two databases. When one restricted them to documents written in English, it was obtained 1,733 articles. In the Scopus database, it was necessary to remove documents whose source's name was conference review (CR), because this type of document is not an article, but a summary of a conference. Information about this removal is in the third column of Table 16 (Except CR). The Web of Science database did not contain this type of document. As a result of this removal, 1,702 papers remained in the selection process. Of these articles, 633 (723-90) were duplicated. Therefore, one obtained 1,069 exclusive documents, of which 458 were published in high-impact scientific vehicles, according to the rules previously defined. In the reading filter, 164 scientific papers were selected. One did not have access to four articles (HOWARD; HOWARD, 1988; MOISE; SANDER; ESTER, 2008; QIN *et al.*, 2013; PATHAK; PAL, 2016), resulting in a selection of 160 documents for analysis.

Table 17 brings the number of documents in each branch and Table 18 shows the co-occurrence of the documents in the branches. Note the name of the branches in both tables are expressed by its first letter. It is important to observe that the total the sum of the document in both tables is 165, which represents the 160 documents selected from the systematic procedures plus 5 documents that were double-classified, as can be seen in Table 18. For example, the work of Hennig and Liao (2013), which belongs to the partitional branch, has in front of the paper citation (first column of the tables of partitional branch presented in Appendix A) the following information: [M], which means this document proposed a partitional and model-based approaches in the same document, justifying that this work is also in the tables of the model-based branch.

Table 17 – Number of documents in each branch

	P	H	M	N	O	I
Scopus	57	17	29	10	28	20
Web of Science	2	0	0	0	1	1
Total	59	17	29	10	29	21

Table 18 – Co-occurrence of documents in the branches

	P	H	M	N	O	I
P	56	2	1	0	0	0
H	2	14	1	0	0	0
M	1	1	26	0	1	0
N	0	0	0	10	0	0
O	0	0	1	0	28	0
I	0	0	0	0	0	21

4.3.1 Document collection

In relation to the 160 documents, Table 19 and Figure 8 show the description of the document collection and the number of documents published in each year, respectively. It was used the Biblioshiny for Bibliometrix, which is an R-tool from the R package called bibliometrix (ARIA; CUCCURULLO, 2017), to generate the mentioned information.

Table 19 – Document collection - Description

Description	Results
Documents	160
Sources (Journals, Books, etc.)	94
Keywords Plus (ID)	1026
Author's Keywords (DE)	457
Period	1983 - 2020
Average citations per documents	26.49
Authors	405
Author Appearances	486
Authors of single-authored documents	12
Authors of multi-authored documents	393
Single-authored documents	12
Documents per Author	0.395
Authors per Document	2.53
Co-Authors per Documents	3.04
Collaboration Index	2.66
Document types:	
Article	149
Conference paper	10
Review	1

Figure 8 – Document collection - Annual scientific production

4.4 Bibliometric analysis

In this section, one details the bibliometric analysis, describing its six main topics, as shown in Figure 7 and explained in Section 4.2. All the figures and tables presented in this section were generated by the bibliometrix package (R language) (ARIA; CUCCURULLO, 2017). The explanation of the figures and tables was supported by the Biblioshiny tutorial³.

4.4.1 Sources

Figure 9 shows the ten most relevant sources, i.e. the sources that published the largest number of documents regarding mixed data clustering, and Figure 10 shows the ten most cited sources, i.e. the sources that received one or more citations considering the document collection (ARIA; CUCCURULLO, 2017).

Figure 9 – Sources analysis - Most relevant sources

Besides, Figure 11 brings the growth of the six most relevant sources throughout the years, considering the number of documents published in each year by each source.

It is possible to note that the journal *Expert system with application* appears in the first two positions in both figures as the most relevant and cited journal. It is also noted that, although the number of publications per year in that journal grew until mid-2011 (Figure 11), it declined after that year until 2019. The source that mostly grew, regarding the number of publications per year, is the journal named *Advances in data analysis and classification*. This source had an exponential growth in 2006, reaching the first position in 2019.

³ <https://bibliometrix.org/biblioshiny/>

Figure 10 – Sources analysis - Most cited sources

Figure 11 – Sources analysis - Source growth

Table 20 shows the sources that belong to zone 1 of the Bradford's law (ALABI, 1979), which contains the sources responsible for 55% of publications with respect to mixed data clustering.

4.4.2 Authors

Figure 12 shows the most relevant authors regarding the number of published documents and Figure 13 shows the first ten authors who have the highest impact,

Table 20 – Sources analysis - Sources in zone 1 of the Bradford's law

Source	Rank	Frequency
Expert systems with applications	1	12
Advances in data analysis and classification	2	5
Applied soft computing journal	3	5
Journal of classification	4	5
Knowledge-based systems	5	5
Mathematical problems in engineering	6	5
Computational statistics and data analysis	7	4
Information sciences	8	4
Pattern recognition	9	4
IEEE Access	10	3
IEEE Transactions on knowledge and data engineering	11	3

considering the H-index. Both figures bring the same authors in the first five positions, showing the authors who have more published documents are also the same that have the highest H-index.

Figure 12 – Authors analysis - Most relevant authors

Figure 14 shows a timeline of the production of the ten most productive authors (number of published documents). In this figure, the red line shows the period (years) in which an author had document published, where the presence of blue balls represents the specific years in which a document was published and the size of the blue balls is proportional to the number of published documents. Therefore, it is possible to note that Basford K.E. is the author who has been contributed to this theme for the longest time, although Hsu C. C. is the author who published more documents in the largest period of years.

Figure 13 – Authors analysis - H-index of the authors

Figure 15 shows the productivity of the authors through the Lotkas's law (LOTKA, 1926), where the dotted line shows the theoretical frequency distribution and the continuous line the real distribution. In this figure, it is possible to notice that more than 75% of the authors published only one document on the studied thematic and approximately 10% of them have, at least, 2 documents published on mixed data clustering.

Figure 14 – Authors analysis - Production of the authors over the time

Figure 16 brings the most cited countries the authors are related. Australia is in the first position because the work of Huang (1998) is the most cited article among the 160 documents, as can be seen in Section 4.4.3.

Figure 15 – Authors analysis - Productivity of the authors through Lotka's law

Figure 17 shows the collaboration among the countries to publish a document in this theme, where the red part shows the international collaboration, i.e. the number of published documents which had at least one author from a different country, and the blue part in regard to the documents produced by authors from the same country.

Figure 16 – Authors analysis - Most cited countries

Figure 18 shows the map of the number of documents per country, where the color intensity is proportional to the number of publications (based on the affiliation of the first author). Table 21 brings the values of Figure 18 for the first ten countries which have more publication and its frequency values.

Figure 17 – Authors analysis - Corresponding author's country, where red means multiple countries publications (MCP) and blue is regard to single country publication (SCP)

Figure 18 – Authors analysis - Scientific production by country

4.4.3 Document

Figure 19 shows the most global cited documents with respect to the number of citations they received from documents contained in the entire database (Scopus and Web of Science) (ARIA; CUCCURULLO, 2017). Figure 20 shows the most local cited documents, i.e. the number of citations the documents received from the document collection (160 documents resulted from this systematic review) (ARIA; CUCCURULLO, 2017). It is possible to note that Huang (1998) and Ahmad and Dey (2007) are in the first positions in both figures.

Table 21 – Authors analysis - Scientific production by country

Country	Frequency
China	56
United States of America	42
India	27
United Kingdom	22
Taiwan	21
Australia	19
France	16
Italy	15
Canada	14
Spain	11

Figure 19 – Documents analysis - Most global cited documents

Figure 21 shows the first ten most cited references included on, at least, one of the reference lists from the document collection. Note that the cited documents (Figure 19 and 20) are a subset of the reference set (ARIA; CUCCURULLO, 2017).

Figure 22 brings the publication year spectroscopy of the references, i.e. the result from a quantitative method that identifying the historical origins of a specific theme. In this figure, it is possible to note the references included in the references lists from the document collection are of the beginning of the 20s century, and most of the documents included in these lists began to be published in the mid-20s century. Currently, the quantity of documents that compose the reference lists of the documents that address the mixed data clustering issue is increasing, evidencing this is an emerging theme in the literature.

Figure 23 shows the ten most frequent *authors' keywords* that appeared in the 160 documents. Figure 24 shows the word cloud of the 20 most frequent *keywords plus* from

Figure 20 – Documents analysis - Most local cited documents

Figure 21 – Documents analysis - Most cited references

the document collection. Note that *authors' keywords* are the words (or a set of words) that the authors chose believing to be the best words to represent the content of their documents and *keywords plus* are words (or a set of words) that have been generated automatically by a computer algorithm to show those that frequently appear in the title of the references of the analyzed articles (GARFIELD, 1990a; GARFIELD, 1990b). Garfield and Sher (1993) explained that it is possible to understand an article's content with greater depth and variety analyzing the *keywords plus* than the *authors' keywords*.

Figure 22 – Documents analysis - Reference publication year spectroscopy, where the red line means the deviation from the 5-year median and the blue line the number of cited reference per year

Figure 23 – Documents analysis - Most frequent author's keywords

4.4.4 Conceptual structure

The conceptual structure is able to capture the relations among the concepts presented in the document collection. Figure 25 shows the co-occurrences network of the authors' keywords. The co-occurrences network shows the number of times a specific keyword appeared in the documents (size of each rectangular word box) and the number of

Figure 24 – Documents analysis - Most frequent keywords plus - Word cloud

co-occurrences of two keywords that appeared simultaneously in the document collection (the thickness of the line that links two word boxes). To generate this figure, one considered the association as the normalization procedure, the walktrap clustering algorithm, 24 nodes, 30 labels, and the minimum of 2 edges.

Figure 26 shows the thematic map (COBO *et al.*, 2011) of the document collection. By clustering the authors' keywords network, the thematic map provides a visualization regarding different themes of a given domain. Following the explication given in the Bibliometrix tool⁴, the map has four regions, separated by two axes, where the horizontal axis represents the centrality, in which is possible to verify the importance of the theme in the entire research field, and the vertical axis represents the density, in which is possible to analyze the development of the theme. The four regions are i) high developed and isolated themes (high density and low centrality); ii) motor themes (high density and high centrality); iii) basic and transversal themes (low density and high centrality), and iv) emerging or declining themes (low density and low centrality). The bubbles on the map represent the groups, whose names are the words that mostly appeared in that specific group and their size is proportional to the occurrences of words in that group.

4.4.5 Intellectual structure

The intellectual structure permits analyzing the relationship among references that were cited in the document collection. Figure 27 shows the co-citation network, where each node represents a reference, the node size is proportional to the number of citations that reference received from the document collection, the link between two nodes shows that both references were cited by a third document belonged to the document collection, and the thickness of the link is proportional to the number of citations both references received from a third document belonged to the document collection. To generate this

⁴ <https://bibliometrix.org/biblioshiny/>

Figure 25 – Conceptual analysis - Co-occurrences network of the authors' keywords

Figure 26 – Conceptual analysis - Thematic map regarding the authors' keywords

figure, one considered the following configuration: papers as the field, walktrap as the

clustering algorithm, 30 nodes, and the minimum of 4 edges.

Figure 28 shows a historiography mapping, where each node represents a reference, each link shows a direct citation (e.g. Cheung and Jia (2013) cited by He, Xu and Deng (2005); Chatzis (2011) cited by Huang (1998)), nodes and edges are plotted on an oriented graph, and the horizontal axis represents the years. In this figure, it is clear the path of the development of this theme, beginning with the work of Huang (1998) and advancing throughout a few other influential documents, such as Ahmad and Dey (2007), He, Xu and Deng (2005) and Hsu, Jao and Chen (2005). To generate this figure, one considered 10 nodes and fast search as the search algorithm.

Figure 27 – Intellectual analysis - Co-citation network

Figure 28 – Intellectual analysis - Historiography mapping

4.4.6 Social structure

The social structure shows the relationship of the authors and institutions from the document collection (PETERS; RAAN, 1991; ARIA; CUCCURULLO, 2017). Figure 29 shows the co-authorship network, where each node represents an author, the node size is proportional to the number of documents that author published among the document collection, the link between two nodes shows that both authors wrote a document together, and the thickness of this link is proportional to the number of documents both authors wrote. To generate this figure, one considered the following configuration: authors as the field, walktrap as the clustering algorithm, 24 nodes, and the minimum of 2 edges.

Figure 30 shows the co-country network, where each node represents a country, the node size is proportional to the number of documents the countries published among the document collection, the link between two nodes shows that both countries were related in a published documents among the document collection, and the thickness of this link is proportional to the number of documents both countries appear. To generate this figure, one considered the following configuration: countries as the field, walktrap clustering algorithm, 24 nodes, and the minimum of 2 edges.

Taking all information showed in the previous topics into consideration, the following section answers the questions defined in Section 4.2 for the bibliometric analysis.

4.4.7 Answers of the questions defined for the bibliometric analysis

The following items bring the answers to each question defined in Table 12.

- **Question B-SR-1** - What are the most relevant sources? Figures 9 and 10 bring the most relevant and cited sources, respectively. With respect to the most relevant sources, the first three are the following journals: *Expert systems with application*, *Advances in data analysis and classification*, and *Applied soft computing journal*. Regarding the most cited sources, one has in the first three positions the following journals: *Biometrics*, *Pattern recognition letters*, and *Expert systems with applications*.
- **Question B-SR-2** - How is the growth of the sources over the years? Figure 11 brings the number of publications of the first six relevant sources that received documents about mixed data clustering over the years. This figure shows that the *Advances in data analysis and classification* and *Journal of classification* are the two sources that have been growing with respect to the number of publications over time. It is also possible to note the other four sources have been declining since mid-2009 and mid-2015.
- **Question B-AT-1** - Who are the most relevant authors? Figures 12 and 13 show the most relevant authors and the authors who have the highest H-index, respectively.

Figure 29 – Social analysis - Co-authorship network

Note the first six positions of both figures are composed of the same authors, which are, in this order: C.-C. Hsu, A. Ahmad, K. E. Basford, J. Ji, Y. M. Cheng, and S. S. Leu.

- **Question B-AT-2** - What are the most relevant countries? Figures 16 and 18, in addition to Table 21, bring the most cited countries, the scientific production per country through a visual map, and the scientific production per country through a table of frequency. One observes that Australia is the most cited country followed by the USA, China, India, Taiwan, United Kingdom, Canada, New Zealand, Hong Kong, and Portugal. With respect to the number of publications, China is in the first position, followed by the USA, India, United Kingdom, Taiwan, Australia, France, Italy, Canada, and Spain.
- **Question B-AT-3** - What is the proportion of documents produced in international partnerships in each country? Figure 17 shows the number of publications per country, where orange bar shows the international partnerships and, the blue bar shows the documents written by authors from a single country. Canada is the country that

Figure 30 – Social analysis - Co-country network

has the highest proportion of international partnerships, followed by China, United Kingdom, USA, and Italy. The remaining countries (Taiwan, Australia, India, Spain, and Germany) do not establish international partnerships to publish their documents in mixed data clustering.

- **Question B-AT-4** - How many documents, on average, does an author publish? Table 19 brings some information regarding the document collection, such as the number of the document per author, which is 0.395. Taking this into consideration and Figure 15, which shows the productivity of the authors through the Lotka's law, one can state that, most authors from the document collection publish a single document, and approximately 10% of them have, at least, 2 documents published on mixed data clustering.
- **Question B-AT-5** - What is the frequency of publications by the most relevant authors? Figure 14 exhibits, in a timeline, the behavior of the first ten most relevant authors, with respect to the number of publications. One observes that, although C.-C. Hsu is the most productive author, K. E. Basford is the author who has contributed to this thematic for the longest time. There are other authors, such as Y. M. Cheng and S. S. Leu who made pointed contributions in 2008 and 2010,

respectively, and do not publish any new documents until that moment. Finally, the remaining authors have begun to contribute to the mixed data clustering problem recently.

- **Question B-DC-1** - What are the most cited documents? Figure 19 and Figure 20 show the most cited documents, globally and locally, respectively. The seminal document in this area, which proposed the k-modes and k-prototypes algorithms (HUANG, 1998), is in the first position in both figures, and the work of Ahmad and Dey (2007). The remaining papers have different positions in both figures, which shows they are cited by documents from another thematic or by documents from the same thematic of which did not match the selection criteria proposed in this systematic review.
- **Question B-DC-2** - What are the most cited references? Figure 21 shows the first ten most cited references by the document collection. One notes that Huang (1998) belongs to the first position of the list, followed by Ji *et al.* (2013). The remaining documents do not belong to this document collection. Most of these references are seminal proposals, which continue to be highly explored by current scientific works.
- **Question B-DC-3** - How is the behavior of the size of the reference lists belonging to the document collection over the years? Figure 22 brings the spectroscopy of the references, year by year, which shows the reference lists from the document collection are increasing over the years. This figure also exhibits the first documents that were cited by the seminal works in this thematic. One observes the documents that are published nowadays have many more papers to base their contributions (references), showing the growth of this theme.
- **Question B-DC-4** - How are the references related over the years? Figure 28 shows a historiography mapping, which brings the relation among the first ten most relevant references. The number of documents in this figure, and in the others, has restricted for the sake of clarity.
- **Question B-DC-5** - What are the most frequent words? Figure 23 and Figure 24 bring the frequency of the authors' keywords and the keywords plus, respectively. One notes the basics words of this theme appear with the highest frequency, such as clustering, mixed data, and data mining, as it was expected. However, one is interested in the more specific words (or set of words) that appeared frequently. In both figures, one can observe the following words: the number of clusters, real-world datasets, k-means, mixture model, and fuzzy clustering. These words show that the concepts connected to them were the ones that most aroused the interest of the authors who contributed to this theme.

- **Question B-CN-1** - What do concepts appear together most often? Figure 25 shows the co-occurrences of the authors' keywords. It is possible to note three groups of words (blue, red, and green). The blue group, the biggest one, is composed of basics words of this thematic (most of them were chosen to compose our search string), such as *clustering*, *data mining*, *mixed data*, *mixed attributes*, *unsupervised learning*, and *mixed-type data* linked to more specific words, like *number of clusters*, *self-organized maps*, *entropy*, *distance hierarchy*, *k-means*, *big data*, *dissimilarity measure*, *k-prototypess*, and *fuzzy clustering*. Some specific words are seminal approaches (k-means, k-prototypes, and self-organized map), processing tasks (number of clusters, dissimilarity measure, and fuzzy clustering), and others are concepts or datasets characteristics (entropy, distance hierarchy, big data). The occurrence of the words like self-organized maps and distance hierarchy may be justified by the many documents published by C.-C. Hsu in the neural network-based branch (see in Appendix B), corresponding 70% of the documents of that branch. Concerning the red group, it is composed of more specific words (feature selection, distance-based clustering), which shows the distance-based algorithms have been exploring more intensely. Finally, the green group shows one single connection between the words *variable selection* and *model-based clustering*.
- **Question B-CN-2** - What are the highly developed/isolated, motor, emerging/decline, and basic themes? Figure 26 shows the thematic map with respect to authors' keywords. Although many groups are described by basics words, such as *clustering algorithms*, *article*, and *mixed type*, it is possible to analyze the others that are described by more specific words. One notes that the green cluster, which is composed of the words *neural networks* and *distance hierarchy*, is in the area that concentrates the *high developed and isolated themes (niches)*. It was understood this group represents the papers from the neural network-based branch, not only because the first word but also because the second descriptor. From our point of view, these proposals were classified in this quadrant, because there are only ten documents in this branch (isolated), written by a few authors, who, in general, apply the distance hierarchy concept in their algorithms (high developed). With respect to *emerging or declining themes*, one notes the blue cluster, which is described by the words *fuzzy clustering*, *optimization*, and *fuzzy systems*. In our analysis, one understands that such words represent emerging themes, once that there are not too many documents addressing these approaches among the document collection. Regarding *motor themes*, both clusters (brown and pink) represent the themes that are in evidence in the document collection, among them are those that express concern about big data, fuzzy sets, and that use the k-prototypes algorithm. Finally, the *basic themes* is represented by two clusters (red and orange), which are composed of basics words that do not express a revelatory concept, although it has already been expected because the

search string induced it.

- **Question B-IN-1** - How are the references cited in the documents related? Figure 27 exhibits the co-citation network. It is possible to note that there are two groups of references (blue and red) in which the first has the work of Huang (1998) as the reference which received more citation and is related with more other references, and the second has the work of Gower (1971) as that was more cited and related. One observes that most of the references are seminal and traditional documents, although there are some of them that are recent. In our point of view, these recent papers are those that receive more attention by the community, therefore important articles to be considered in future contributions.
- **Question B-SC-1** - Which did the authors and countries work together? Figure 29 and Figure 30 show the co-authorship and the co-country networks, respectively. One observes six groups of the most common partnerships established among authors (blue, red, purple, orange, green, and brown). It is possible to note a group composed of *China, United Kingdom, USA, Netherlands, Canada, Poland, Australia, New Zealand, and Hong Kong* that has countries that wrote, at least, two documents together. There are also countries not linked, which did not establish a partnership in this theme that resulted in more than 2 papers. The partnership produces citations for the countries that establish it, as one can note in Figure 16, excepting for the case of India, Taiwan, and Portugal, which are on the list of the top ten most cited countries, but wrote articles without international partnerships.

In this section, it was possible to conclude the bibliometric analysis, achieving

4.5 Individual analysis - Part 1

In this section, one details the individual analysis, providing the information about each document in each branch. In Appendix A, one brings the documents from partitioned to IAPRA branch (Tables from 37 to 42). The tables inform the following information in each column: i) the described document; ii) the names of the approaches proposed by the described document; iii) the main algorithms, methods and/or concepts the document based its contributions, the strategy used by the described work to cope with mixed data (1: data transformation and 2: a specific function for each type of attribute), and iv) type of objective the described document announced.

4.5.1 Answers of the questions defined for the individual analysis - Part 1

The following items bring the answers to each question of the individual analysis, part 1, defined in Section 4.2.3.

- **Question I-IN-1** - On what algorithms, methods, and/or concepts do the documents base its proposals or applications? Many documents based their proposals in seminal approaches, such as k-means and k-prototypes algorithm, mixture statistical models, and self-organized maps. Many of them also used the concept of distance hierarchy to propose their contributions, e.g. Tai, Hsu and Chen (2010), Hsu, Lin and Tai (2011). With respect to the branch *others*, many authors addressed the graph-based algorithms to solve the mixed data clustering problem, which is the case of Zhao *et al.* (2013).
- **Question I-ST-1** - What is the most used type of strategy to deal with mixed data? Strategy number two, which delegates specific functions for each type of attribute, is the most used by the authors, independently of the branch. The readership can find an extensive material regarding distance-based methods for mixed data clustering in Foss, Markatou and Ray (2019).
- **Question I-OB-1** - What is the proportion of pure-proposal documents? Generally, the pure-proposal documents are the majority, representing 60% of the documents analyzed in this review. However, this proportion is different when the branches are analyzed individually. In the following, it is presented the proportion of pure-proposal documents in each branch, partitional (61%), hierarchical (18%), model-based (62%), neural network-based (100%), others (66%), and IAPRA (62%).
- **Question I-OB-2** - What is the proportion of pure-application documents? Generally, the pure-application documents are not the majority, representing 17% of the documents analyzed in this review. However, this proportion is different when the branches are analyzed individually. In the following, it is presented the proportion of pure-application documents in each branch, partitional (4%) , hierarchical (65%), model-based (17%), neural network-based (0%), others (0%), and IAPRA (29%)
- **Question I-OB-3** - What is the proportion of proposal-inspired-on-application documents? Generally, the proposal-inspired-on-application documents are not the majority, representing 23% of the documents analyzed in this review. However, this proportion is different when the branches are analyzed individually. In the following, it is presented the proportion of pure-application documents in each branch, partitional (10%) , hierarchical (18%), model-based (21%), neural network-based (0%), others (34%), and IAPRA (10%)
- **Question I-OB-4** - What are the most addressed themes by pure-application and proposal-inspired-on-application documents? Generally, the following themes were highly addressed, which are health, social, marketing, theoretical, agriculture, and industry. In each branch the most addressed themes were, partitional (health),

hierarchical (social, marketing and agriculture, with the same proportion), model-based (health), neural network-based (-), others (health), and IAPRA (theoretical).

- **Question I-QN-1** - Which branches do have the most and fewer documents? Undoubtedly, the partitional branch has the highest number of documents (59), in opposition of the neural network-based branch, which has the fewest number of documents (10).
- **Question I-CS-1** - How is the publication constancy in each branch over time? Generally, the publications in this thematic have been growing throughout the years, especially in the partitional and model-based branches. However, this is not the scenario for some branches, such as hierarchical and neural network-based, which have their frequency of publication in the same values since they began.
- **Question I-DV-1** - Is there diversity among authors who contributed to each branch? In the partitional and model-based branches, there are many authors who contribute for the publications. However, in the neural network-based branch, predominates documents wrote by Hsu, C.-C.

4.6 Individual analysis - Part 2

In the following subtopics, one provides a brief description of each document analyzed in this review, as well as a discussion about the information from the filled standard representation string, which are the rows of Tables from 43 to 48) shown in the Appendix B, where one also brought the meaning of the acronyms used to build the mentioned representation.

4.6.1 Partitional

The first work in this branch that was in line with the objectives of this systematic review is the paper of Huang (1998) in which the author proposed two modifications in the k-means algorithm so that it could work with categorical data and mixed data, resulting in the k-modes and k-prototypes algorithms, respectively. Ng and Wong (2002) proposed a fuzzy version of k-modes and k-prototypes algorithms, and applied the tabu search to define the centroids of the clusters. He, Xu and Deng (2005) proposed two generalization of Squeeze algorithm (HE; XU; DENG, 2002), called dSqueezer and unsSqueezer. Both algorithms work with mixed datasets so that the former uses a discretization method to transform numerical attributes into categorical ones, and the last algorithm applies a compound similarity function to cope with both types of features. The authors provided an analysis regarding scalability of both algorithms, running them on large datasets. Although the authors stated they used simulated datasets in the experiments, the paper only uses real-word datasets. Ahmad and Dey (2007) proposed a new objective function and a

new way to measure dissimilarity of the observations considering the co-occurrence of values. Bushel, Wolfinger and Gibson (2007) presented the modk-prototype algorithm, a modification of k-prototypes algorithm to deal with medical data. The proposed algorithm considers simultaneously a microarray of gene expression and a classes of phenotypic variables, such as clinical assessments and histopathological observations. The authors also proposed an adaptation in a numerical validation measure, called DVI-CU.

Cheng and Leu (2008) proposed a new k-prototypes algorithm to deal with the constraints imposed by the bridge maintenance inspection problem, called Contract Packaging Technique (CPT). Brouwer (2008) presented a method to build a fuzzy membership matrix for mixed datasets. The authors applied this method in the fuzzy c-means algorithm. Cheng and Leu (2009) generalized the proposal made by them in the article published in 2008, naming the proposed constrained k-prototypes (CKP) algorithm. Lee and Pedrycz (2009) presented a new algorithm based on k-prototypes, called p-mode prototype, which considers the frequency of classes of categorical attributes to determine centroids. Lee and Pedrycz (2010) developed an adaptation method to create ordinal-numerical mappings for ordinal features from a mixed dataset. From this method, the authors modified the fuzzy c-means algorithm, calling their approach of adaptive fuzzy c-means (AFCM). Ahmad and Dey (2011) proposed a modification in the k-means algorithm so that it could work with mixed data in subspaces. The authors inspired themselves in the work of Jing, Ng and Huang (2007) and Ahmad and Dey (2007). Cheng and Leu (2011) combined the KJ method and the CKP algorithm, which they had previously proposed in 2009. Pham, Suarez-Alvarez and Prostov (2011) used the bee meta-heuristic (bee algorithm) to initialize the centroid of the k-prototypes algorithm and prevent it from getting stuck in local minimums. The authors called their algorithm of RANKPRO (random search with k-prototypes algorithm).

Irigoiien, Sierra and Arenas (2012) applied the proposal made by themselves in the work of 2008 ((IRIGOIEN; ARENAS, 2008), classified as IAPRA - Section 4.6.6), to detect relevant clusters and outliers in gene expression datasets. Moreover, they used as intern validation measure the quality index called INCA (see more in Section 4.6.6), and carried out the partitioning around medoids algorithm (PAM) (KAUFMAN; ROUSSEEUW, 1990) with the Gower distance (GOWER, 1971). Ji *et al.* (2012) proposed a new dissimilarity measure based on the co-occurrence of values to make the k-prototypes algorithm able to cope with the fuzzy sets. Liang *et al.* (2012) determined the number of groups with a new dissimilarity and validation measure (category utility function for mixed data - CUM). In addition, the authors developed an approach to penalize the worst group, reallocating their objects in the remaining clusters.

Cheung and Jia (2013) presented an algorithm, called iterative clustering learning based on object-cluster similarity metric (OCIL), that calculates the number of groups,

penalizing those that are redundant. Hennig and Liao (2013) addressed the social stratification problem using two types of data clustering algorithms, partitional and model-based. The partitional one, k-medoids, is the reason for this work is listed here, although the same paper is presented in Section 4.6.3 as well. The authors brought an in-deep analysis to carry out the social stratification, discussing data transformation in a semantic way, some basic concepts of the data clustering technique, interpretation of the groups, and the differences regarding performance measures in clustering analysis. It is an excellent paper which is highly cited by the scientific community. Ji *et al.* (2013) based on two documents ((KIM; LEE; LEE, 2004; HUANG *et al.*, 2005)) to propose a new way to obtain prototypes for categorical data and a new dissimilarity function for the k-prototypes algorithm based on the relevance of different attributes towards the clustering process. Ovesná *et al.* (2013) analyzed the genetic structure of a barley collection by clustering the data using k-means algorithm with the Gower distance. Han *et al.* (2014) generalized the k-anonymity algorithm to mixed data, proposing two new approaches, MAGE (microaggregation and generalization) and TSCKA (two steps clustering k-anonymization algorithm). The k-anonymity algorithm aims to protect data privacy during the mining process.

Pasquier, Naciri and Jeanmonod (2015) applied the principal component analysis (PCA) with the Gower distance and k-means algorithm to identify groups in a agriculture dataset. Harikumar and Surya (2015) proposed a modification of k-medoid to work with mixed data, through a new dissimilarity function. Ji *et al.* (2015) presented a method based on density and distance to address the initialization and updating of the centroids in the k-prototypes algorithm, called cluster centers initialization based on density and distance (CCIDD). Although the authors stated that, as far as they knew, there was no method that addressed the initialization of the centroids for mixed data clustering in the literature until that moment, there was the work of Cheung and Jia (2013) that had already been suggested a method for the same problem. Sangam and Om (2015) proposed a hybrid similarity coefficient and used it in the proposed method, called hybrid data labeling algorithm (HDLA), which aims to assign a new record to one of the existing groups. The authors shared the idea that in large datasets, or in a data stream scenario, it is usual to cluster a sample of objects and then assign the remaining objects to the clusters previously found. This assignment was the concern of this paper. Wei, Chow and Chan (2015) brought a new proposal for data transformation approaches. The authors proposed a method that transforms non-numerical data into numerical ones, which according to them, avoids information loss. The proposed method, called unsupervised feature transformation (UFT), is based on information theory.

Ahmad and Hashmi (2016) generalized the k-harmonic algorithm to mixed data, taking inspiration from the work of Ahmad and Dey (2007) to do this adaptation. The k-harmonic algorithm is a numerical data algorithm that proposes a better method to

initialize the centroids instead of choosing them randomly. The authors called the proposed algorithm as k-means type clustering algorithm for mixed datasets (KMCMD). Amorim and Makarenkov (2016) presented a modification of the k-means for mixed data, assigning different weight values for the attributes in a subclustering scenario. The authors proposed a method called Minkowski weighted k-means with distributed centroids (MWK-DC). Bishnu and Bhattacharjee (2016) presented a generalization of the k-modes algorithm for mixed data in order to apply the proposed algorithm in a software cost analysis. Drakopoulos *et al.* (2016) addressed the problem of the initialization of the centroids and the definition of the number of clusters in the k-means algorithm for mixed data based on density and graph theory, respectively. The authors, who were inspired by the work of Gourgari and Makris (2015), defined a grid on the original vector space to estimate the density used in the centroids initialization proposal. Besides, they constructed a fuzzy graph based on a new similarity function for mixed data to determine the number of clusters. They also proposed an external validation index (Tversky asymmetric index) to quantify the relevance of clusters. Gocsik *et al.* (2016) carried out the k-means algorithm in a mixed dataset, which was composed of data from a survey with broiler and fattening pig farmers. The authors did not mention how they handled the mixed data in the k-means algorithm. Thong and Son (2016) modified the fuzzy clustering algorithm on picture fuzzy set (FC-PFS) by defining multiple centroids using the particle swarm optimization (PSO) and proposing a new dissimilarity measure to cope with categorical data as well. The author's proposal was called picture fuzzy clustering algorithm for complex data (PFCA-CD).

Huerta-Muñoz, Ríos-Mercado and Ruiz (2017) modified the k-means algorithm for the costumer segmentation problem. In their proposal, the authors defined a specific dissimilarity measure for each attribute from a specific dataset. Then, they sum all the dissimilarities to compose the objective function of k-means. In addition, they also proposed a new heuristic based on the meta-heuristic iterated greedy (IG), called the iterated greedy search for customer segmentation (IGACS), to improve the centroids initialization process. The algorithm that includes all this proposal is called modified k-means (MKM). Kampert, Meulman and Friedman (2017) extended the COSA algorithm from Friedman and Meulman (2004). The main output of rCOSA is a dissimilarity matrix, which can be imputed in distance-based clustering algorithms, such as partitional and hierarchical ones, in addition to some other methods, such as self-organized maps. This is the reason this paper is in this subsection and in the hierarchical branch (Section 4.6.5). Kim (2017) proposed a method, called fast k-prototypes (FKPT), to accelerate the execution time of the k-prototypes algorithm by reducing the distance calculation. Liang, Wang and Lai (2017) proposed a new algorithm, called WKStream, to cope with mixed data stream with different shapes. They utilized the concepts of micro- and macro-clusters and also developed another algorithm, called weighted k-means+ (Wk-means+), which

weights numerical and categorical attributes based on non-uniformity concept. Özbakır and Turna (2017) modeled two meta-heuristics, called ions motion optimization (IMO) and weighted superposition attraction (WSA), to define the number of clusters. Wangchamhan, Chiewchanwattana and Sunat (2017) addressed the centroid initialization problem in the k-means algorithm with the Gower distance. The authors proposed three algorithms, all based on league championship algorithm (LCA), called i) chaotic league championship algorithm (C-LCA); ii) search clustering using chaotic league championship algorithm (SC-LCA), and iii) hybrid k-means and chaotic league championship algorithm (KSC-LCA).

Bathla, Aggarwal and Rani (2018) proposed the implementation of the k-prototypes algorithm in MapReduce to work with big data. Davoodi and Moradi (2018) used a mixed data clustering algorithm based on fuzzy k-prototypes algorithm to determine the number of rules in each layer from a deep learning framework, called deep rule-based fuzzy system (DRBFS), with the objective of predicting mortality at an emergency unit. Deng *et al.* (2018) proposed a new method based on the k-medoids algorithm to analyze the rural migrant population in China. The authors used PAM algorithm with the Gower distance. Gu *et al.* (2018) proposed a modification of the k-prototypes algorithm so that it could handle multi-objective optimization in a real-world problem, subject to restrictions. Hong *et al.* (2018) presented a modification in the dissimilarity function of the k-prototypes algorithm to manage the charging behavior of electric vehicles. The authors combined a hierarchical clustering algorithm with the k-prototypes to develop a method, called hierarchical clustering and improved k-prototypes method based on average density and rough set theory (HKPDR), which determines the number of clusters and initial centers of the clusters. Jia and Cheung (2018) extended the OCIL algorithm, proposing weights for the attributes in different groups (subspace). In addition, the authors presented a method, called WOCIL, to determine the centroids initialization and the number of groups.

Li *et al.* (2018) developed an algorithm to determine the number of groups, considering a penalty method to avoid redundant groups, and a method for centroids initialization to perform analysis of exceptional events in warehouses. The proposed method was called penalized competitive clustering algorithm based on the standardized initial clustering centers (SIC-PCC). Najjar *et al.* (2018) proposed a framework to cope with medical data that came from many sources, such as hospital stays, consultations, and visits. The algorithm, which was based on the k-prototypes algorithm proposed by Najjar, Gagné and Reinharz (2014), was used to cluster a medical data. Sangam and Om (2018) presented a new dissimilarity function for the k-prototypes algorithm to equalize the scales of both dissimilarities (categorical and numerical), without losing the scalability of the algorithm. Sangam and Om (2018) generalized the HDLA algorithm for data streams. The authors proposed a framework, called Equi-Clustream, which is composed of three algorithms to cope with mixed time-evolving data. The first algorithm, called a hybrid drifting concept detection, identifies the clusters on the current time window considering the clusters of

the previous time window. The second algorithm, called hybrid data labeling, assigns the clustering labels for the remaining data, and the last algorithm designs the visualization of the clusters in the different time windows. Szepannek (2018) presented an implementation of the k-prototypes algorithm on the R programming language. The authors also proposed a default value for the weight parameter for categorical variables, allowing the user changes this value. Yao *et al.* (2018) presented a method that initializes the centers of the clusters and updates the number of groups in the k-prototypes algorithm. Yuan, Zhang and Feng (2018) proposed a new method, called neighborhood information entropy-based outlier detection algorithm (NIEOD), to detect outliers based on entropy theory and maximum likelihood search.

HajKacem, N'cir and Essoussi (2019) proposed a method, called accelerated MapReduce-based k-prototypes algorithm (AMRKP), to accelerate the k-prototypes algorithm by the minimization of the distance calculation. Dutta, Sil and Dutta (2019) proposed a multi-objective optimization algorithm to define the optimal number of clusters using the k-prototypes as the clustering algorithm and genetic algorithm as the optimizer, called multi-objective genetic algorithm for k-prototypes (MOGA-KP). Noia *et al.* (2019) used the k-means algorithm with a specific dissimilarity function for each type of attribute in a health-inspired application. After clustering the data, the authors used two supervised algorithms with the labels resulted from the clustering algorithm. Ji *et al.* (2019) used the meta-heuristic artificial bee colony (ABC) and chaos theory to initialize the centers of the groups in the k-prototypes algorithm. The proposed method was called the artificial bee colony clustering based on k-prototypes (ABC k-prototypes). Lv *et al.* (2019) proposed a new method, called optimizing and differentially private clustering algorithm (ODPCA), to protect personal information in the data mining process by adding noise, of each iteration when updating the cluster centers. Although there is the k-prototypes algorithm to cope with mixed datasets, the authors used the k-means and k-modes algorithms to handle both types of features, numerical and categorical, respectively. The justification for this decision was not clear. Maione, Nelson and Barbosa (2019) used the PAM algorithm to cluster a mixed dataset whose content consisted of data from small producers and farmer families. The authors used the silhouette index to determine the number of clusters and the Gower distance as the dissimilarity function. The proposed method also determines a supervised learning algorithm to assign the new data into one of the clusters defined in the first step. The work of Shukla and Nanda (2019) aimed to design a risk assessment model for the credit card score problem. Therefore, the authors proposed a parallel social spider optimization algorithm (P-SSO), which was an extension from the work of Shukla and Nanda (2016), to optimize the dissimilarity function, looking for good initial centers in a data clustering algorithm based on the k-prototypes. Yao *et al.* (2019) proposed a modified k-prototypes algorithm, inspired on the work of Liang *et al.* (2012), to determine the number of groups automatically using the entropy theory. Their proposal was inspired

by an internet of things (IoT) application.

4.6.1.1 Discussion

From Table 43 showed in the Appendix B, we can note that five of the 59 documents (8.5%) proposed new solutions for large datasets, four (6.8%) for a dynamic scenario, and only one article (1.7%) worked with high-dimensional datasets. There were also a few articles (8.5%) which made contributions to preprocessing tasks. These contributions addressed the transformation of attributes, outliers detection, and feature scaling. Regarding the processing tasks, many studies (41%) proposed contributions in the dissimilarity function, 23.8% addressed the problem of defining the number of groups, 20.3% performed the contribution in the definition of the centroids. To a lesser extent, there were documents that addressed the problem of feature weighting (11.9%), constraints in the clustering process (8.5%), the clustering in data subsets (6.8%), the importance of the variables (3.4%), clusters updating (1.7%), and privacy protection in data clustering activity (1.7%). Besides, there were papers that used the fuzzy logic (8.5%), and rough sets (1.7%). It is important to note that a single paper may have addressed more than one contribution, which makes the sum of the above percentages greater than one hundred. Only six documents contributed to post-processing activities, such as the development of validation measurements (3.4%), visualization of the groups (3.4%) and interpretation of the groups (3.4%). Regarding the auxiliary approaches, 10% of the documents used a pure application of a previously proposed data clustering algorithm, 1.7% contributed proposing a framework for a specific real-world problem, 1.7% contributed with writing and providing code of an already proposed algorithm, and 23.7% used some method (optimization and competitive learning) to compose a contribution made in one of the three steps from the KDD process (preprocessing, processing and post-processing activities).

4.6.2 Hierarchical

Philip and Ottaway (1983) explored the usage of the data clustering algorithm with the Gower distance and the idea of feature weighting. The authors also applied the clustering technique to analyze the age of old weapons. Frisvad (1994) used a data clustering algorithm (unweighted pair-group method using arithmetic averages) in a chemical application. Harch *et al.* (1996) used the Gower distance with the ward hierarchical algorithm in an agriculture application. Sieber, Petrini and Greenacre (1998) transformed mixed data into single-type data and applied a categorical hierarchical algorithm to define fungal taxonomy. Friedman and Meulman (2004) proposed a new mixed data clustering algorithm, called clustering of objects on subsets of attributes (COSA), which addresses the subspace clustering problem with different attribute weights for each sub clustering, the missing data issue, and the optimization of the weights of the attributes. Shin and Sohn (2004) analyzed a large dataset of mobile telecommunication information. Although the authors stated the dataset

was composed of mixed-type attributes, they did not mention which dissimilarity function was used or if they transformed the mixed data into single-type attributes. After clustered the data with the single linkage clustering algorithm, the authors evaluated each segment with data envelopment analysis (DEA), classifying each group as efficient or inefficient. Then, a neural network was applied considering as the label the DEA classification.

Hsu, Chen and Su (2007) proposed an algorithm, called agglomerative hierarchical clustering based on distance hierarchy concept (DHC) to cope with mixed data. Sheu (2007) used a hybrid fuzzy-hierarchical clustering as part of a proposed framework that clusters costumers to facilitate the logistic distribution operations. In the usage of the clustering algorithm, the authors transformed categorical attributes into a binary representation. Herr (2010) proposed a strategy to analyze mixed data resulted from research surveys. To explain its proposal, the author took into consideration the agent-based model, although the author affirmed the proposed strategy could be applied in other types of research. From that, the proposal consisted of clustering the data, applying a decision tree analysis to allocate the new registers into clusters, and extracting the most important variables from the clusters through a visual inspection process using an extended random forest method. To identify the optimal number of clusters, the proposed method used the gap statistic, which determines the optimal number of clusters in relation to a reference distribution and the within-cluster dispersion. Chitradevi *et al.* (2012) proposed a new method, called distributed agglomerative cluster-based anomaly detection algorithm (DACAD), to assist wireless sensor networks in identifying data. Their algorithm, which was based on the k-nearest neighborhood (KNN) algorithm, is a distributed agglomerative hierarchical clustering algorithm, which aims to detect faulty network reading. Fonseca (2013) proposed a comparison among clustering algorithms and latent class models in social sciences. The authors applied both methods in three datasets of which one was composed of mixed features. Sopan *et al.* (2013) applied hierarchical clustering algorithms in a proposed framework to explore data distributions and visualization.

Horta and Camanho (2014) used a hierarchy algorithm (balanced iterative reducing and clustering using hierarchies, BIRCH), which is able to cope with mixed data and is implemented in SPSS software, to identify the competitive positioning of companies from the construction sector. Afterward, the authors used DEA to evaluate the performance of each found cluster. Brida *et al.* (2014) used the ClustOfVar algorithm, the PCA technique, and the classification and regression tree (CART) to analyze an official survey dataset done with tourists who disembarked in Uruguayan ports. The work of Kampert, Meulman and Friedman (2017) has already been described in Section 4.6.1. However, the authors also used hierarchical clustering algorithms to address the feature weighting problem, justifying the fact this paper also appears in this section. In the work of Najjar *et al.* (2018), the authors proposed a method to analyze disease evidences, such as medical visits, consultations, and hospital stays, to deliver treatment pathways for the patients. The first

proposal consisted of applying the k-prototypes algorithm to cluster the medical events and constructing treatment pathways. The second clustering approach corresponded to a hierarchical clustering algorithm, justifying the selection of this paper in this branch. To classify the soil, Bragato, Ganis and Feoli (2019) applied a complete linkage hierarchical clustering algorithm and compared two dissimilarities (Burnaby and Goodall) in the same clustering method to analyze the influence of each one in the clustering results. They found that the Burnaby dissimilarity considered association among features, assigning higher weights to independent attributes, and Goodall dissimilarity considered more similar the observations that share infrequent values than those that share frequent values.

4.6.2.1 Discussion

From Table 44 showed in the Appendix B, we can note that four of the 17 documents (23.5%) addressed proposals that contemplate different datasets, such as the large (11.8%), dynamic (5.9%), distributed (5.9%) and high-dimensional (5.9%) ones. It is important to note that Najjar *et al.* (2018) addressed two (distributed and large) scenarios for the dataset. There were also a few articles (11.8%) that make contributions to preprocessing tasks, such as missing data (5.9%) and outliers detection (5.9%). Regarding the processing tasks, the papers contributed to the feature weighting (11.8%), the clustering in data subsets (11.8%), dissimilarity function (11.8%), clustering the data considering fuzzy logic (5.9%), constraints in the clustering process (5.9%) and the definition of number of groups (5.9%). Only 3 documents contributed to post-processing activities, such as the development of performance measures of the clustering (5.9%), visualization of the groups (5.9%) and interpretation of the groups (5.9%). Regarding the auxiliary approaches, 64.7% of the documents do pure applications of a previously proposed data clustering algorithm, and 35.3% use optimization algorithms (11.8%) as an auxiliary method to be part of their theoretical proposals.

4.6.3 Model-based

Wong and Chiu (1987) proposed a method to cope with incomplete mixed datasets using a discretization method based on entropy theory, which transforms the numerical attributes into categorical ones, and an inference method to fill missing values. In addition, the authors analyzed the initialization of the clusters based on nearest-neighbor distance. Harch *et al.* (1999) used a procedure from SAS software, called FASTCLUST, and an algorithm, called MULTIMIX (HUNT, 1996) (JORGENSEN; HUNT, 1996), to analyze a groundnut germplasm data. Hunt and Basford (1999) proposed the first algorithm of finite mixtures to cope with a three-mode three-way array. The authors considered that the numerical data were normally distributed in the clusters, inspired on the work of Basford and McLachlan (1985), and categorical data followed a mixture of multinomial distributions.

Bacher (2000) proposed a probabilistic clustering model, based on the latent class model (LCM), which determines the number of groups, using statistic tests and goodness-of-fit measures. Hunt and Jorgensen (2003) extended the MULTIMIX algorithm, which was proposed by their own (see in Harch *et al.* (1999)). The paper with the MULTIMIX proposal did not return in this survey because it did not fill our criteria. Regarding the newest paper (2003), the authors addressed the missing and mixed-type data problem. McLachlan and Chang (2004) proposed a theoretical discussion about mixture models for clustering analysis, addressing both single and mixed-type data. Gondek and Hofmann (2007) extended the information bottleneck framework (IBF), which is a method based on mutual information, to avoid of redundancy of knowledge in different groups. The authors named their proposal coordinated conditional information bottleneck (CCIB). In addition to their mentioned proposal, the authors extended the IBF for semi-supervised learning. This is the reason this paper also appears in Section 4.6.5.

Chatzis (2011) combined the fuzzy c-means algorithm (FCM) with Gauss-multinomial distribution and extended the gath-geva (G-G) algorithm. The authors' proposal was called the Kullback-Leibler fuzzy c-means algorithm for Gauss-multinomial-distributed data (KL-FCM-GM). Hunt and Jorgensen (2011) provided a wide discussion about model-based mixed data clustering approaches. In addition, the authors presented a few options of softwares that can be used to cluster mixed datasets. Fonseca (2013) provided a cluster analysis in three datasets from the social science area, using hierarchical and model-based algorithms. As the authors used two types of clustering algorithms, we have also described this paper in Section 4.6.2. The work of Hennig and Liao (2013) has already been described in Section 4.6.1, because the authors provided an analysis of social-economic stratification from the perspective of two types of clustering algorithms, partitional and model-based. Regarding the model-based algorithm, the authors used a latent class clustering to carried out the stratification.

Zhang and Ip (2014) proposed a new measure to evaluate the absolute and relative impacts of mixed attributes in discriminant latent class models. To evaluate the proposal, they considered two datasets, the first from the work of Moustaki and Papageorgiou (2005) and another from Frazer *et al.* (2007). Baudry *et al.* (2015) proposed an information criterion based on integrated completed likelihood (ICL) criterion, called SILC criterion, to define the number of groups, which considers an external categorical variable to support this decision. This external attribute is regarding information from the domain of the specialist. Blomstedt *et al.* (2015) addressed the mixed data clustering problem, proposing a Bayesian predictive framework (BPF) with an optimizer, called greedy stochastic search (GSS), to search the optimal partition. Ingrassia *et al.* (2015) introduced a generalized linear mixed cluster-weighted model (GLMCWM), based on cluster-weighted models (CWM).

Foss *et al.* (2016) proposed a new statistical model based on a semi-parametric approach (k-means and finite mixture models (FMM)) to fill the gap regarding the unbalanced contribution of numerical and categorical attributes. In addition, their algorithm, which called KAymeans for M_Ixed L_Arge data (KAMILA), can handle non-normal distribution data. Hunt and Basford (2016) evaluated the MULTIMIX algorithm using many classic criteria. McParland and Gormley (2016) proposed a new statistical model, called clustMD, which is based on the latent variable framework and a suite of parsimonious mixture models, to cope with mixed datasets.

McParland *et al.* (2017) proposed a new method based on the Bayesian finite mixture model (BFMM) to cluster a real-world dataset from the pan European LIPGENE-SU.VI.MAX (supplementation en vitamines et minéraux antioxydants) study⁵, which is a studying that considers both phenotypic and genetic factors to analyze a metabolic syndrome. The authors also proposed a feature selection method to cope with high-dimensional data and a novel approximation of the likelihood function incorporated with the Bayesian information criterion (BIC) and the Markov chain Monte Carlo (BIC-MCMC) criterion to select the number of clusters and latent factors. Their proposal was called mixture of factor analyser for mixed data (MFA-MD). Tekumalla, Rajan and Bhattacharyya (2017) extended the vine copula (VC) method for mixed datasets, proposing two new inference algorithms that fitting vines on mixed data, called Ext-DVine (EDV), and Ext-Vine-MVC (EVMVC). Abidin and Westhead (2017) inspired themselves by a health area problem to propose a new statistical model for mixed data.

Foss and Markatou (2018) proposed a package in the R environment for the KAMILA algorithm and provided a function to generate simulated mixed datasets with two clusters and the code for the work of Modha and Spangler (2003). The authors also provide an adaptation of the KAMILA algorithm for Hadoop. Guglielmi *et al.* (2018) proposed a statistical model, based on Bayesian statistic models (BSM), to apply it in a problem in the health area. Storlie *et al.* (2018), motivated by the usage of data clustering technique in a real-world problem in health area (patients with potentially autism spectrum disorder), proposed a new method to cope with missing values and to carry out feature selection. Their proposal, called DPM-vs, is a method based on the Dirichlet process model (DPM). Wang, Zhu and Dong (2018) proposed a probabilistic model, based on FMM, called the Poisson Gaussian multinomial mixture model (PGMM), to capture the compactness and the size distribution of the clusters.

Biernacki and Lourme (2019) proposed a new statistical model for clustering and co-clustering, considering both types of data. Marbac and Sedki (2019) presented a code package of the methods that were proposed a previous paper of the authors (see in Marbac and Sedki (2017)), which did not return in this review. The package, called VarSelLCM,

⁵ www.ucd.ie/lipgene

implements a statistical model that can cope with missing values and carry out the feature selection, optimizing two criteria, BIC and maximum integrated complete-data likelihood. In the work of Marbac, Sedki and Patin (2019), the authors applied the VarSelLCM package in a real-world dataset and in other datasets as well. Marbac and Vandewalle (2019) proposed a framework that assumes the existence of independent blocks of variables that follow a latent class model. The authors' proposal consists of finding these blocks of features and thereafter clustering the individuals in each block. This method is not a co-clustering algorithm, because in co-clustering the attributes and the observations are clustering simultaneously, as pointed out by the authors.

4.6.3.1 Discussion

From Table 45 showed in the Appendix B, we can note that four of the 29 papers (13.8%) proposed new solutions for large datasets, one (3.4%) for high-dimensional datasets, and one (3.4%) for three-mode three-way datasets. There were also some articles (27.6%) that made contributions to preprocessing tasks. These contributions addressed data transformation, missing data, and dimensionality reduction. Regarding the processing tasks, many studies (72.4%) proposed contributions in statistical models, and 10.3% addressed the problem of defining the number of groups. To a lesser extent, there were documents that addressed the importance of variables (3.4%), variables correlation (3.4%), clusters updating (3.4%). It is important to note that a single paper may have addressed more than one contribution, which makes the sum of the above percentages greater than one hundred. Only two documents (6.9%) contributed to post-processing activities, developing performance measures of the clustering. Regarding the auxiliary approaches, 10.3% contributed proposing a framework for a specific real-world problem, 10.3% made a comparison and discussion about model-based mixed data clustering algorithms, 6.9% of the documents used a pure application of a previously proposed data clustering algorithm, 6.9% contributed with writing and providing code of an already proposed algorithm, 6.9% of the documents used parallelism, 3.4% applied optimization methods and 3.4% used specialist knowledge to build the proposals.

4.6.4 Neural network-based

Hsu and Wang (2006) proposed a framework to cope with mixed data, based on self-organized maps (SOM) algorithm, distance hierarchy concept, and attribute-orientated induction (AOI). The authors' framework has two main steps, in which the first consists in using a proposed generalized SOM (GSOM) to cluster the data, providing visualization and exploration tasks; while the second one consists in using a proposed extended attribute-orientated induction (EAOI) to extract the characteristics from each cluster. The authors used the average category utility (ACU) and entropy to verify the clustering quality. Hsu and Huang (2008) expanded the adaptive resonance theory network algorithm type I

(ART1) and type II (ATR2) to cope with mixed data, called modified adaptive resonance theory network (M-ART). Their proposal was also based on the distance hierarchy concept.

Tai, Hsu and Chen (2010) proposed a growing mixed-type SOM (GMixSOM) to cope with mixed-type data in a dynamic scenario. The authors based their proposal on the GenSOM algorithm from the work of Hsu and Wang (2006). Hsu, Lin and Tai (2011) has concerned about incorporating the semantics inherent in categorical values in the dissimilarity function used in a SOM algorithm. The authors proposed an extended SOM, called MixSOM, which is based on the generalized visualization-induced self-organizing map (GViSOM) algorithm proposed by Chung-Chian Hsu, Kuo-Min Wang and Sheng-Hsuan Wang (2006). The MixSOM incorporates the distance hierarchy concept to calculate the dissimilarities of categorical data. Hsu and Lin (2012) proposed a very similar approach they had already proposed in the work of 2011 (see in Hsu, Lin and Tai (2011)). Here, the authors extended the SOM algorithm and used the distance hierarchy concept to propose the same algorithm, called MixSOM, although the computational experiments were different.

Tai and Hsu (2012) presented a very similar work they had already presented in 2010 (see in Tai, Hsu and Chen (2010)). We decided to maintain both papers because, although they are very similar, they are not the same (the computational experiments are different). In addition, the first paper was published in a conference and the second one was published in a journal. Liao and Hsu (2012) proposed a new dissimilarity method for an adapted SOM algorithm so that the relevancy information could be taken from the distance hierarchy concept. In addition, the authors based the proposed algorithm in a tree growing adaptation method and integration of the U-Matrix for visualization. Noorbehbahani, Mousavi and Mirzaei (2014) proposed an adaptation in the SOM algorithm, adjusted self-organizing incremental neural network (ASOINN), so that it can calculate the number of clusters automatically, considering a new dissimilarity function and updated rules for mixed data. Coso *et al.* (2015) proposed a new dissimilarity function so that categorical data could be used in its original form. The authors based on batch SOM (NCSOM) to propose the frequency neuron mixed self-organized map (FMSOM), which considers the dissimilarity of categorical data as the probability of the reference vector that does not contain the category present on the input vector. Malondkar *et al.* (2019) adapted the growing hierarchical self-organizing map algorithm (GHSOM), which is a dynamic version of the SOM algorithm that produces multi-levels of maps (hierarchically), to handle mixed features and large datasets in the Spark platform. The authors' proposal called Spark-GHSOM.

4.6.4.1 Discussion

From Table 46 showed in the Appendix B, we can note that nine of the 10 documents (90%) addressed proposals that contemplate different datasets, such as the large (20%), dynamic (40%), distributed (10%) and high-dimensional (60%) ones. It is important to note that many documents address more than one scenario for the dataset, which is the reason why the sum of the individual percentages is more than one hundred. There was also one article (10%) that made contributions to the preprocessing task (noisy reduction). Regarding the processing tasks, all the papers contributed to improve the dissimilarity function, 30% clusters updating, 10% proposed a new way to update the rules of the algorithm and 10% proposed a new way to determine the number of the groups. With respect to post-processing activities, 7 papers contributed to the visualization of the results (this is because they used the SOM algorithm, which is an algorithm that already delivers this feature for the decision-maker), and one article improved the way the groups are interpreted. Regarding the auxiliary approaches, only 10% of the documents used parallelism.

4.6.5 Others

Ichino and Sklansky (1985) proposed an adaptation of the relative neighborhood graph (RNG) and Gabriel graph (GG) so that it could handle mixed data, specifically the ordinal features. In their proposal, the authors changed the concept of circle influence graph to rectangular influence graph (RIG) and adopted the Minkowski distance instead of the Euclidean distance. Hence, they presented the interclass RIG (IRIG) and the mutual neighborhood graph (MNG). It is important to note that they adapted the algorithm considering continuous and ordinal (discrete) variables, but not the nominal ones. Veiga (1996) presented a new conceptual clustering algorithm based on the ITERATE algorithm (BISWAS; WEINBERG; LI, 1994). The author's proposal consists of three main contributions, which are: i) preordering process to reduce the computational cost of the ITERATE algorithm; ii) improving the control strategy, and iii) handling the missing and mixed data. Hai and Susumu (2004) proposed an improvement in the k-sets algorithm so that it could handle categorical and mixed data and achieve high clustering quality. Besides, the authors implemented the proposed algorithm in a parallel scenario to cope with large datasets, calling their proposal the parallel clustering algorithm for categorical and mixed data (PCACMD). Hsu, Jao and Chen (2005) proposed a clustering technique, called Clus-Tree, to construct conjecture rules for unlabeled datasets in decision trees. The authors used the k-nearest neighborhood graph (K-NNG) and clustering tree concept (CTC) to build their proposal.

Puthongsiriporn *et al.* (2006) proposed a new technique, called evolved pattern-based knowledge induction (ePKI), to develop neighbor hierarchies at the attribute level.

This method, which was based on pattern-based knowledge induction (PBK), is used when an exact answer can not be given for a specific question; therefore, using ePKI, the user can obtain approximate answers considering mixed-type attributes. In the experimental section, the authors brought an industrial problem, providing a component-machine table to define the clusters of machines in a process layout. Although the paper affirmed that the ePKI can handle mixed-type attributes, the authors did not provide any example or dataset with both types of attributes. Hsu and Chen (2007) proposed an algorithm based on variance, entropy and distance hierarchy concept to cope with mixed features. The algorithm, called CAVE, measures the dissimilarity of numerical attribute using variance and the dissimilarity of categorical features by entropy and distance hierarchy. The work of Gondek and Hofmann (2007) has already been mentioned in Section 4.6.3, because the authors proposed a model-based algorithm. However, at the end of the article, they extended the original proposal into two branches, which consider a semi-supervised scenario and successive non-redundant clustering. This is the reason their paper appears in this section.

Ye, Li and Orlowska (2009) proposed a new algorithm, called projected outlier detection in high-dimensional mixed-attributes datasets (PODM), to detect projected outlier in high-dimensional mixed-type datasets. Their proposal, which is based on information entropy, identify and explain anomaly subspace, reporting where the outliers are located. Al-Shaqsi and Wang (2010) proposed a clustering ensemble algorithm to cope with mixed-type attributes. Yang and Wu (2011) proposed a mixed data clustering algorithm, based on k-modes, for large datasets. The authors based on graph theory to assign the observations for each center of the clusters. They addressed the feature weighting and number of clusters problems and used the nearest-neighbor clustering and genetic algorithm.

Zhao *et al.* (2013) proposed a new graph-based clustering method for multidimensional datasets. The authors' proposal allows a multimembership for each observation. According to the authors, the fuzzy clusters and the resulted tree hierarchy are what differ their method from the other ones. The experiments did not consider any mixed dataset, only one pure numerical set of data. Wang, Machiraju and Huang (2014) proposed a new consensus clustering algorithm, called molecular regularized consensus patient stratification (MRCPS), inspired by a problem in the health area. Zhang and Gu (2014) proposed a new method based on the affinity propagation clustering algorithm (APC) and the work of Ahmad and Dey (2007). Niu *et al.* (2015) proposed a coupled user clustering algorithm based on spectral clustering, called CUCA-MD, for web-based learning systems. The method presents three main contributions, which are: i) it considers the coupling relationship of the attributes, ii) it copes with mixed-type data, and iii) it delivers a customized learning service. Chen and He (2016) proposed a density-based data stream clustering algorithm which determines the centers of the groups automatically. The method,

called the self-adaptive peak density clustering algorithm for data with mixed attributes (ACC-FSFD), defines a new dissimilarity function for mixed data and introduces a new concept of the microcluster to cope with the dataset dynamic. The dissimilarity function considers a balance factor that represents the ratio of numerical and categorical attributes in a dataset. This factor allows the algorithm amplifies dominant attributes effects and reduces the effects of non-dominant ones. A semi-supervised clustering method is considered in case of non-dominant attributes are more important for clustering results than the dominant ones.

Bhatnagar *et al.* (2017) proposed a data-aware scalable clustering (DASC) algorithm, which is an incremental algorithm inspired by a grid-based stream clustering algorithm called the ExCC, proposed by Bhatnagar, Kaur and Chakravarthy (2013). This method is able to handle three V's of big data, working with: i) distributed and large files (volume); ii) data stream (velocity), and iii) mixed-type attributes (variety). Ding *et al.* (2017) proposed an entropy-based density peaks clustering algorithm for mixed data employing a fuzzy neighborhood (DP-MD-FN) algorithm. Their proposal is able to find the centers of the groups automatically and, as it is based on density peaks, it is suitable for non-spherical clusters. Du, Ding and Xue (2017) proposed a modified density peaks clustering (DPC) algorithm, called density peaks clustering for mixed-type data (DPC-MD). Hummel, Edelmann and Kopp-Schneider (2017) proposed a new co-clustering algorithm, called CluMix, motivated by the biomedical studies necessity. The proposed method clusters samples and variables to create an overview visualization of the whole relations existed in the dataset. This visualization is a *heatmap*, which receives hierarchical clustering results from samples and variables. Iam-On and Boongoen (2017) proposed two new ensemble clustering algorithms, called WCT-T and WTQ-T, based on link-based cluster ensembles (LCE) and graph concepts, to improve the prediction of the student dropout in a Thailand university. Their method generates an ensemble-information matrix from ensemble clustering in order to input this matrix in a classification model. Jinyin *et al.* (2017) proposed a new density clustering, called fast density clustering algorithm (FDCA). The authors defined a novel dissimilarity function for mixed data and a new procedure to define the centers of the groups automatically. Liu, Yang and He (2017) proposed a density-based clustering algorithm that integrates entropy and probability distribution to cope with mixed attributes, called EPDCA. Their method is a modified DBSCAN (Density-based spatial clustering of applications with noise) algorithm so that it can determine automatically the number of clusters and calculate the dissimilarities by entropy. Liu *et al.* (2017) proposed a new density peaks clustering algorithm with a new unified distance metric. Their method, called fast search and find of density peaks algorithm (DPC-M), is based on the local density and relative distance. Noorbehbahani *et al.* (2017) proposed a semi-supervised clustering algorithm, called the incremental semi-supervised flow network-based intrusion detection system (ISF-NIDS). The method has two main

phases, an offline one, and another online phase. In the first one, a training set (unlabeled dataset) is clustered by the proposed unsupervised algorithm, called mixed self-organizing incremental neural network (MSOINN). In the second phase, the supervised proposed algorithm, called cluster adjustment (CA), receives as input the initial cluster set and the initial classification model resulted from the offline phase. Valenzuela *et al.* (2017) proposed a method to carry out a cluster network analysis of health and disease phenotypic in aging people. The method clusters the variables of a dataset using the Louvain community detection algorithm (LCDA) and a minimum spanning tree (MST).

Jang *et al.* (2018) proposed a grid-based k-prototypes (GK-prototype) algorithm to cope with spatial data in reduced execution time. To reduce the computational time, the authors proposed a grid-based indexing technique, which calculates the distance from the centers of the groups to the cells (grids), instead of observations. Lin, Azarnoush and Runger (2018) proposed a three ensemble clustering to cope with high-dimensional mixed datasets, named clustering algorithm using random forest-based iterations (CRAFTER). The proposed algorithm, which was inspired by random forest clustering (RFC), starts with the RFC algorithm in a small sample of the dataset because it considers the resulted groups from the previous step as the labels of the supervised problem. Zhao, Cao and Liang (2018) proposed a new method to generate base clustering in ensemble clustering approaches. The proposed method, called sequential ensemble clustering generation algorithm for mixed data (SECG), was based on information entropy and aimed to consider the correlation between different base clusters during the process of base clustering generation. Duan *et al.* (2019) proposed a new method based on DPC and stacked denoising autoencoders, called DPC-SDAE, which is able to determine the number of clusters automatically and identify non-spherical clusters.

4.6.5.1 Discussion

From Table 47 showed in the Appendix B, we can note that four of the 29 papers (13.8%) proposed new solutions for large datasets, four (13.8%) for a dynamic scenario, two for distributed datasets, and four articles (13.8%) for high-dimensional datasets. There were also a few articles (10.3%) which made contributions to preprocessing tasks. These contributions addressed the missing data problem and outlier detection. Regarding the processing tasks, most of the studies (69%) proposed contributions in the dissimilarity function. There were also documents, in minor numbers, that addressed the problem of graphs, clustering correlation, statistical models, feature weighting, number of clusters, ensemble procedures, fuzzy sets, centroids definition, clusters updating, the subsets clustering. Many papers addressed more than one contribution. Only two documents contribute to post-processing activities, such as the development of performance measures of the clustering structure (3.4%) and visualization of the groups (3.4%). Regarding the auxiliary approaches, 1.7% used the idea of framework for a specific real-world problem,

1.7% contributed using optimization methods, and 10.3% parallelism.

4.6.6 IAPRA

Gordon (1987) presented a review of some automated methods used in the analysis of soil data, among them are dissimilarity functions for mixed data. Cox and Cox (2000) proposed an extension of the Gower distance, which yields, simultaneously, a set of weighted dissimilarities between pairs of objects and variables. The weights of the objects and variables are automatically defined by the proposed method. Shi and Horvath (2006) explained, using practical and synthetic examples, the random forest dissimilarities (RFD) used to distinguish observed data from synthetic ones. The authors proposed a discussion and an in-depth analysis of RF dissimilarities and the Euclidean distance in many simulated datasets in order to guide the practitioners and researchers in their choices with respect to dissimilarity functions. Irigoien and Arenas (2008) proposed a new statistic index, called index number clusters atypical (INCA), to determine the number of clusters automatically and identify outliers or atypical units.

Barceló-Rico, Díez and Bondia (2011) proposed a polar coordinate-based data transformation to encode categorical data into numerical data. Besides, the author presented a broad comparison of the existing coding methods on that date. In the work of Barceló-Rico and Díez (2012), the authors generalized the categorical codifying proposal made by themselves in 2011 (see in Barceló-Rico, Díez and Bondia (2011)). In the paper of 2012, the proposed codification, which is based on polar and spherical coordinates, was used with the k-means algorithm for the sake of comparison with other codifications and mixed data clustering algorithms. Suarez-Alvarez *et al.* (2012) proposed a new normalization method for mixed data. In the experiment results, the authors showed the normalized dataset processed by their method reached better performance than the non-normalized datasets.

Bernard *et al.* (2014) proposed a visual-interactive system to assist researchers in finding relations between aggregated bins of multivariate attributes in large mixed datasets. The authors applied statistical dependency tests and hierarchical clustering to provide the results. Hennig and Lin (2015) proposed a method to investigate the distribution of a given validation index, thereafter this index can be used as a statistic test to verify the homogeneity hypothesis against the clustering alternative. Lam, Wei and Wunsch (2015) proposed the unsupervised feature learning with fuzzy ART (UFLA), which transforms categorical features into binary attributes, considering large and sparse datasets. After carrying out the data transformation, the authors used the fuzzy adaptive resonance theory (FART) to cluster the dataset. Prostov, Suarez-Alvarez and Prostov (2015) provided a mathematical proof for the normalization method proposed in Suarez-Alvarez *et al.* (2012), showing the properties of the sample estimators used in that approach.

Hsu and Huang (2016) proposed an extension of the t-distributed stochastic neighbor embedding (t-SNE) technique, called dimensionality reduction with distance hierarchy (DRDH), to cope with mixed data, using the distance hierarchy concept to measure the dissimilarity of categorical data, using for this the agglomerative hierarchical clustering. Rahman and Islam (2016) proposed a new fuzzy method to cope with missing values in mixed datasets, called fuzzy expectation maximization and fuzzy clustering-based missing value imputation framework for data pre-processing (FEMI). Taha and Hadi (2016) proposed two new measures of the association between a quantitative and categorical variables. The authors did not apply the proposed measures in any clustering algorithm, but mentioned it can be used in such methods. Kuppusamy and Paramasivam (2017) proposed a neural network-based clustering algorithm, called grey fuzzy neural network (GFNN), to fill missing data in a mixed dataset. The proposed approach is a result of the integration of the grey wolf optimizer (GWO) and adaptative neuro-fuzzy inference system (ANFIS).

Audigier *et al.* (2018) proposed a comparison and discussion about imputation method, a strategy to cope with missing values, for multilevel data with continuous and binary features. Nowosad and Stepinski (2018) adapted the V-measure to geographic community, changing its perspective from information theory to variance analysis. The proposed method, called spatial association between regionalizations (SABRE), can determine the optimal number of regions, handle with mixed data and be carried out on the R package provided by the authors. Ahmad and Khan (2019) proposed an extensive and wide traditional revision regarding mixed data clustering, describing many scientific documents in each class of clustering algorithm and discussing the challenges and opportunities in this theme. Foss, Markatou and Ray (2019) presented a review and a discussion about some dissimilarity functions, data transformation methods and partitional and model-based algorithms for mixed data. Golinko and Zhu (2019) proposed a feature embedding method, called generalized feature embedding approach (GEL), for supervised and unsupervised learning. The proposed method; which was based on previous work of the same authors (see in Golinko and Zhu (2017)), called GFEL (Generalized Feature Embedding Learning); can also integrate label information of the data and accommodate new data points. Velden, Enza and Markos (2019) presented a review bringing some distance-based clustering algorithms that handle mixed datasets and some dissimilarity measures used in this kind of dataset.

4.6.6.1 Discussion

From Table 48 showed in the Appendix B, we can note that two (9.5%) of the 21 papers proposed new solutions for large datasets, one (4.8%) for a dynamic scenario, one (4.8%) for sparse datasets, and one (4.8%) article considers high-dimensional datasets. As already expected, in this branch there were many articles that the main proposal was

regarding preprocessing and post-processing tasks. With respect to the former, there were papers that worked for outliers detection (4.8%), feature scaling (9.5%), data transformation (14.3%), missing data (14.3%), and dimensionality reduction (9.5%). Regarding the last, 3 papers proposed new performance measures and 2 addressed visualization tasks. Regarding the processing tasks, many studies (9.5%) proposed contributions in the dissimilarity function, 9.5% addressed the problem of defining the number of groups, 14.3% used fuzzy sets, 4.8% worked with clusters updating and 4.8% work with variable correlation. Concerning the auxiliary approaches, 23.8% of the documents discussed the theme, 14.3% reviewed a specific topic, 4.8% made a comparison of techniques, 4.8% contributed proposing a framework for a specific real-world problem, 4.8% used optimization methods, and 4.8% used as auxiliary approach a mathematical proof.

4.6.7 Answers of the questions defined for the individual analysis - Part 2

The following items bring the answers to each question of individual analysis, part 2, defined in Section 4.2.3. Note that the meaning of the acronyms showed in the mentioned figures ahead is defined in the Appendix B.

- Question I-SR-1 - Which are the main characteristics of datasets considered in the documents? Figure 31 shows the most addressed types of datasets in the document collection. As can be observed, the following characteristics of the datasets are the most addressed in the mixed data clustering documents: large, high-dimensional, dynamic, distributed, three-mode and three-way, and sparse.
- Question I-SR-2 - Which are the preprocessing tasks that received the most contributions? Figure 32 shows the six most addressed preprocessing tasks in the document collection. As can be observed, the following preprocessing tasks are the most addressed in the mixed data clustering documents: missing data, data transformation, dimensionality reduction, outliers detection, feature scaling and noisy reduction.
- Question I-SR-3 - Which are the processing tasks that received the most contributions? Figure 33 shows the six most addressed processing tasks in the document collection. As can be observed, the following processing tasks are the most addressed in the mixed data clustering documents: dissimilarity function, definition of the number of groups, statistical model, centroids definition, feature weighting, and fuzzy sets.
- Question I-SR-4 - Which are the post-processing tasks that received the most contributions? Figure 34 shows the most addressed post-processing tasks in the document collection. As can be observed, the following post-processing tasks are the most addressed in the mixed data clustering documents: visualization, definition of performance measures, interpretation of the clusters.

- Question I-SR-5 - Which are the auxiliary approaches most used? Figure 35 shows the six most addressed auxiliary approaches in the document collection. As can be observed, the following auxiliary approaches are the most addressed in the mixed data clustering documents: optimization, pure application, framework, parallelism, discussion, and comparison.

Figure 31 – Individual analysis, Part 2 - Most addressed datasets

Figure 32 – Individual analysis, Part 2 - Most addressed preprocessing tasks

Figure 33 – Individual analysis, Part 2 - Most addressed processing tasks

Figure 34 – Individual analysis, Part 2 - Most addressed post-processing tasks

4.7 Individual analysis - Part 3

In the Appendix C, one provides, in Tables from 52 to 69, the information regarding the datasets, validation measures, programming language and code availability from documents resulted in the systematic review, which are separated in the branches defined in this thesis. In addition to this, it is also provides, in the same appendix, the meaning of the acronyms used to fill those tables. In the following paragraphs, one makes a

Figure 35 – Individual analysis, Part 2 - Auxiliary approaches

summarization of the individual information of each document returned in the systematic review, bringing the datasets each document used, the performance measures that were chosen, the programming languages were used, and code availability. Note that a single document can have contributed with more than one item in each of the information we detailed, resulting in a sum higher than the number of documents in each branch.

From the studies resulted in the partitional branch, it was possible to note that less than half of the studies (28.9%) used simulated datasets in their computational experiments, as opposed to those (95%) that used real-world datasets. Most of the simulated datasets were created by the authors, who described the methods they used to generate them. Only 6.7% of the documents that used simulated datasets, made them available in packages in the R environment or in the supplementary material of their articles. Most of the real-world datasets were extracted from the UCI machine learning repository. Regarding the performance measures, in general, the documents used the three types of indices (internal, external and relative) to evaluate the clustering algorithm, with more emphasis for the external and internal ones. The usage of external indices, such as accuracy and the adjusted Rand index, stood out. In this branch, it was quite common the authors to measure the execution time of the algorithm, a fact that we did not observe in the other branches. Concerning the programming language and code availability, we note that some documents (8, i.e. 13.5%) provided the code they used in their proposals. This availability was generally done through a package in the R environment. Many of the articles (32, i.e. 54.2%) informed the programming language they used to implement the algorithms, such as MATLAB (11/32), C/C#/C++ (8/32), R (6/32), Java (5/32), and Python (2/32).

With respect to hierarchical branch, it was possible to note that only three documents (17.6%) used simulated datasets in their computational experiments, as opposed to those (100%) that used real-world datasets. Most of the simulated datasets were brought by the authors, and only one was available in package in the R environment. Most of the real-world datasets were brought by the authors, and some of them were extracted from the UCI machine learning repository. Concerning the performance measures, we can affirm that, in general, the articles used internal indices to evaluate the clustering algorithms followed by external and relative ones, which is justified by the objective of most of the documents. Many papers did not even use evaluation indices, showing pure application papers did not test the quality of the clusters found by the algorithms. No paper measures the execution time and accuracy of the approaches, which were the two measures wide used in the partitional branch. With respect to the programming language and code availability, only two documents (11.8%) made the codes available, which were done through packages in the R environment. There were also two documents that informed the programming language they used to implement their algorithms. It is important to observe that many papers in this branch informed the software they used to run an already proposed algorithm, which did not be included in this analysis.

Regarding the model-based branch, it was possible to note that approximately half of the studies (51.7%) used simulated datasets in their computational experiments, as opposed to those (82.8%) that used real-world datasets. Most of the simulated datasets were created by the authors, who described the methods they used to generate them in their documents. Only 6.9% of the documents that used simulated datasets made them available in packages in the R environment or in the supplementary material of articles. The real-world datasets were generally extracted from the UCI machine learning repository or brought by the authors. In relation to the performance measures, we can affirm that, in general, the authors used more the relative indices to evaluate the clustering algorithms followed by the external and relative ones. The usage of the Akaike information index and its variations, and the Bayesian information index and its variation, stood out. In this branch, only 5 (17.2%) of the articles measured the execution time of the algorithms. Respecting the programming language and code availability, some documents (9, i.e. 31%) provided the code they used in their proposals. This availability was done through packages in the R environment and github website. Many of the articles (9, i.e. 31%) informed the programming language they used to implement the algorithms, such as R (88.9%), and Java (11.1%).

From documents classified into the neural network-based branch, one notes that more than half of the studies (70%) used simulated datasets in their computational experiments, as opposed to those (100%) that used real-world datasets. All the simulated datasets were created by the authors, who described the methods they used to generate them. Most of the real-world datasets were extracted from the UCI machine learning

repository, and some others were brought by the authors. Concerning the performance measures, we can assert that, in general, the articles used more internal indices to evaluate the clustering algorithms followed by external and relative ones. The usage of the entropy validation stood out. In this branch, it is not quite common to measure the execution time of the algorithms. In respect to programming language and code availability, only one work (10%) made the code the authors used in their proposals available on the authors' website. Five documents informed the programming language the authors used to implement their algorithms. In this branch, C++ and Java were the most used languages.

Regarding the works classified into the branch called others, one notes that less than half of the documents (34.4%) used simulated datasets in their computational experiments, as opposed to those (89.6%) that used real-world datasets. Most of the simulated datasets were created by the authors, who described the methods they used to generate them in their documents. Only 3.4% of the documents that used simulated datasets, made them available in online repositories. Most of the real-world datasets were extracted from the UCI machine learning repository. With respect to performance measures, in general, the articles used more external indices to evaluate the clustering algorithms, especially the accuracy and the adjusted Rand index, followed by the internal and relative ones. In this branch, it was not quite common to measure the execution time of the algorithms. Concerning programming language and code availability, we can note that only 2 documents provided the code they used in their proposals. This availability was done through packages in the R environment and on MATLAB website. Many of the articles (14, i.e. 48.3%) informed the programming language they used to implement the algorithms, such as MATLAB (7/14), Java (3/14), R (2/14), Python (1/14), C (1/14).

With respect to the branch called IAPRA, one notes that less than half of the studies (33.3%) used simulated datasets in their computational experiments, as opposed to those (81%) that used real-world datasets. All the simulated datasets were created by the authors, who described the methods they used to generate them in their documents. Most of the real-world datasets were extracted from the UCI machine learning repository. Concerning the performance measures, it was possible to say that, in general, the authors used external indices to evaluate the clustering algorithms followed by the relative and internal ones. The usage of external indices, such as accuracy and the adjusted Rand index, stood out. It was not quite common to measure the execution time of the algorithms. Regarding the programming language and code availability, some documents (5, i.e. 26.3%) provided the code they used in their proposals. This availability was done through packages in the R environment and in the supplementary material of the articles. Only 8 articles (38.1%) informed the programming language they used to implement the algorithms, such as MATLAB (1/8), C++ (1/8), R (6/8), and Java (1/8). Foss et al. (FOSS *et al.*, 2016) used C++ and R languages.

4.7.1 Answers of the questions defined for the individual analysis - Part 3

The following items bring the answers to each question of individual analysis, part 3, defined in Section 4.2.3. Note that the meaning of the acronyms showed in the mentioned figures ahead is defined in the Appendix C.

- **Question DTS-1** - Do the documents use more simulated or real-world datasets? Real-world datasets are more considered (89%) in the computational experiments of the document collection than simulated ones (35.7%), as shown in Figures 36 and 37.
- **Question DTS-2** - What are the sources of the simulated and real-world datasets? With respect to simulated datasets, it is more common the authors describe the method they used to generate these type of datasets and do not share the files. Rarely, simulated datasets are in some repository or package. Regarding real-world datasets, many of them come from the UCI machine learning repository.
- **Question DTS-3** - What are the most used simulated and real-world datasets? With respect to simulated datasets, there is not the most used datasets because the authors generate them in their experiments, thereafter these types of datasets are very specific in each article. Regarding the real-world datasets, Figure 38 shows the five most used real-world datasets in the document collection. As can be observed, the following real-world datasets are the most used in the mixed data clustering documents: CRD-APP (Credit approval, from UCI), ADL-DTS (Adult dataset, from UCI), HRD-DTS (Heart disease dataset, from UCI), STT-HRD (Statlog (Heart), from UCI), and STT-ACA (Statlog (Australian credit approval), from UCI).
- **Question VM-1** - Do the documents more use internal, external, or relative validation measures? Considering all the branches, the documents more use external performance measures (172 occurrences), followed by internal (91 occurrences), and relative ones (91 occurrences).
- **Question VM-2** - What are the most used validation measures? Figures 39, 40, and 41 show the five most used internal, external and relative performance measures, respectively.
- **Question PLC-1** - Is it common for the documents to inform the programming language that was used to carry out the experiments? Figure 42 shows, for each branch, whether the document informed the programming language chosen to implement its algorithms in the computational experiment or not.

- **Question PLC-2** - What is the most used programming language in this area? Figure 44 shows the five most language programming used by the authors, which are R, MATLAB, Java, C++, and Python.
- **Question PLC-3** - Is it common for the documents to make the proposal code available to the community? Figure 43 shows, for each branch, whether the document make the code available or not. We note that 16.9% (i. e. 27) of the document collection make the code of their proposals available, through the following programming languages: R, MATLAB, Java, and others.
- **Question PLC-4** - Where are most of these codes available? Most of them are available in packages developed in R language and deposited in the CRAN repository, github, MATLAB files exchange, and author's website.

Figure 36 – Individual analysis, Part 3 - Simulated datasets per branch

Taking the analysis presented so far into consideration, the following section brings some paths for future research on this topic.

4.8 Challenges and paths for future research

In the following topics, one highlights some challenges and paths for future research on mixed data clustering technique and related fields.

Figure 37 – Individual analysis, Part 3 - Real-world datasets per branch

Figure 38 – Individual analysis, Part 3 - Most used real-world datasets

4.8.1 Bibliometric analysis and individual analysis, part 1

- More long-term specialists: from the bibliography analysis, one can note that this thematic has few specialists that always bringing a new study on mixed data clustering, such as the first authors of the following works Hsu, Jao and Chen (2005), Ahmad and Dey (2007), Ji *et al.* (2012), Hennig and Liao (2013), Foss *et al.* (2016).

Figure 39 – Individual analysis, Part 3 - Internal performance measures

Figure 40 – Individual analysis, Part 3 - External performance measures

The average number of the document per author is 0.395, showing this field requires specialists that might elect this thematic as one of the main research topics in their careers. Long-term specialists tend to acquire more in-depth knowledge in a specific topic, which may make their contributions more relevant to the scientific community.

- More international partnerships: from the bibliography analysis, one can note that few scientific works have resulted from international partnerships. Here, one might have

Figure 41 – Individual analysis, Part 3 - Relative performance measures

Figure 42 – Individual analysis, Part 3 - Programming languages per branch

an opportunity to strengthen this thematic through the cooperation of researchers from different countries, mainly in the application and development of mixed data clustering in real-world international problems.

- Developing more studies on emerging themes: from the bibliography analysis, one can note that there are many topics that need to be developed and explored by the scientific community. One has specifically pointed out the following approaches:

Figure 43 – Individual analysis, Part 3 - Availability of the code

Figure 44 – Individual analysis, Part 3 - Most used programming languages

fuzzy clustering, optimization, and fuzzy systems. Indeed, those topics are extremely important to bring a wide scope and more efficient performance for the mixed data clustering technique and related fields. Some examples of papers returned in our review that dedicated their objective to the fuzzy topic are Sheu (2007), Thong and Son (2016), Davoodi and Moradi (2018).

- The objective of the papers: from the individual analysis, one can note that in some branches, such as the hierarchical, there are many documents that addressed pure application approaches as the goal of the scientific papers. On the other hand, in the model-based, neural network-based, others, and IAPRA branches, there are so many papers whose main objective was to propose a pure theoretical proposal in their works. Here, one can have a special opportunity to contribute with this field, proposing scientific studies that have as objective pure application of some already proposed method to solve a real-world problem, especially in the branches which have so many documents that proposed pure theoretical proposal approaches. The same is valid to the opposite, i.e. proposing theoretical contributions for hierarchical approaches can handle mixed data clustering.
- Strategies to cope with mixed data: from the individual analysis, one can note that so many documents adopted strategy number two to handle mixed attributes. Developing more studies to transform the mixed data into single-type data is very important in this thematic, because, with effective transformation methods the community can use single-type data clustering algorithms to cluster mixed datasets. As one could see, the transformation methods present many drawbacks nowadays, which might be solved with a more in-depth investigation in this regard.
- New algorithms that belong in the branch called others and IAPRA: from the individual analysis, one can note that, in the branch called others and IAPRA, there are few documents that propose algorithms in some types of approaches. One can observe that graph-based and density-based algorithms are quite adapted to cope with mixed datasets; however, other approaches are not so exploring by the scientific community than the methods previously cited, such as spectral clustering (e.g., Niu *et al.* (2015)), semi-supervised algorithms (e.g., Gondek and Hofmann (2007)), and co-clustering methods (e.g., Hummel, Edelmann and Kopp-Schneider (2017)). The same happens in the IAPRA branch, which has many works proposing new dissimilarity measures, but studies with respect to other kinds of preprocessing or post-processing methods, such as new indices, reduction of noisy data, or dimensionality reduction. Some examples of works that returned in our review on these themes are Irigoien and Arenas (2008), Hsu and Huang (2016), Golinko and Zhu (2019).

4.8.2 Individual analysis, part 2

- Dataset scenarios: There are many challenging dataset scenarios, such as large, dynamic, high-dimensional, distributed among others, which are closer to real-world problems than the scenarios considered by the most of articles, i.e. scenarios of centralized, static, small or medium datasets and which do not present high-dimensionality. As can be seen in the individual analysis, there are few studies that

address datasets with such a characteristics. It is important to remind that the individual analysis considered as default the static, centralized, small or medium datasets, and those that did not present high-dimensionality.

- Preprocessing tasks: The development of approaches that reduce the loss of information in the transformation from single-type to mixed-type data can be further explored, as well as noise reduction and outlier detection algorithms. In addition, feature scaling methods are not so explored as well.
- Processing tasks: There is still a lot to explore in terms of improving the data clustering algorithms themselves, such as the development of algorithms that is able to update the groups as new data arrive (dynamic scenario), and find groups in subspaces, besides ensemble approaches, and co-clustering, among others.
- Post-processing tasks: First, the development of validation measures for more specific objectives is very important to this area since the data clustering technique is highly dependent on these measures, and for that, they must be well-defined for specifically problems. In a second demand, the development of visualization approaches, and algorithms that is able to help the interpretation of the groups have been showing a challenge. There are only few documents that directly address this issue.
- Auxiliary approaches: Depend on the branch, there are many opportunities to explore some auxiliary approaches than others. Generally, we can cite a few of them, such as the usage of parallelism, the knowledge of the specialist, competitive learning, and deep learning.

4.8.3 Individual analysis, part 3

- Datasets chosen by the authors: we observed that the authors did not usually share the simulated datasets they used in their experiments. Although they described the procedure that was carried out to generate the simulated datasets, they did not make the files available. It would be a huge contribution if the authors share the files on the supplementary material of the articles. This observation was also made by Ahmad and Khan (2019), who suggested a centralized repository to upload such datasets in order to facilitate future research. We also noted that many real-world datasets, which are in the UCI machine learning repository, are indicated to classification tasks, and not to the data clustering technique. This is worrying because there is not any evidence that the classes of classification datasets are correspondent to clusters. Hence, in our point of view, this is an open question to be explored more deeply.
- Performance measures: it is known that the performance measures are an important and sensitive question in data clustering algorithms, especially the internal ones,

because, most of the time, the specialists do not have the truth clusters (golden truth), therefore needing to use internal measures. However, we noted that not all the articles use internal measures, in fact, most of them chose external criteria to evaluate the quality of the clustering. We also noted that the criteria based on execution time were not taken into consideration by the majority of the authors, even though being important to show the scalability of the algorithms.

- **Programming language and code availability:** we can affirm that few documents informed the programming language the authors used in the computational experiments and made the codes of the proposed algorithms available. It would very be important if the codes could be shared with the community so that the area can develop faster. We think that it is an operational opportunity making the codes of the algorithms that were already been proposed available for the community, through a package in the R environment or in MATLAB, which are programming languages widely used in this thematic.

In the work of Ahmad and Khan (2019), the authors also provided a detailed list of many opportunities on the mixed data clustering technique, some of which as already highlighted above. Hence, to have a richer view of paths for future research, one suggests for the readership the reading of the mentioned work.

4.9 Final remarks

This systematic review provided a broad and in-depth overview of the development mixed data clustering technique, especially through a bibliometric and individual analysis.

The bibliometric analysis provided a holistic view of this theme from six perspectives, which covered information about sources, authors, and documents, besides an analysis of the structure of knowledge, with conceptual, intellectual, and social aspects of the documents. From this, one could answer the questions and bring some path for future research.

The individual analysis has been done in three parts, always analyzing the documents respecting their classification in the branches defined in this chapter. The first individual analysis highlighted, among other information, the objective of the document, on which method it based its proposal, the name of the proposed approach, the strategy to cope with mixed datasets. The second individual analysis shed light on defining the standard representation string for each document and gave a brief description of them. Finally, the third analysis mapped some information with respect to the computational experiments of the documents, such as which datasets the authors used in their studies, the programming language that was chosen, whether the code was made available by the authors or not.

Therefore, this chapter fulfilled its objectives, among which was highlighted some research opportunities in mixed data clustering technique and related topics. This theme offers countless opportunities in its several branches, which can be explored by the scientific community so that the advancement of this technique offers society its benefits. One hopes this review contribute to the development of this theme, and that it encourage more researchers to contribute on it, because the demand for more effective and efficient approaches to mixed data clustering is growing and it comes from relevant areas to the advancement of society, such as health, engineering, finance, among others.

Part IV

Patterns behind optimization

5 FEATURE BALANCING

“There can be three main purposes in studying truth: one, to find it when we seek it; the other, to demonstrate it when we have it; the last, to distinguish it from the false when we examine it (Free translation)”

BLAISE PASCAL, De l’Esprit géométrique.

5.1 Motivation

Feature balancing is the term defined in this thesis to refer to approaches that aim to find a weight to balance the contribution of numerical and categorical features in distance-based mixed data clustering algorithms. The reason for this balance is detailed in the next section. Such weigh has already been considered in some distances for mixed attributes and mixed data clustering algorithms, such as Gower distance and k-prototypes algorithm, respectively. The same problem is indirectly faced when dummy coding is used to transform mixed datasets into single-type ones, because the choice of the constant c , usually considered as $c = 1$, in an arbitrary way, represents the weight that is applied to the set of categorical features to balancing the contribution of both types of attributes.

Foss *et al.* (2016), Foss and Markatou (2018), Foss, Markatou and Ray (2019) show how difficult is to determine the correct weight in such approaches since this definition is highly dependent on the characteristics of the dataset and the data clustering algorithm chosen by the user. Modha and Spangler (2003) also addressed this thematic and proposed a brute-force method that searches a good weight from a range of values provided as a parameter.

Although the mentioned works came this issue to light, they did not provide appropriate solutions to this problem, especially using the optimization methods that are available nowadays. These methods are able to find good solutions for this problem considering the specificities of the dataset and data clustering algorithm, besides carrying out the search in a good manner, since the brute-force demands a considerable execution time.

Hence, this chapter aims to solve the feature balancing problem in distance-based data clustering algorithms using an evolutionary optimization method, called Biased Random-Key Genetic Algorithm (BRKGA). The objective involves the study of the Gower distance, the k-prototypes algorithm, and the dummy coding.

The outline of this chapter is as follows. In Section 5.2, one presents basic concepts that involves feature balancing. Section 5.3 describes some studies that addressed the problem of balancing the contribution of the numerical and categorical attributes in the

distance-based mixed data clustering algorithms. In Sections 5.4 and 5.5, one presents the proposed optimizer and the proposed algorithms, respectively. Section 5.6 presents the methodology of this proposal. Sections 5.7 and 5.8 present the results and discuss them, respectively. Finally, Section 5.9 concludes this chapter.

5.2 Main concepts

In the following topics, one describes the main concepts surrounding the feature balancing.

5.2.1 Differences among balancing, weighting and selection

This subtopic is dedicated to express the understanding this thesis presents with respect to feature balancing, weighting, and selection in data clustering algorithms, because one has been observing that one might confuse these activities. However, before that, let us define the following scenario that our thoughts are valid.

One considers the types of attributes can contribute in an unbalanced way for the dissimilarity function when the distributions of the distance values of each type of attribute are very different. An example can be given in Table 5. As already explained the possible values for the dissimilarity of the binary vectors is zero *or* one, whereas the values for the dissimilarities of the first and second attributes might vary from 0 to 1 if the min-max normalization has been performed or might vary from the upper to the lower bound if other normalization procedure has been carried out. In this situation, it is clear that both types of features will contribute differently with the objective function of distance-based data clustering algorithms only by the fact they pass through different ways to make the dissimilarity being calculated.

Thus, in our point of view, when one addresses the feature balancing problem, the aiming is regarding assigning a unique weight into the entire set of the categorical attributes in order to minimize the average difference of the distance values of both types of attributes. In this balancing situation, it does not have reasons to assign different weights for *each* categorical attribute, neither to the numerical ones, because if it is done, the task will be not the balancing, but the weighting since the weights will emphasize, not balance, the contribution of each categorical features in the clustering process. When we are handling feature weighting strategies, we are looking for a set of weights of each attribute in the dataset in order to highlight the clustering structure that might be covered by irrelevant features. Finally, one addresses the feature selection methods when it aims to make a binary choice of which attributes must stay in the clustering process, eliminating (and not weakening the contribution) those that are highly irrelevant or redundant and maintaining those whose contribution is really relevant. Hence, we can claim that feature

selection is a specification of feature weighting, that in its turn, is a generalization for feature balancing.

Undoubtedly the feature balancing might emphasize the contribution of the weighted features depending on the number and the quality of each type of features present in the dataset, but since this strategy assigns only one weight applied into the entire set of the same type of features, in an aggregate way, the individual emphasis is minimized, giving a place the balancing goal. We believe that is important to make this clearer because it is quite common to note works comparing feature balancing methods to feature weighting approaches (e.g., Foss, Markatou and Ray (2019, p. 91), and Modha and Spangler (2003, p. 2648)), and, as they have different objectives, this comparison is not fair.

With this in hand, it is possible to conclude that feature balancing is an exclusive problem of distance-based methods, whereas feature weighting and feature selection not, because these methods can be applied in all data clustering algorithm classes. An example is the feature weighting strategy proposed in Martarelli and Nagano (2019) for KAMILA algorithm, where the authors modeled an optimization algorithm to choose the quasi-optimum weights for each attribute in order to emphasize the contribution of relevant features and weakness the contribution of the irrelevant ones. The authors obtained excellent results with their strategy, statistically tested, in comparison to the KAMILA algorithm with the default setting, i.e. with all the weights equal to one. As KAMILA algorithm is a model-based proposal, only feature weighting or feature selection makes sense, since model-based approaches already balance the contributions of both types of attributes. More about KAMILA algorithm can be found in Section 5.3.

5.2.2 Feature balancing relies on the numerical data normalization method

In subsection 2.2 from Foss *et al.* (2016), the authors showed the expectation of the contribution of numerical and categorical features to the distance between two previously defined populations, which were built by mixed data values (continuous and binary vectors). In their proof, the authors considered that numerical attributes had been normalized by the z-score method, and thereafter, they claim that the expectation of the contribution of the numerical features is higher than the categorical ones, concluding the weight that is applied to the set of the categorical attributes might composed of values more than 1.

However, if the numerical attributes are normalized by the min-max method, the range of the distance values will vary from 0 to 1. Therefore, in min-max normalization, the weight applied to the set of categorical attributes, which will have as distance values only the number zero or one, must be defined in the $[0,1]$ interval. Thus, this weight will minimize the difference between the average distance values from the numerical attributes and the categorical ones.

As could be seen, the choice of the normalization procedure for the numerical features defines the solution space of the weight in the feature balancing problem. In this chapter, we adopted the min-max normalization, because this method does not infer about the distribution of the numerical attributes, and because of this it is more comprehensive, and as already pointed out in other scientific works, min-max normalization is preferred when one handles mixed datasets (see more in Larose (2005, p. 101)). Although Suarez-Alvarez *et al.* (2012, p. 2636) highlighted that the min-max normalization “does not achieve equalization of the attributes means” here we are not interested in equalizing the mean of the attributes, we are concerned about the contribution each type of them will have in the dissimilarity function.

5.2.3 Why does one improve distance-based algorithms?

If you are new in the data clustering technique, you might be thinking why it is important to improve distance-based algorithms since the model-based methods do not face the feature balancing problem, and thereafter can cluster the dataset effectively as KAMILA algorithm does. In fact, it is an excellent question to make. The answer is given by the comprehension of what means clustering data. Hennig (2015) provides a philosophical and comprehensive discussion about the meaning of clustering in which the author reflects about the *truth* in cluster analysis, and concludes that it depends on the context and the objectives of the clustering. The author claimed the cluster analysis does not have the same meaning for everyone, as can be seen as follows (HENNIG, 2015, p. 53).

The term “cluster” does not mean the same to all researchers in all situations. This is acknowledged in general overviews and books about cluster analysis, but seems to be ignored by many authors of specialist work who try to convince readers that a certain method is best for finding the “true/natural/real” clusters.

In line with some other experts in cluster analysis, the author affirms that there is not any the best method nor the best class of algorithm that can be used in all the situations, although many researchers insist on defending one or another class of clustering algorithms. Hennig (2015, p. 8) reiterates that the analyzer needs to reflect on the data, the proprieties of the dissimilarity measures, and the main characteristics of the clusters it is looking for or expecting to. Such a reflection will support the proper choice of a clustering algorithm.

Hence, it is known that distance, model, density, and graph-based data clustering methods have different aims, advantages, and drawbacks. Improving whichever class of the clustering algorithms is the same as contributing to the advance of this thematic, which is only in its beginning (SUAREZ-ALVAREZ *et al.*, 2012; HENNIG; LIAO, 2013; HENNIG; LIN, 2015; FOSS; MARKATOU; RAY, 2019).

5.3 Related works

In this section, we describe six works that addressed the feature balancing problem in data clustering algorithms. We firstly introduce the work of Modha and Spangler (2003), which proposed a brute-force method to find a good weight to solve the feature balance issue. Next, we briefly describe a normalization procedure for mixed datasets, which can be used to balance the contribution of both types of features. Another method is the `clustMixType` algorithm, an approach to automatically define the λ value in the k-prototypes algorithm. Finally, we describe three papers of Foss, which the first proposed KAMILA algorithm, the second introduced the Kamila package in R language, and the last made a comprehensive discussion regarding distance metrics and clustering methods for mixed datasets.

Modha and Spangler (2003) proposed the feature weighting strategy for the k-means algorithm, optimizing a generalized Fisher ratio, which consists of a division between the average within-cluster distortion (dissimilarity) and the average between-cluster distortion. The authors proposed a heuristic that carries out a brute-force search into a space that is restricted by the user-defined parameter. Then, the authors deliver the so-called optimum weights for each type of attribute in a mixed data clustering context. The term *so-called optimum weights* was used to indicate that their proposal aimed to find the optimum weight, although the method does not do this, because, by being a heuristic that limits the solutions' domain, it can not guarantee that the optimum weights will be found. As the authors pointed out at the end of their article, the method needs to be improved using more efficient strategies to search the weights. In our method (see in Section 5.4), we implement efficient choices with respect to optimization strategies to the feature balancing problem that improves the Modha and Spangler idea as can will be seen in Section 5.7.

Suarez-Alvarez *et al.* (2012) proposed a normalization method for mixed datasets with the objective of equalizing the contribution of the features in distance-based clustering algorithms by transforming the mixed vector in order that the mean of each attribute could be the same. The authors claimed that until that moment there was not any unbiased and consistent approach that normalized this kind of dataset. Their proposal, which was built considering the Minkowski metric for numerical attributes and the simple matching dissimilarity for the categorical features, multiplies a specific weight to each feature, where these weights were calculated as being the inverse of the dissimilarity expectation for that attribute. In the experiment step, the authors compared the accuracy of the k-prototypes algorithm with and without the normalization procedure, taking into consideration datasets extracted from the UCI machine learning repository. As a result, they obtained good accuracy values from the normalized datasets. Prostov, Suarez-Alvarez and Prostov (2015) is an extension of Suarez-Alvarez *et al.* (2012), because the authors proposed the same normalization procedure, providing, in this paper, a mathematical justification of the

statistical normalization method proposed in Suarez-Alvarez *et al.* (2012), especially with regard to the estimators of the dissimilarity expectation used to calculate the weights.

Szepannek (2018) proposed an automatic method to determine the λ in the k-prototypes algorithm, called clustMixType algorithm. They set the weight for categorical attributes as

$$\lambda = \frac{\sigma^t}{h_{cat}}, \quad t \in 1, 2 \quad (5.1)$$

where, σ represents the standard deviation of the numerical attributes, t defines whether the λ will consider the average of the standard deviation ($t = 1$) or the variance ($t = 2$) of the numerical attributes, and h_{cat} is calculated as the average of $h_m = 1 - \sum c p_{mc}^2$ for $t = 1$, and $h_m = 1 - \max_c \{p_{mc}\}$ for $t = 2$, where c are the categories of the categorical attributes, m is the amount of all variables, and p is the probability. As default, the method considers $t = 1$.

The next three articles that we will describe ahead were written in the same general ideas, and although the first two proposed a model-based data clustering algorithm, the authors made an in-depth exploration about feature balancing problem, showing the performance of some seminal algorithms, such as the k-means, k-prototypes, and partition around medoids (PAM), in mixed datasets using the dummy coding and the Gower distance. These were the reasons we selected these papers.

The first article is Foss *et al.* (2016) in which the authors proposed a semi-parametric model-based mixed data clustering algorithm, called KAY-means for MIXed LARge data (KAMILA). Among the methods selected for the comparison, they elected the Modla-Spangler algorithm (MODHA; SPANGLER, 2003), k-prototypes (HUANG, 1998), weighted k-means (HARTIGAN; WONG, 1979) with dummy coding, and PAM algorithm with dummy coding and Gower distance. The authors found the PAM algorithm was unduly influenced by dummy coding and the Gower distance, whereas the k-means did not have the same sensibility. In addition, they noted that when one fixed the weights that obtained good results in one scenario and applied this weighting setting to other datasets, the performance of the weighted-based algorithm suffered, evidencing that the weights must be defined considering the characteristics of each dataset. In our proposed method, the weights are defined for each dataset, therefore, such a challenge will not be a question in our algorithms.

The second article is Foss and Markatou (2018). In this paper, the authors implemented the KAMILA algorithm and made the code available through an R package (FOSS; MARKATOU, 2018), besides describing the KAMILA package, and provided many experiments comparing the semi-parametric algorithm to some distance-based approaches that relied on a weight set. In addition, the authors showed a way to adapt the KAMILA

algorithm to a map-reduce framework, implementing the algorithm using Hadoop. Besides the finding from Foss *et al.* (2016), in this paper, the authors claimed that in some scenarios, Modla-Spangler method was more appropriate when the datasets have small sizes, specifically in well-separated clusters and with categorical sparsity, i.e. when the dataset size was small relative to the number of categorical levels.

The third article is Foss, Markatou and Ray (2019). In this paper, the authors proposed a wide discussion with respect to distance metrics and data clustering algorithms for mixed datasets. They described data transformation methods, highlighting the discretization approaches and numerical coding mechanisms, and shed light on the difficulty of properly defining the weights to balance the contribution of both types of features, simulating the behavior of the distance ratio while some parameters changed. These parameters were regarding the overlap level of the numerical and categorical attributes had in the simulated datasets and the weight values in weight-based approaches. The authors highlighted the sensibility of the k-means algorithm with the dummy coding method when the overlap level in numerical attributes was high, and its insensibility when categorical features were not informative. They also concluded that when the assumptions regarding the data distributions are severely violated, maximum-likelihood estimation using the expectation-maximization (EM) algorithm of a normal-multinomial mixture model with equal spherical covariance structure had a good performance, although it was outperformed by k-means using Hennig-Liao weighting (HENNIG; LIN, 2015) and Modha-Spangler clustering (MODHA; SPANGLER, 2003). Concerning the choice of the weights, Foss *et al.* (2016), Foss and Markatou (2018), Foss, Markatou and Ray (2019) concluded that it was too hard to define them manually, and even normalization methods could not handle the feature balancing problem because they were "acceptable in certain restricted settings".

Although the KAMILA method achieved a good performance in some comparison tests, this approach requires some attention points. First, one of the assumptions the authors established is independence among the categorical variables given the cluster membership. If this condition is not respected, then they say that is necessary the creation of a new categorical variable resulted from the combination of dependent features. Second, if the numerical and categorical attributes are not conditionally independent given the cluster membership, the authors suggest using the location model, which models the numerical variables as being multivariate normal with unique parameters for each cluster and class from categorical attributes. As highlighted by the authors, the drawbacks of the location model become this option intractable as the number of features increases.

5.4 Proposed optimizer

In this section, we detail how we adapted each step of the BRKGA to solve the feature balancing problem on mixed data clustering. We called the proposed optimizer

Unsupervised Feature Balancing for Mixed Datasets through the Biased Random-Key Genetic Algorithm (UFBMD-BRKGA), and its pseudocode is shown in Algorithm 1.

Algorithm 1: Unsupervised Feature Balancing for Mixed Datasets through Biased Random-Key Genetic Algorithm

```

Define the parameters  $npop$ ,  $p_e$ ,  $p_m$ ,  $\rho_s$ ,  $maxIter$ ;
Build the initial population;
Evaluate each individual;
while the stopping criteria are not reach do
    Divide the population into elite ( $p_e$ ) and non-elite groups ( $1 - p_e$ );
    Build the next population, copying elite group (TOP), generating mutant individuals
    (BOT), and creating new individuals by the crossover procedure;
    Evaluate each individual;

```

5.4.0.1 Initial population, coding, decoding scheme, and fitness function

We code the solutions as follows. Each individual is a binary representation of two numbers that vary from 0 to 9. The binary representation has always four positions, therefore to represent the two numbers the string has eight binary positions. The first four are transformed into a decimal number, and it is multiplied by 10. The other four positions are also transformed into decimal representation and summed to the result of the first number, resulting in a number from 0 to 99. Let pW be the mentioned four-binary-positions number. Because the binary representation has four positions that might happen that pW is more than 10. In these cases, we make that pW be equal to 10. We also define a factor, which can be equal to 100, 1.000, or 10.000, which divides the final decimal number, resulting in the weight that will be assigned to a distance-based mixed data clustering algorithm. Hence, the individual is composed of 9 positions, where the first eight is regarding the two binary representation numbers and the ninth is the factor. Figure 45 shows an example of the decoding process of an individual.

With respect to generating the initial population, we build it creating the most part of the individuals randomly, and only five deterministically. The random part is generated by using the Bernoulli distribution to fill each position corresponded to the binary representation with the same probability for zero and one. We also randomly generate the factor that we divide the pW value. In this case, we consider the following probabilities, 70%, 20%, and 10%, for each possible value of this factor, which are 100, 1.000, or 10.000, respectively. Note that the number of individuals in the initial population is defined by a parameter $npop$.

Concerning the deterministic part, it is generated according to the following strategies. The first deterministic individual is determined as being equal to the the

Figure 45 – Decoding process of an individual

average of all weights assigned to the categorical attributes obtained from the mixed normalization procedure proposed by Suarez-Alvarez *et al.* (2012), divided by the sum of the average of all weights of the numerical features and the average of the weights of the categorical features obtained from the same normalization procedure. We code this weight following the coding rules we mentioned in the first paragraph of this section. The second deterministic individual is the weight calculated in the work of Szepannek (2018). The third individual represents the average of numerical attributes, and the next two are 0.1, and 0.01.

As the fitness function, this method considers the external performance measure for clustering algorithms, called adjusted Rand index (ARI) (HUBERT; ARABIE, 1985). Because we will use only simulated datasets in which we already know the number of groups and the belonging of each register to the groups we can use this kind of index, as it will be seen in Section 5.6. The closer the ARI index values are to 1, the higher the quality of the clustering, as opposed to a ARI value equal to zero, which represents a poor clustering quality.

5.4.0.2 Selection, crossover, mutation, and stopping rule

The first step consists of classifying the individuals from the current population in descending order by their fitness values, thereafter one divides the population into two groups, wherein the former, called the elite, receives the p_e (percentage) of the best individuals, and the second one, called non-elite, receives the remaining individuals. Next, a new population is built in three steps by copying the full elite group from the old population to the newest one, renaming it as the TOP group; creating mutant individuals (p_m percent of the entire population), calling these individuals the BOT group; and generating the remaining individuals by the crossover procedure.

The mutant individuals are created as follows. The worst individual from the previous population is selected; thereafter, a random position of this individual is chosen to be substituted by a value generated randomly following the same rules we respected to build the random part of the initial population. The same procedure is executed for all mutant individuals, updating the worst coded solutions to the previous second-worst individual, and so on.

The individuals resulting from the crossover procedure are built as follows. Two individuals (parent 1 and parent 2) are randomly selected from the previous population so that one comes from the TOP group (parent 1), and the other is selected from the entire population (parent 2). The new individual (offspring) is built with the help of an auxiliary string, which has the same size as an individual. Each position (allele) of this string, which is filled with numbers from the $[0,1]$ continuous interval, and is compared to a predetermined parameter, called ρ_s . Thus, the parent 1 allele is selected to compose the child individual if the corresponding allele of the auxiliary string is smaller or equal to ρ_s . Otherwise, the corresponding allele from parent 2 feeds the child sequence. It is noteworthy that in both the mutation and in the crossover stage, we allow only feasible solutions.

The UFBMD-BRKGA stops when it achieves *maxIter* iterations or when its improvement of quality is less than *minInc* after *minIt* successive generations.

5.5 Proposed algorithms

We use the UFBMD-BRKGA algorithm in the following approaches in order to assign good weights for each method.

- K-means with dummy coding: The optimizer is used to define the value of c from the dummy coding. We use the *kmeans* function from the *stats* package to carry out the k-means algorithm proposed by Hartigan and Wong (1979);
- Weighted k-means with dummy coding: The optimizer is used to define the weights of the set of numerical and categorical attributes. In this case, the weight of numerical attributes is defined to be $1-wCat$, where $wCat$ is the weight of categorical features. We use the *wkmeans* function from the KAMILA package to carry out this algorithm;
- Partitional Around Medoids with Gower distance: The optimizer is used to define the weights of the set of numerical and categorical attributes. In this case, the weight of numerical attributes is defined to be $1-wCat$, where $wCat$ is the weight of categorical features. We use the *daisy* function from the *cluster* package to calculate the dissimilarity matrix and the *pam* function from the same package to carry out the PAM algorithm;

- K-prototypes: The optimizer is used to define the lambda value from the k-prototypes algorithm. We use the *kproto* function from the *clustMixType* package to carry out this algorithm;

5.6 Methodology

In this section, we describe the experiments we designed to understand the behavior of the proposed methods and the approaches we selected from the literature. We also present the methods we used to generate the datasets, the methods we selected to compare the proposed algorithms, the manner we defined the parameters of the methods for each set of experiments, the performance measure we chose to evaluate the algorithms, and the computational setting we used.

5.6.1 Experiments

The following experiments were designed to obtain a better understanding of the algorithms when they cope with different characteristics of datasets, such as when the dataset has more numerical than categorical attributes, or vice-versa; when the number of numerical and categorical attributes are balanced; when the number of attributes increases 10 times; when the number of registers increases; when a different number of attributes is overlapped; when the number of levels from categorical attributes increases, and when the data are generated from normal and non-normal distributions. Therefore, we designed four main experiments (from A to D), as described in the following topics.

- Experiment A: in this experiment, we aimed to observe the effects, with respect to the performance of the algorithms, a dataset that has unbalanced number of numerical and categorical attributes produces. We also objective understanding the behavior of the methods when the number of all attributes are ten times higher and when the data are generated by the normal and a non-normal distribution.
- Experiment B: in this experiment, we aimed to observe the effects, with respect to the performance of the algorithms, a dataset that has balanced number of numerical and categorical attributes produces. We also objective understanding the behavior of the methods when the number of all attributes are ten times higher and when the data are generated by the normal and a non-normal distribution.
- Experiment C: in this experiment, we aimed to understand the behavior of the methods when the number of levels from the categorical attributes increases in balanced and unbalanced datasets, and when the data are generated by the normal and a non-normal distribution.
- Experiment D: in this experiment, we aimed to understand the behavior of the methods, with respect to the performance of the algorithms and computation time,

when the size of the datasets increases both in the number of attributes and in the number of instances.

We elected the lognormal with skew equal to 9 as the non-normal distribution for all experiments, which is correspondent to a log standard deviation of 1.131. Concerning the specific characteristics of each experiment, Table 22 brings all the information.

Table 22 – Details of the experiments

Experiment	Number of instances	Distribution		Attributes combination (numerical/categorical)						Number of levels from categorical attributes		Percentage of overlapping		
		Normal	LogNormal	Balanced			Unbalanced			2	10	0%	30%	70%
				2/2	20/20	200/200	2/8	8/2	20/80	80/20				
A1	250	x					x	x	x	x		x	x	x
A2	250		x				x	x	x	x		x	x	x
B1	250	x		x	x							x	x	x
B2	250		x	x	x							x	x	x
C1	250	x		x			x	x			x	x	x	x
C2	250		x	x			x	x			x	x	x	x
D1	250, 500, 750, 1000, 2000, 3000	x		x							x			
D2	250	x		x	x	x					x			

All the datasets were generated using the *genMixedData* function from Kamila package. To generate the datasets that considered the logNormal distribution, we adapted this function. All the simulated datasets have two groups, and 50% of the attributes in each type have the defined percentage of overlapping. The function also requires, at least, one numerical and one categorical attribute with 0.01% of error level (overlapping). However, to generate the experiments that require 0% of overlapping, we added one more attribute on the generation function and deleted that attribute that had 0.01%, obtaining a dataset that has effectively 0% of attributes with overlapping.

5.6.2 Methods

Table 23 presents all the algorithms we consider as baseline, literature-proposed, and proposed in this paper.

In order to evaluate those algorithms fairly, we cluster them in four groups, as shown in Table 24. Note that the statistical test and all comparisons we will make in the result and discussion section is regarding each of those groups, separately. To generate the results, all methods were carried out 100 times.

5.6.3 Parameters definition

Some methods of those presented in the previous section require parameters defined by the users. As known, such definition can benefit one or other methods if the values of the parameters do not be chosen properly. Because of this, it is very important to use techniques that automatically define the parameters facing different instances, such as the design of experiments methods, and race algorithms.

Table 23 – Details of the experiments

Method	Function package ¹	Baseline		Literature-proposed		Proposed	
		Name	Configuration	Name	Configuration	Name	Configuration
k-means with dummy coding	<i>kmeans</i> function	KMDBS	$c = 1$			KMDBR	c defined by UFBMD-BRKGa
	from <i>stats</i> package						
	<i>dummyCodeFactorDf</i> function from <i>kamila</i> package						
Weighted k-means with dummy coding	<i>wkmeans</i> function	WKMBs	$wCon = 0.5$ and $wCat = 0.5$	WKMMs	weights defined by Modha and Spangler (2003)	WKMBR	weights defined by UFBMD-BRKGa
	<i>gmsClust</i> function						
	from <i>kamila</i> package						
PAM with Gower distance	<i>daisy</i> function	PMGBS	$wCon = 0.5$ and $wCat = 0.5$	-	-	PMGBR	weights defined by UFBMD-BRKGa
	<i>pam</i> function						
	from <i>cluster</i> package						
k-prototypes	<i>kproto</i> function	KPTBS	$\lambda = NULL$ $\lambda = 1$	KPTNM KPTeL	λ defined by Szepannek (2018) dataset normalized by Suarez-Alvarez <i>et al.</i> (2012)	KPTBR	λ defined by UFBMD-BRKGa
	from <i>clustMixType</i> package						
Mixed Gaussian multinomial mixtures	<i>flemixedruns</i> function from <i>fpc</i> package	FMMMD	default	-	-	-	-
KAMILA	<i>kproto</i> function from <i>kamila</i> package	-	-	KAMILA	default	-	-

Table 24 – Groups of methods for comparisons

Group	KAMILA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTeL	KPTBR
1 (K)	x	x	x	x									
2 (W)	x	x			x	x	x						
3 (P)	x	x						x	x				
4 (Y)	x	x								x	x	x	x

Therefore, in this chapter, we defined four parameters of the UFBMD-BRKGa algorithm ($p_e, p_m, \rho, npop$), and the *searchDensity* of the Modha-Spangler algorithm through the *irace* package (LÓPEZ-IBÁÑEZ *et al.*, 2016), considering the elite option as true, and the maximum of iteration as 1.000. We considered the following range of values for each parameter: (i) $npop$: from 9 to 15; (ii) p_e : from 0.1 to 0.9 (iii) p_m : from 0.1 to 0.9 (iv) ρ : from 0.1 to 0.9 (v) *searchDensity*: from 2 to 20. The following parameters we set as fixed: $maxIter = 10$, $minInc = 5$, and $minIt = 0.01$. In Appendix D (Table 72), we provide a table with all the values defined by the *irace* method for each parameter in each experiment.

5.6.4 Performance measure and statistical test

As we are working with simulated datasets, we evaluate the performance of all the algorithms with ARI measure. We used the R package called *mclust* to measure the ARI values. As the statistical test, we elected the Wilcoxon rank-sum test to evaluate whether there was a relevant difference between the best result of an algorithm in a specific dataset and the other algorithms. For that, we used the *wilcox.test* function from *stats* package considering one-side (greater option), with a confident level of 95%.

5.6.5 Computational setting

The computational experiments were carried out in R language, on a personal desktop computer with Intel Core i7-4790K CPU @ 4 GHz along with 16 GB of RAM under Windows 10.

5.7 Results

In this section, we present the results we obtained in each experiment. Tables from 25 to 32 present the average of ARI values and the standard deviation of each measure of the experiments A1 to D2, respectively. Note that bold values are statistically better than the others in the same group. When KMILA or FMMD values are in bold, it means that the respective algorithm is the best algorithm of all methods of the four groups, and when one of both methods has a letter as an exponent, it means that the respective algorithm is better than the group represented by that letter, where group 1 is represented by letter the "K", group 2 by the "W", group 3 by the "P", and group 4 by the "Y". Also note that the first column provides the information about some characteristics of the dataset, which the first two-element that compose the string means the number of numerical and categorical attributes, respectively, and after the backslash, the two elements represent the percentage of overlapping in the numerical and categorical attributes, respectively.

In Appendix D, we provide the information about the average and the standard deviation of the weights assigned to the numerical and categorical attributes in each method, and the average and the standard deviation values of the execution time of each algorithm (Tables from 73 to 106).

5.8 Discussion

As one can be observed in the tables presented in the previous section, the ARI average values were very different in each experiment, showing the mixed data clustering algorithms have an expressive sensibility with respect to the characteristics of the datasets. In the following topics, we discuss the results of each experiment.

5.8.1 Experiment A

Concerning experiment A1, we can note that when the number of attributes increases ten times, all algorithms perform very well, achieving the highest ARI value in all repetition, except in two datasets handled by the k-prototypes algorithm. However, when the datasets are small regarding the number of attributes, we note that the mixed Gaussian multinomial mixtures algorithm achieved the best result in the groups P and Y when the dataset considered 30% of overlapping in the numerical attributes and 70% in

Table 25 – Results of experiment A1 - Normal distribution, 250 instances and two levels in each categorical attribute

Dataset scenario	All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTBL	KPTBR
02-08/0.0-0.0	1 ±0.002	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
02-08/0.0-0.3	1 ±0.002	1 ±0	1 ±0.002	1 ±0	1 ±0.002	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
02-08/0.0-0.7	1 ±0.002	1 ±0	1 ±0.002	1 ±0	1 ±0.002	1 ±0	1 ±0	0.999 ±0.003	1 ±0	0.999 ±0.003	1 ±0.002	1 ±0	1 ±0
02-08/0.3-0.0	0.999 ±0.004	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
02-08/0.3-0.3	0.999 ±0.004	1 ±0	1 ±0.002	1 ±0.002	1 ±0.002	1 ±0.002	1 ±0.002	1 ±0	1 ±0	1 ±0.002	1 ±0	1 ±0.002	1 ±0.002
02-08/0.3-0.7	0.999 ±0.005	1 ^{PY} ±0	1 ±0.002	1 ±0.002	1 ±0.002	1 ±0.002	1 ±0.002	0.999 ±0.004	1 ±0.003	0.997 ±0.007	0.999 ±0.004	0.997 ±0.007	0.997 ±0.006
02-08/0.7-0.0	0.999 ±0.004	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
02-08/0.7-0.3	0.997 ±0.008	1 ±0	1 ±0.002	1 ±0.002	1 ±0.002	1 ±0.002	1 ±0.002	1 ±0	1 ±0	1 ±0.003	1 ±0.002	0.999 ±0.004	1 ±0.003
02-08/0.7-0.7	0.997 ±0.008	1 ^{PY} ±0	0.99 ±0.096	1 ±0.002	0.99 ±0.096	1 ±0.002	1 ±0.002	0.999 ±0.004	0.999 ±0.003	0.975 ±0.1	0.998 ±0.005	0.98 ±0.016	0.985 ±0.014
08-02/0.0-0.0	1 ±0	1 ±0	1 ±0.002	1 ±0	1 ±0.002	1 ±0.002	1 ±0	1 ±0	1 ±0	1 ±0.002	1 ±0	1 ±0	1 ±0
08-02/0.0-0.3	1 ±0	1 ±0	0.997 ±0.007	1 ±0	0.997 ±0.007	1 ±0	1 ±0	1 ±0	1 ±0	0.997 ±0.007	1 ±0	1 ±0	1 ±0
08-02/0.0-0.7	1 ±0	1 ±0	0.746 ±0.402	1 ±0	0.746 ±0.402	1 ±0	1 ±0	1 ±0	1 ±0	0.994 ±0.01	1 ±0	1 ±0	1 ±0
08-02/0.3-0.0	1 ±0.002	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
08-02/0.3-0.3	0.999 ±0.003	1 ±0	0.933 ±0.151	1 ±0	0.933 ±0.151	0.998 ±0.006	1 ±0	1 ±0.003	1 ±0	0.998 ±0.007	1 ±0	1 ±0	1 ±0
08-02/0.3-0.7	1 ±0.003	1 ±0	0.717 ±0.405	1 ±0	0.717 ±0.405	0.997 ±0.008	1 ±0	0.999 ±0.004	1 ±0	0.993 ±0.01	1 ±0	1 ±0.002	1 ±0.002
08-02/0.7-0.0	0.988 ±0.014	1 ^{KW} ±0	0.996 ±0.008	0.998 ±0.005	0.996 ±0.008	0.997 ±0.006	0.998 ±0.006	1 ±0	1 ±0	0.998 ±0.006	1 ±0	0.945 ±0.037	0.998 ±0.005
08-02/0.7-0.3	0.977 ±0.034	1 ^{KW} ±0	0.847 ±0.189	0.989 ±0.014	0.847 ±0.189	0.984 ±0.018	0.984 ±0.016	0.999 ±0.005	1 ±0	0.959 ±0.027	1 ±0	0.904 ±0.046	0.961 ±0.026
08-02/0.7-0.7	0.974 ±0.062	1 ^{KW} ±0	0.725 ±0.397	0.982 ±0.016	0.725 ±0.397	0.981 ±0.018	0.982 ±0.016	0.995 ±0.009	1 ±0	0.891 ±0.049	1 ±0.002	0.828 ±0.105	0.897 ±0.046
20-80/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
20-80/0.0-0.3	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
20-80/0.0-0.7	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
20-80/0.3-0.0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
20-80/0.3-0.3	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
20-80/0.3-0.7	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
20-80/0.7-0.0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
20-80/0.7-0.3	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
20-80/0.7-0.7	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
80-20/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
80-20/0.0-0.3	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
80-20/0.0-0.7	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
80-20/0.3-0.0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
80-20/0.3-0.3	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
80-20/0.3-0.7	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	0.99 ±0.1	1 ±0	1 ±0	1 ±0
80-20/0.7-0.0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
80-20/0.7-0.3	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
80-20/0.7-0.7	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	0.99 ±0.1	1 ±0	1 ±0	1 ±0

the categorical features, and when the overlapping was 70% in both types of attributes. The same algorithm also achieved the highest ARI value in the groups K and W when the overlapping of the numerical attributes was 70% and the overlaps of the categorical features were 0%, 30%, and 70%. The remaining methods had a close performance, and the statistical test did not indicate any difference.

Experiment A2 had similar behavior to experiment A1 in relation to when the number of attributes increases, i.e. the performance of the algorithms increases expressively

Table 26 – Results of experiment A2 - LogNormal distribution, 250 instances and two levels in each categorical attribute

Dataset scenario	All groups		Group 1 (K)			Group 2 (W)			Group 3 (P)		Group 4 (Y)			
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTBL	KPTBR	
02-08/0.0-0.0	0.71	0.994	1	1	1	1	1	1	1	0.88	0.869	0.88	1	
	±0.456	±0.013	±0	±0	±0	±0	±0	±0	±0	±0.328	±0.339	±0.328	±0	
02-08/0.0-0.3	0.639	0.984	1	1	1	1	1	0.999	1	0.988	0.988	0.968	0.998	
	±0.482	±0.024	±0.002	±0.002	±0.002	±0.002	±0.002	±0.004	±0.003	±0.1	±0.1	±0.172	±0.005	
02-08/0.0-0.7	0.512	0.961	0.97	1	0.97	1	1	0.964	0.971	0.953	0.952	0.943	0.965	
	±0.499	±0.102	±0.172	±0.002	±0.172	±0.002	±0.002	±0.025	±0.021	±0.099	±0.1	±0.138	±0.022	
02-08/0.3-0.0	0.73	0.993	1	1	1	1	1	1	1	0.909	0.9	0.88	1	
	±0.446	±0.014	±0	±0	±0	±0	±0	±0	±0	±0.289	±0.303	±0.328	±0	
02-08/0.3-0.3	0.629	0.971	1	1	1	1	1	0.999	1	0.968	0.988	0.968	0.998	
	±0.485	±0.101	±0.002	±0.002	±0.002	±0.002	±0.002	±0.005	±0.003	±0.172	±0.1	±0.172	±0.006	
02-08/0.3-0.7	0.56	0.938	0.99	1	0.99	1	1	0.964	0.969	0.945	0.945	0.925	0.965	
	±0.498	±0.168	±0.096	±0.002	±0.096	±0.002	±0.002	±0.022	±0.021	±0.138	±0.138	±0.192	±0.023	
02-08/0.7-0.0	0.71	0.993	1	1	1	1	1	1	1	0.92	0.9	0.89	1	
	±0.456	±0.014	±0	±0	±0	±0	±0	±0	±0	±0.274	±0.303	±0.316	±0	
02-08/0.7-0.3	0.639	0.971	1	1	1	1	1	0.999	1	0.988	0.988	0.978	0.998	
	±0.482	±0.101	±0.002	±0.002	±0.002	±0.002	±0.002	±0.005	±0.003	±0.1	±0.1	±0.141	±0.006	
02-08/0.7-0.7	0.569	0.935	0.99	1	0.99	1	1	0.964	0.969	0.944	0.954	0.934	0.965	
	±0.497	±0.169	±0.096	±0.002	±0.096	±0.002	±0.002	±0.023	±0.021	±0.138	±0.099	±0.167	±0.022	
08-02/0.0-0.0	0.832	0.002	0.981	0.989	0.981	0.98	0.99	0.98	0.985	0.903	0.823	0.735	0.983	
	±0.352	±0.008	±0.018	±0.011	±0.018	±0.017	±0.012	±0.019	±0.017	±0.269	±0.362	±0.414	±0.016	
08-02/0.0-0.3	0.52	0	0.726	0.98	0.726	0.721	0.98	0.722	0.734	0.691	0.637	0.553	0.761	
	±0.368	±0.005	±0.183	±0.016	±0.183	±0.195	±0.016	±0.06	±0.06	±0.17	±0.242	±0.283	±0.054	
08-02/0.0-0.7	0.399	0	0.526	0.98	0.526	0.465	0.98	0.424	0.44	0.385	0.352	0.256	0.473	
	±0.437	±0.005	±0.407	±0.016	±0.407	±0.416	±0.016	±0.066	±0.065	±0.129	±0.153	±0.193	±0.063	
08-02/0.3-0.0	0.834	0.001	0.98	0.989	0.98	0.98	0.99	0.978	0.981	0.882	0.804	0.682	0.984	
	±0.353	±0.007	±0.019	±0.012	±0.019	±0.019	±0.011	±0.02	±0.017	±0.297	±0.381	±0.439	±0.016	
08-02/0.3-0.3	0.547	0.001	0.723	0.98	0.723	0.738	0.98	0.714	0.727	0.685	0.565	0.49	0.747	
	±0.369	±0.006	±0.179	±0.016	±0.179	±0.179	±0.016	±0.054	±0.053	±0.146	±0.287	±0.318	±0.047	
08-02/0.3-0.7	0.407	0.001	0.471	0.98	0.471	0.455	0.98	0.418	0.434	0.381	0.313	0.21	0.476	
	±0.437	±0.005	±0.407	±0.016	±0.407	±0.417	±0.016	±0.061	±0.062	±0.122	±0.174	±0.192	±0.058	
08-02/0.7-0.0	0.815	0.002	0.98	0.989	0.98	0.979	0.99	0.98	0.983	0.883	0.805	0.683	0.984	
	±0.371	±0.007	±0.02	±0.012	±0.02	±0.02	±0.011	±0.019	±0.017	±0.297	±0.38	±0.44	±0.016	
08-02/0.7-0.3	0.548	0.001	0.725	0.98	0.725	0.736	0.98	0.72	0.73	0.679	0.566	0.493	0.749	
	±0.37	±0.006	±0.177	±0.016	±0.177	±0.169	±0.016	±0.058	±0.056	±0.162	±0.288	±0.32	±0.047	
08-02/0.7-0.7	0.409	0	0.481	0.98	0.481	0.442	0.98	0.42	0.437	0.383	0.32	0.21	0.476	
	±0.442	±0.005	±0.408	±0.016	±0.408	±0.416	±0.016	±0.067	±0.064	±0.117	±0.173	±0.192	±0.059	
20-80/0.0-0.0	0.79	0.998	1	1	1	1	1	1	1	0.97	0.91	0.89	1	
	±0.409	±0.006	±0	±0	±0	±0	±0	±0	±0	±0.172	±0.289	±0.316	±0	
20-80/0.0-0.3	0.48	0.996	1	1	1	1	1	1	1	1	0.97	0.97	1	
	±0.502	±0.008	±0	±0	±0	±0	±0	±0	±0	±0	±0.172	±0.172	±0	
20-80/0.0-0.7	0.5	0.995	1	1	1	1	1	1	1	1	0.99	0.97	1	
	±0.502	±0.009	±0	±0	±0	±0	±0	±0	±0	±0	±0.1	±0.172	±0	
20-80/0.3-0.0	0.74	0.999	1	1	1	1	1	1	1	0.98	0.94	0.92	1	
	±0.441	±0.004	±0	±0	±0	±0	±0	±0	±0	±0.142	±0.239	±0.273	±0	
20-80/0.3-0.3	0.48	0.997	1	1	1	1	1	1	1	1	0.98	0.97	1	
	±0.502	±0.007	±0	±0	±0	±0	±0	±0	±0	±0	±0.141	±0.172	±0	
20-80/0.3-0.7	0.46	0.996	1	1	1	1	1	1	1	0.99	0.99	0.98	1	
	±0.501	±0.008	±0	±0	±0	±0	±0	±0	±0	±0.1	±0.1	±0.141	±0	
20-80/0.7-0.0	0.71	0.999	1	1	1	1	1	1	1	0.98	0.96	0.92	1	
	±0.456	±0.004	±0	±0	±0	±0	±0	±0	±0	±0.142	±0.198	±0.273	±0	
20-80/0.7-0.3	0.47	0.997	1	1	1	1	1	1	1	1	0.98	0.97	1	
	±0.502	±0.007	±0	±0	±0	±0	±0	±0	±0	±0	±0.141	±0.172	±0	
20-80/0.7-0.7	0.44	0.996	1	1	1	1	1	1	1	0.99	0.99	0.98	1	
	±0.499	±0.008	±0	±0	±0	±0	±0	±0	±0	±0.1	±0.1	±0.141	±0	
80-20/0.0-0.0	0.97	0.751	1	1	1	1	1	1	1	0.9	0.79	0.69	1	
	±0.171	±0.163	±0	±0	±0	±0	±0	±0	±0	±0.302	±0.41	±0.465	±0	
80-20/0.0-0.3	0.12	0.292	1	1	1	1	1	1	1	1	0.8	0.699	1	
	±0.327	±0.213	±0	±0	±0	±0	±0	±0	±0	±0	±0.402	±0.461	±0	
80-20/0.0-0.7	0.11	0.054	1	1	1	1	1	1	1	1	0.8	0.668	1	
	±0.314	±0.08	±0	±0	±0	±0	±0	±0	±0	±0	±0.402	±0.472	±0	
80-20/0.3-0.0	0.97	0.756	1	1	1	1	1	1	1	0.91	0.81	0.7	1	
	±0.171	±0.16	±0	±0	±0	±0	±0	±0	±0	±0.288	±0.395	±0.461	±0	
80-20/0.3-0.3	0.14	0.28	1	1	1	1	1	1	1	1	0.85	0.719	1	
	±0.349	±0.239	±0	±0	±0	±0	±0	±0	±0	±0	±0.359	±0.451	±0	
80-20/0.3-0.7	0.13	0.047	1	1	1	1	1	1	1	0.98	0.83	0.648	1	
	±0.338	±0.084	±0	±0	±0	±0	±0	±0	±0	±0.141	±0.377	±0.478	±0	
80-20/0.7-0.0	0.97	0.751	1	1	1	1	1	1	1	0.88	0.81	0.67	1	
	±0.171	±0.161	±0	±0	±0	±0	±0	±0	±0	±0.327	±0.395	±0.473	±0	
80-20/0.7-0.3	0.14	0.281	1	1	1	1	1	1	1	1	0.86	0.719	1	
	±0.349	±0.232	±0	±0	±0	±0	±0	±0	±0	±0	±0.349	±0.451	±0	
80-20/0.7-0.7	0.13	0.044	1	1	1	1	1	1	1	0.98	0.82	0.668	1	
	±0.338	±0.08	±0	±0	±0	±0	±0	±0	±0	±0.141	±0.386	±0.471	±0	

when the datasets have a higher number of attributes. In this case, we can affirm that most of the algorithms have the same performance, except the algorithms from group 4 (Y), which had the k-prototypes algorithm with UFBMD-BRKGA the best algorithm in four datasets (20-80/0.0-0.0, 80-20/0.0-0.0, 80-20/0.3-0.0, 80-20/0.7-0.0).

When the datasets have two numerical attributes and eight categorical, most of the results had not statistical difference, except in three datasets handled by the group

Table 29 – Results of experiment C1 - Normal distribution, 250 instances and ten levels in each categorical attribute

Dataset	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)			Group 4 (Y)		
	KAMLA	FMMMD	KMDBS	KMDBR	WKMSBS	WKMSMS	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
02-08/0.0-0.0	1	1	1	1	1	1	1	0.995	1	0.992	1	1	1
	± 0.002	± 0	± 0	± 0	± 0	± 0	± 0	± 0.009	± 0.002	± 0.011	± 0	± 0	± 0
02-08/0.0-0.3	1	1	1	1	1	1	1	0.912	1	0.903	0.999	1	1
	± 0.002	± 0	± 0.002	± 0	± 0.002	± 0	± 0	± 0.042	± 0	± 0.041	± 0.004	± 0.002	± 0.002
02-08/0.0-0.7	1	1	1	1	1	1	1	0.748	1	0.706	0.994	0.999	1
	± 0.002	± 0	± 0.002	± 0	± 0.002	± 0	± 0	± 0.07	± 0.002	± 0.104	± 0.009	± 0.004	± 0.002
02-08/0.3-0.0	0.999	1^{PY}	1	1	1	1	1	0.99	0.998	0.979	0.992	0.966	0.986
	± 0.004	± 0	± 0	± 0	± 0	± 0	± 0	± 0.013	± 0.006	± 0.02	± 0.012	± 0.025	± 0.016
02-08/0.3-0.3	0.999	1^{PY}	1	1	1	1	1	0.886	0.993	0.886	0.982	0.953	0.973
	± 0.004	± 0	± 0.002	± 0.002	± 0.002	± 0.002	± 0.002	± 0.045	± 0.01	± 0.044	± 0.016	± 0.029	± 0.021
02-08/0.3-0.7	0.999	1^{PY}	1	1	1	1	1	0.691	0.988	0.681	0.954	0.93	0.947
	± 0.005	± 0	± 0.002	± 0.002	± 0.002	± 0.002	± 0.002	± 0.065	± 0.014	± 0.091	± 0.026	± 0.038	± 0.029
02-08/0.7-0.0	0.997	1^{PY}	1	1	1	1	1	0.987	0.998	0.877	0.987	0.399	0.906
	± 0.008	± 0	± 0	± 0	± 0	± 0	± 0	± 0.015	± 0.006	± 0.045	± 0.016	± 0.099	± 0.036
02-08/0.7-0.3	0.997	1^{PY}	1	1	1	1	1	0.88	0.992	0.733	0.964	0.353	0.816
	± 0.008	± 0	± 0.002	± 0.002	± 0.002	± 0.002	± 0.002	± 0.043	± 0.01	± 0.116	± 0.022	± 0.083	± 0.045
02-08/0.7-0.7	0.997	1^{PY}	1	1	1	1	1	0.671	0.984	0.457	0.904	0.271	0.628
	± 0.008	± 0	± 0.002	± 0.002	± 0.002	± 0.002	± 0.002	± 0.063	± 0.015	± 0.142	± 0.038	± 0.077	± 0.047
08-02/0.0-0.0	1	1	1	1	1	1	1	1	1	0.994	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0	± 0	± 0	± 0	± 0.064	± 0	± 0	± 0
08-02/0.0-0.3	1	1	0.97	1	0.97	1	1	1	1	0.994	1	1	1
	± 0	± 0	± 0.17	± 0	± 0.17	± 0	± 0	± 0	± 0	± 0.059	± 0	± 0	± 0
08-02/0.0-0.7	1	1	0.98	1	0.98	1	1	1	1	0.984	1	1	1
	± 0	± 0	± 0.139	± 0	± 0.139	± 0	± 0	± 0	± 0	± 0.104	± 0	± 0	± 0
08-02/0.3-0.0	1	1	0.99	1	0.99	1	1	1	1	0.983	1	1	1
	± 0	± 0	± 0.096	± 0	± 0.096	± 0	± 0	± 0	± 0	± 0.081	± 0	± 0.002	± 0.002
08-02/0.3-0.3	0.999	1	0.97	1	0.97	1	1	1	1	0.862	1	1	1
	± 0.003	± 0	± 0.162	± 0	± 0.162	± 0	± 0	± 0.002	± 0	± 0.119	± 0	± 0.002	± 0.002
08-02/0.3-0.7	1	1	0.895	1	0.895	1	1	1	1	0.696	1	1	1
	± 0.003	± 0	± 0.295	± 0	± 0.295	± 0	± 0	± 0.002	± 0	± 0.126	± 0	± 0.002	± 0.002
08-02/0.7-0.0	0.921	1^{KW}	0.915	0.998	0.915	0.997	0.996	1	1	0.771	1	0.651	0.816
	± 0.145	± 0	± 0.249	± 0.005	± 0.249	± 0.007	± 0.009	± 0	± 0	± 0.096	± 0.002	± 0.095	± 0.057
08-02/0.7-0.3	0.636	1^{KW}	0.686	0.98	0.686	0.961	0.963	0.998	1	0.626	1	0.639	0.774
	± 0.205	± 0	± 0.364	± 0.018	± 0.364	± 0.103	± 0.036	± 0.005	± 0	± 0.168	± 0.002	± 0.09	± 0.056
08-02/0.7-0.7	0.285	1^{KW}	0.491	0.986	0.491	0.913	0.976	0.997	1	0.463	1	0.625	0.715
	± 0.185	± 0	± 0.416	± 0.016	± 0.416	± 0.218	± 0.032	± 0.008	± 0	± 0.154	± 0.002	± 0.09	± 0.063
02-02/0.0-0.0	1	1	0.97	1	0.97	1	1	0.993	1	0.984	0.999	1	1
	± 0.002	± 0	± 0.165	± 0	± 0.165	± 0	± 0	± 0.009	± 0.002	± 0.056	± 0.004	± 0.002	± 0.002
02-02/0.0-0.3	1	1	0.969	1	0.969	0.996	1	0.922	1	0.85	0.999	1	1
	± 0.002	± 0.002	± 0.168	± 0	± 0.168	± 0.036	± 0	± 0.038	± 0	± 0.138	± 0.003	± 0.002	± 0.002
02-02/0.0-0.7	1	1	0.933	1	0.933	1	1	0.801	1	0.704	0.998	0.999	1
	± 0.002	± 0.002	± 0.222	± 0	± 0.222	± 0	± 0	± 0.066	± 0	± 0.131	± 0.006	± 0.003	± 0.002
02-02/0.3-0.0	0.994	1^{WPY}	0.948	0.999	0.948	0.998	0.998	0.971	0.984	0.906	0.967	0.906	0.934
	± 0.011	± 0.002	± 0.206	± 0.005	± 0.206	± 0.006	± 0.006	± 0.023	± 0.017	± 0.079	± 0.024	± 0.045	± 0.033
02-02/0.3-0.3	0.984	0.999	0.885	0.993	0.885	0.979	0.987	0.885	0.982	0.786	0.961	0.898	0.919
	± 0.018	± 0.004	± 0.248	± 0.01	± 0.248	± 0.091	± 0.016	± 0.049	± 0.019	± 0.142	± 0.025	± 0.046	± 0.032
02-02/0.3-0.7	0.983	0.999	0.79	0.993	0.79	0.976	0.987	0.745	0.979	0.622	0.946	0.887	0.907
	± 0.018	± 0.004	± 0.324	± 0.011	± 0.324	± 0.098	± 0.016	± 0.061	± 0.021	± 0.15	± 0.029	± 0.05	± 0.04
02-02/0.7-0.0	0.979	1	0.841	0.99	0.841	0.976	0.987	0.965	0.982	0.286	0.936	0.174	0.463
	± 0.019	± 0.003	± 0.318	± 0.011	± 0.318	± 0.089	± 0.014	± 0.028	± 0.02	± 0.163	± 0.039	± 0.06	± 0.069
02-02/0.7-0.3	0.852	0.998	0.567	0.921	0.567	0.831	0.895	0.868	0.976	0.206	0.92	0.167	0.404
	± 0.098	± 0.005	± 0.34	± 0.047	± 0.34	± 0.174	± 0.063	± 0.047	± 0.022	± 0.145	± 0.039	± 0.059	± 0.062
02-02/0.7-0.7	0.682	0.997	0.368	0.962	0.368	0.709	0.935	0.723	0.973	0.135	0.873	0.149	0.313
	± 0.235	± 0.006	± 0.321	± 0.045	± 0.321	± 0.287	± 0.092	± 0.066	± 0.023	± 0.105	± 0.05	± 0.054	± 0.062

attributes than numerical, most of the data clustering algorithms considered in this paper do not present an expressive sensibility concerning the overlapping of the attributes, i.e. the datasets might even have irrelevant information in the categorical features that it will not disturb the main objective of these types of algorithms, which is finding groups.

Here we can conclude a very important point of this study in practical terms, i.e. if the analyzers question themselves whether they can add more categorical features into the dataset because they might not know which categorical attribute could be more relevant to find good clusters, then they can add whatever they want respecting the unbalanced characteristic of the datasets in terms of the numerical features, and the results tend to be insensible to irrelevant categorical features. This conclusion is strongly valid to groups 1 and 2, and less strong to group 3, i.e. for weighted k-means, k-means, and PAM algorithm. It is not valid to the k-prototypes algorithm, which is more sensitive to the choice of lambda value by a good feature balancing strategy.

On the other hand, when the datasets have a few attributes and more numerical

Table 32 – Results of experiment D2 - LogNormal distribution, two levels in each categorical attribute

Dataset	All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTBL	KPTBR
02-02/0.0-0.0	1 ±0	1 ±0	1 ±0.001	1 ±0	1 ±0.001	1 ±0.001	1 ±0	1 ±0.001	1 ±0	1 ±0.001	1 ±0.001	1 ±0	1 ±0
20-20/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
200-200/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0

Figure 46 – Execution time in different numbers of instances

algorithms had not significantly different performance.

Figure 47 – Execution time in different numbers of attributes

5.8.2 Experiment B

With respect to experiment B1, when the datasets are balanced concerning the number of numerical and categorical attributes, one can note that the algorithms had different performances when the datasets were small, with two features for each type. In this scenario, the FMMD algorithm achieved the best ARI values in four datasets considering all the groups, and in two datasets such a method achieved the best performance in group K, and in the groups K, W, and Y. When datasets were increased ten times with respect to the number of attributes, we note that all the algorithms performed equally, i.e. there were not any statistical differences concerning the ARI value. Comparing the experiment B1

with experiment A1, in which the datasets were also generated by the normal distribution, we can claim that the fact the dataset is unbalanced or balanced with respect to the number of numerical and categorical attributes does not influence the performance of the algorithms if the datasets are not too small. However, if the dataset has a few attributes, then the performance of the algorithms is influenced by the overlapping, and the choice of which algorithm to use also makes a difference in the clustering performance.

Regarding the experiment B2, we can note that when the datasets have a few attributes, our algorithms achieved the best performance in all the datasets in the groups K and W. In the groups P and Y, our algorithms just not achieved the best ARI values in the datasets which had 0% for overlapping in the categorical attributes. In the remaining datasets, our methods obtained the best performance, statistically tested. However, when the datasets are larger with respect to the number of attributes, we note that, generally, the algorithms achieved the same performance, except in the first dataset, in which the k-prototypes algorithm with our optimization proposal to choose the lambda obtained the best performance.

5.8.3 Experiment C

In relation to experiment C1, where we maintained all the configurations chosen in the other experiments except the number of levels in the categorical attributes, we note that the FMMMD algorithm obtained the best performance in many datasets. When the datasets have two numerical attributes and eight categorical, the mentioned method achieved the best ARI values in the groups P and Y. However, when the datasets have eight numerical attributes and two categorical, the FMMMD algorithm obtained the best performance in the groups K and W. When the datasets were balanced concerning the number of attributes of each type, one observes that the FMMMD algorithm in the best method of all groups in five of nine datasets, and it is also the best algorithm in one dataset for the groups W, P, and Y.

In experiment C2, we note a very interesting result. This experiment is equal to the experiment A2 for datasets with two and eight numerical and categorical attributes, respectively, and equal to the experiment B2, for datasets with twenty attributes of each type, except to the number of levels in the categorical features, which in this experiment the datasets have ten levels, and in the other experiments, the datasets have two, in both cases.

When we analyze the datasets with two numerical and eight categorical attributes, we note that when the number of levels of categorical attributes was two, the algorithms that achieved the best ARI were the k-prototypes with our proposed feature balancing strategy (in three of nine datasets), and the PAM also with the proposed BRKGA optimizer (in one of nine datasets). However, when the number of levels increased from two to ten,

the FMMD algorithm was the only method that achieves the best ARI values, in three of nine datasets, the same set of data the k-prototypes obtained the best performance in the experiment A2.

When we analyze the datasets with eight numerical and two categorical attributes, we note that when the number of levels of categorical attributes increases from two to ten, the Kamila algorithm took the place of the best performance from PAM and k-prototypes, both with our proposed strategy, in six of nine datasets, though the groups 1 and 2 maintain intact with respect to the best algorithms, which are the KMDBR, and WKMBR.

When the datasets are balanced regarding the number of attributes of each type of feature, i.e. when both types have two attributes, we can compare the respective datasets with the datasets of experiment B2. With respect to groups 1 and 2, nothing changed, i.e. the best algorithms continued to be our proposed algorithms. Concerning groups 3 and 4, in experiment C2, our proposals achieved the best ARI values in fewer datasets than in experiment B2, for datasets with two numerical and categorical attributes.

5.8.4 Experiment D

Experiment D1 shows the behavior of the execution time from all the methods when the datasets increase in a number of instances. It is important to highlight that the direct comparison is not possible because some algorithms have part of their code in C language, which is faster than the R language. Although the direct comparison is not possible, we can analyze how the execution time changes as the number of instances increases. In groups 1 (K) and 2 (W), we observe that the FMMD and KAMLA algorithms present a higher slope than the other methods, followed by the k-means with our proposed approach and the k-means algorithm with binary dummy coding. With respect to group 3 (P), the highest slope is from PAM with the Gower distance, the method we selected as a baseline, followed by the Kamila, FMMD algorithms, and our proposal. In this group, it is interesting to note that datasets with more than 1,000 instances make our proposal decreasing the execution time. Concerning group 4 (Y), the following algorithms have similar behavior, the KPTBS, KPTNM, KTPEL. The highest slope is the Kamila algorithm, followed by the FMMD and, our proposed algorithm.

From experiment D1, we can observe that, although our proposed optimizer required more time to execute the algorithms, in a relative analysis concerning the increase of the number of attributes, our proposal maintains an attractive time slope as the number of instances is increasing.

Regarding experiment D2, in all the groups, our proposed approaches achieved an attractive time slope in comparison to the other methods, except in group 2, in which the Modha-Spangler method obtained a decreasing time slope.

5.9 Final remarks

As highlighted in the Subsection 5.2.2, determining an optimum and unique weight to the entire set of categorical attributes is the main challenge of feature balancing approaches, especially because the choice of this weight relies on the characteristics of each dataset. As could be seen in this chapter, many scientific works have already addressed this problem that comes out in distance-based data clustering algorithms, pointing out that defining such weight is a hard (FOSS *et al.*, 2016). The proposed methods aimed to assign a quasi-optimum weight to four distance-based mixed data clustering algorithms, which are an excellent strategy since doing it manually or using a brute-force search is not effective.

Hence, we proposed an optimizer to address the feature balancing problem in four mixed data clustering algorithms. We selected nine methods to compare to our proposals. The methodology was composed of four experiments, which simulated 423 datasets with many characteristics in order to analyze the behavior of the algorithms in different scenarios. We defined all the main parameters of the algorithms using the irace package. The results showed that as the datasets are increasing in the number of attributes, the performance of the algorithms tends to not differ. However, when the datasets have a small dimensionality, the feature balancing strategy chosen to handle this problem affects significantly the performance of the algorithms. Another interesting observation came from the fact that when there are categorical attributes overlapped, the algorithms are not so sensible as when the overlapping is in the numerical features. Regarding the execution time, we analyzed the time curve over different attributes and instances. We found that the curve slope of our proposed methods is attractive compared to other approaches.

6 FEATURE WEIGHTING

“Shall I refuse my dinner because I do not fully understand the process of digestion?”

Oliver Heaviside

6.1 Motivation

In Chapter 5, one presented optimization methods to balance the contribution of numerical and categorical attributes in distance-based mixed data clustering algorithms under the assumption that all features had the same importance in the clustering process. However, there are many situations that weighting a set of attributes, or all of them is desired in order to emphasize the contribution of the most important ones while decreasing the impact of irrelevant features. Therefore, feature weighting is a preprocessing task highly addressed in the pattern recognition field, especially in data clustering algorithms. Nonetheless, too few studies have addressed such an issue in the mixed data clustering problem. In this select set of works, only Martarelli and Nagano (2019) proposed to handle feature weighting problems in a model-based mixed data clustering algorithm using optimization methods. The authors modeled the BKRGGA algorithm for this problem, using the KAMILA algorithm as the clustering method and the KAMILA objective function as the fitness function of the meta-heuristics.

Hence, in this chapter, one presents an extension of the Martarelli and Nagano (2019) work, proposing a set of optimization algorithms to handle the feature weighting problem in KAMILA algorithm. This set of optimization methods comprehends modeling three heuristics and meta-heuristics (Biased Random-Key Genetic Algorithm, Evolutionary Clustering Search, Iterated Local Search) and hybridizing them.

The outline of this chapter is as follows. In Section 2 one presents basic concepts of the main heuristics used in this chapter. Section 3 describes the proposed algorithms. Section 4 presents the methodology of this proposal. Section 4 and 5 present the results and discuss them, respectively. Finally, Section 6 concludes this chapter.

6.2 Related works

In the following paragraphs, one brings a brief remind-description of the studies that addressed the feature weighting problem from all data clustering branches, and that have already been presented in Chapter 4.

From the partitional branch, Amorim and Makarenkov (2016) presented a modification of the k-means for mixed data, assigning different weight values for the attributes in a subclustering scenario. The authors proposed a method called Minkowski weighted

k-means with distributed centroids (MWK-DC). Liang, Wang and Lai (2017) proposed a new algorithm, called WKStream, to cope with mixed data stream with different shapes. They utilized the concepts of micro- and macro-clusters and also developed another algorithm, called weighted k-means+ (Wk-means+). In addition, the authors weighted numerical and categorical attributes based on *non-uniformity* concept. Özbakır and Turna (2017) modeled two meta-heuristics, called ions motion optimization (IMO) and weighted superposition attraction (WSA), to define the number of clusters. Jia and Cheung (2018) extended the OCIL algorithm, proposing weights for the attributes in different groups (subspace). In addition, the authors presented a method, called WOCIL, to determine the centroids initialization and the number of groups.

Concerning the hierarchical branch, Philip and Ottaway (1983) explored the usage of the data clustering algorithm with the Gower distance and the idea of feature weighting. The authors also applied the clustering technique to analyze the age of old weapons. Friedman and Meulman (2004) proposed a new mixed data clustering algorithm, called clustering of objects on subsets of attributes (COSA), which addresses the subspace clustering problem with different attribute weights for each sub clustering, the missing data issue, and the optimization of the weights of the attributes. The work of Kampert, Meulman and Friedman (2017) brought an up-to-date software package that extended the COSA software proposed by Friedman and Meulman (2004), adding functions for clustering methods, which feature weighting problem was addressed by a proposed weighted dissimilarity function. To classify the soil, Bragato, Ganis and Feoli (2019) applied a complete linkage hierarchical clustering algorithm and compared two dissimilarities (Burnaby and Goodall) in the same clustering method to analyze the influence of each one in the clustering results. They found that the Burnaby dissimilarity considered association among features, assigning higher weights to independent attributes, and Goodall dissimilarity considered more similar the observations that share infrequent values than those that share frequent values.

With respect to the model-based branch, Ingrassia *et al.* (2015) introduced a generalized linear mixed cluster-weighted model (GLMCWM), based on cluster-weighted models (CWM). Regarding others branch, Yang and Wu (2011) proposed a mixed data clustering algorithm, based on k-modes, for large datasets. The authors based on graph theory to assign the observations for each center of the clusters. They addressed the feature weighting and number of clusters problems and used the nearest-neighbor clustering and genetic algorithm. Finally, from IAPRA branch, Cox and Cox (2000) proposed an extension of the Gower distance, which yielded, simultaneously, a set of weighted dissimilarities between pairs of objects and variables. The weights of the objects and variables were automatically defined by the proposed method.

6.3 Proposed algorithms

In the following sections, one describes the proposed optimization algorithms, providing the theoretical explanation for the methods that were not explained in the previous chapters, which are Evolutionary Clustering Search, Iterated Local Search, and the hybrid method that was designed to handle the feature weighting problem.

6.3.1 Evolutionary Clustering Search

The Evolutionary Clustering Search (ECS) is a specification of the Clustering Search meta-heuristic, which were proposed by Oliveira and Lorena (2004), Oliveira and Lorena (2007). While both approaches are designed to explore promise solutions belonging to systematically defined regions from the solutions space, the former considers the usage of an evolutionary method in one of its steps and the last not. The main idea is to have a method to generate the solutions, cluster them, analyze the clusters to identify the promise ones, and then investigate those prominent groups in the previous step throughout a local search method. Thus, four steps compose both methods, which are: i) Meta-heuristic (ME); ii) Iterative Clustering Process (IC); iii) Cluster Analyzer (CA), and iv) Local Optimization Algorithm (LO). In the following paragraphs, one gives a brief description of each step and how they were related, as well as how ECS was modeled to handle the feature weighting problem in mixed data clustering technique.

First, the solutions were generated in the ME step by a meta-heuristic. For the feature weighting problem handled by this chapter, we selected the Genetic Algorithm (GA) for this step, in which the solutions were coded as being a vector with values from 0 to 1, randomly generated, and with the number of positions equal to the number of attributes of the dataset. Such a values represent the weights of the attributes. The crossover procedure considered the two-cut method and the crossover rate p_c , and the mutation step checked for each individual the mutation rate p_m , and for those the mutation occurred, it was replaced p_{mm} percent of the alleles by random values from 0 to 1. Then, these solutions (initial population from the GA) were clustered into $n_{Clusters}$ clusters (predefined parameter) by a clustering approach. In this work, the clusters were created choosing the $n_{Clusters}$ centers randomly, and assigning the solutions to closest center throughout the Euclidean distance calculation.

Next, the solutions were evaluated and the crossover and mutation procedures from GA were executed, as follows. The crossover was done by the two-cut method and the mutation was carried out by executing two local search approaches, as follows. The first local search method checked if a random value from 0 to 1 is less than the mutation ratio (p_m) for each offspring resulted from the crossover procedure. In an affirmative case, it replaced a predefined percentage (p_{MILS1}) of alleles in the individuals' string by values generated randomly, following the same procedure adopted to generate the initial

population. The second local search procedure firstly checked if a random value from 0 to 1 is less than the mutation ratio (p_m). In an affirmative case, this method replaced a predefined percentage (p_{MILS2}) of alleles in the individuals' string by values generated randomly, following the same procedure adopted to generate the initial population. These individuals were selected randomly from the offspring resulted from the crossover procedure and also from the the first local search method, and the number of selected individuals was the rounded result of the multiplication between the number of offspring so far and the mutation rate. Next, one evaluated all offspring resulted from the crossover and mutation approaches, adding into the clusters only those that had a fitness value greater than the minimum value of the population. When a new solution was added into a cluster, its centroid was updated so that each one of its allele representing the average value of that allele from all solutions belonging into that cluster. Thus, the IC step was carried out.

The CA step consisted of analyzing each cluster to identify promise ones. In this work, this analysis was done as follows. First, one calculated the average fitness value of each cluster and the entire population, dividing the first by the second, which resulted in a ratio. If this ratio was greater than 0.7, one indicated the respective cluster was promissory. For those clusters that already achieved an average fitness value that represent the best one, the analysis was not done because it was not necessary.

The last step was the LO, which established the local search algorithm to be executed in the promissory clusters. For this problem, we used the Iterated Local Search (ILS) method, which consisted of generating initial solutions, carrying out a local search algorithm on them, applying a perturbation on the initial solutions, carrying out a local search algorithm on them, and finally verifying whether the perturbation resulted in better values or not. As initial solutions, one considered p_{MILS31} percent of the total of individual belonging of the cluster. As those solutions had already been resulted of local search procedures carried out in previous steps of the ECS, then the first local search algorithm did not be executed. For each selected individual, the perturbation was done by replacing p_{MILS32} percent of the alleles by randomly generated values. Then, if the resulted fitness value was greater than that of the selected individual (see the fitness function one used in this chapter in section 6.4.4), then the selected individual was replaced by the one generated from the permutation procedure.

Therefore, Algorithm 2 presents the pseudocode.

6.3.2 Iterated Local Search

The ILS algorithm modeled in this chapter to handle the feature weighting problem was the same one that was modeled in the IC step of ECS algorithm. Therefore, Algorithm

Algorithm 2: Evolutionary Clustering Search

Define the parameters $n_{pop}, p_c, p_m, n_{Clusters}, PMILS1, PMILS2, PMILS31, PMILS32$;
 Build the initial population by the GA algorithm;
 Define the centers of the clusters;
 Evaluate each individual;
while *the stopping criteria are not reach* **do**
 Build the next population carrying out the crossover and mutation procedures by the
 GA algorithm;
 Add the best new individuals into the clusters;
 if *promissory cluster is identified* **then**
 Carry out the local search algorithm (ILS);

3 presents the pseudocode.

Algorithm 3: Iterated Local Search

Define the parameters $n_{pop}, p_{IndILS}, p_{AttrILS}$;
 Build the initial population;
 Evaluate each individual;
while *the stopping criteria are not reach* **do**
 for p_{IndILS} *percent of the individuals* **do**
 Select an individual randomly from the entire population;
 Replace random values for $p_{AttrILS}$ percent of alleles from the selected
 individuals;
 Evaluate the modified individual;
 Insert the individual into the population if it is better than the selected one;

6.3.3 Hybridized algorithm

The hybridized method proposed in this chapter was the same of the ECS algorithm, however, instead of applying the GA method, it was applied the BRKGA proposed by Martarelli and Nagano (2019). Therefore, Algorithm 4 presents the pseudocode.

The predefined parameters that were mentioned in the description of the methods are shown in section 6.4.3, as well as the explanation of how they were defined. In the next section, one describes the methodology of this work.

6.4 Methodology

In this section, one describes the experiments that were designed to understand the behavior of the proposed methods. It is also presented the methods that were used to

Algorithm 4: Unsupervised Feature Weighting for Mixed Datasets through Hybrid Algorithm

```

Define the parameters  $n_{pop}, p_c, p_m, n_{Clusters}, PMILS1, PMILS2, PMILS31, PMILS32$ ;
Build the initial population by the BRKGA algorithm;
Define the centers of the clusters;
Evaluate each individual;
while the stopping criteria are not reach do
    Build the next population carrying out the crossover and mutation procedures by the
    BRKGA algorithm;
    Add the best new individuals into the clusters;
    if promissory cluster is identified then
        Carry out the local search algorithm (ILS);

```

generate the datasets, the methods that were selected to compare the proposed algorithms, the manner one defined the parameters of the methods for each set of experiments, the performance measure that was chosen to evaluate the algorithms, and the computational setting used.

6.4.1 Experiments

We designed three experiments, in which the first and the second (A and B, respectively) differ by the fact the data from the former were generated by the normal distribution and from the last by non-normal distribution (logNormal). The first two experiments generated 24 datasets each, which present a balanced and unbalanced number of numerical and categorical attributes. All categorical features had two levels of categories. In addition, in these experiments when the overlap of the attributes occurred, this was done overlapping them in 50%. There were three values of overlapping percentages for numerical and categorical features, which were 20%, 40%, and 80%. The last experiment (C) generated 5 datasets, all from the normal distribution, 50% of overlap in those attributes, and 40% of overlapped attributes of each type.

We elected the lognormal with skew equal to 9 as the non-normal distribution for all experiments, which is correspondent to a log standard deviation of 1.131. Concerning the specific characteristics of each experiment, Table 33 brings all the information.

Table 33 – Details of the experiments

Experiment	Number of instances	Distribution		Attributes combination (numerical/categorical)					
		Normal	LogNormal	Balanced			Unbalanced		
				5/5	50/50	5/10	10/5	50/100	100/50
A	250	x		x	x	x	x	x	x
B	250		x	x	x	x	x	x	x
C	500, 1000, 2000, 3000, 5000	x		x					

6.4.2 Methods

In the following topics, one describes each baseline methods that were selected, as well as the proposed algorithms. Note that for all method, the Kamila algorithm (FOSS *et al.*, 2016; FOSS; MARKATOU, 2018) was carried ou as the clustering approach.

- KAMBSLN - Baseline algorithm 1: All weights equal to 1
- KAMGAFW - Baseline algorithm 2: Genetic Algorithm, which was modeled as follows. The solutions were coded as being a vector with values from 0 to 1, randomly generated, and with the number of positions equal to the number of attributes of the dataset. The crossover procedure considered the two-cut method and the crossover rate p_c , and the mutation step checked for each individual the mutation rate p_m , and for those the mutation occurred, it was replaced p_{mm} percent of the alleles by random values from 0 to 1. It is important to highlight that this GA algorithm is the same of those used in the modeling of the ECS algorithm;
- KAMBRKG - Baseline algorithm 3: BRKGA proposed by Martarelli and Nagano (2019), which was modeled in the same way the authors did;
- KAMECSF - Proposed method 1: ECS, which was modeled as presented in section 6.3;
- KAMILSF - Proposed method 2: ILS, which was modeled as presented in section 6.3;
- KAMECBR - Proposed method 3: the hybridized method, which was modeled as presented in section 6.3.

The stopping criteria for all the algorithms was defined as follows. The algorithm stops when it achieves *maxIter* iterations or when its improvement of quality is less than *minInc* after *minIt* successive generations.

6.4.3 Parameters definition

The parameters were defined using the *irace* package (LÓPEZ-IBÁÑEZ *et al.*, 2016), and their values are presented in Appendix E (Tables from 107 to 109). We considered the following range of values for each parameter: (i) *npop*: from 9 to 15; (ii) *nClusters*: from 2 to 5; (iii) p_c , p_m , p_{mm} , p_e , ρ , p_{MILS1} , p_{MILS2} , p_{MILS31} , p_{MILS32} , p_{IndILS} , $p_{AttrILS}$: from 0.1 to 1, except p_{mm} to GA, which is from 0.1 to 0.5.

6.4.4 Performance measure and statistical test

The objective function of the Kamila algorithm was considered to be the performance measure in this chapter. As the statistical test, we elected the Wilcoxon rank-sum test to evaluate whether there was a relevant difference between the best result of an algorithm in a specific dataset and the other algorithms. For that, we used the *wilcox.test* function from *stats* package considering one-side (greater option), with a confident level of 95%.

6.4.5 Computational setting

The computational experiments were carried out in R language, on a personal desktop computer with Intel Core i7-4790K CPU @ 4 GHz along with 16 GB of RAM under Windows 10.

6.5 Results and discussion

In this section, we present the results we obtained in each experiment. Tables from 34 to 36 present the average of ARI values and the standard deviation of each measure of the experiments A to C, respectively. Note that bold values are statistically better than the other ones.

One notes that the first column provides the information about some characteristics of the dataset, which the first two-element that compose the string means the number of numerical and categorical attributes, respectively, and after the backslash, the two elements represent the number of numerical and categorical attributes overlapped, respectively.

In Appendix E, we provide the average and standard deviation of the execution time of each algorithm (Tables from 110 to 112).

As one can be observed, the hybridized method (KAMECBR) performed better than the other algorithms in almost all the datasets of all the experiments, showing that the union of the BRKGA, ECS, and ILS provided a significant improvement in the solution quality, which did not occur when the meta-heuristics were carried out individually. It is also important to highlight that the BRKGA (KAMBRKG) achieved a better performance than the GA (KAMGAFW), which was already expected since it was reported by to the work of Martarelli and Nagano (2019), and the performance of ILS algorithm was worst than the BRKGA, and even GA, which one attributes by the fact the ILS considered only random generated values to replace the alleles in the individuals. It could be better if the replacement considered some information of the problem to create the new values for the selected alleles. The ECS method achieved excellent results, which were as good as the hybridized meta-heuristic in datasets number 16 and 24 from experiment B.

Regarding the execution time, the hybridized method assumed similar values to

Table 34 – Results of experiment A - Average and standard deviation of the value of the objective function of the Kamila algorithm (the greater the better). Bold values indicate the best method with respect to performance (solution quality).

Dataset scenario	Baseline			Proposed algorithms		
	KAMBSLN	KAMGAFW	KAMBRKG	KAMILSF	KAMECSF	KAMECBR
[1] 05-10/00-00	-92.709	-24.903	-20.946	-26.728	-16.219	-13.329
	± 6.922	± 5.032	± 4.169	± 4.398	± 2.882	± 3.033
[2] 05-10/04-02	-1585.817	-242.862	-181.56	-275.01	-110.085	-82.368
	± 125.616	± 60.105	± 57.446	± 72.32	± 38.887	± 31.117
[3] 05-10/02-04	-1330.461	-160.051	-119.112	-167.419	-79.522	-63.748
	± 89.009	± 55.667	± 36.99	± 50.474	± 20.437	± 17.912
[4] 05-10/01-08	-1423.616	-221.714	-175.797	-225.161	-137.211	-106.662
	± 79.559	± 38.794	± 37.44	± 41.981	± 25.752	± 26.871
[5] 10-05/00-00	-47.9	-9.319	-7.527	-10.441	-5.388	-4.233
	± 4.922	± 2.571	± 2.472	± 2.765	± 1.658	± 1.51
[6] 10-05/08-01	-808.151	-101.419	-74.2	-106.33	-52.057	-38.751
	± 56.71	± 33.755	± 22.165	± 30.888	± 17.746	± 14.864
[7] 10-05/04-02	-711.553	-79.734	-54.819	-83.022	-33.369	-26.178
	± 40.021	± 27.736	± 20.793	± 33.129	± 12.91	± 12.518
[8] 10-05/02-04	-762.964	-99.248	-77.24	-115.076	-48.694	-36.104
	± 41.284	± 31.841	± 30.183	± 35.078	± 15.309	± 13.4
[9] 05-05/00-00	-46.605	-8.897	-7.596	-9.999	-4.493	-3.905
	± 5.32	± 2.481	± 2.49	± 2.548	± 1.419	± 1.375
[10] 05-05/04-01	-788.189	-86.279	-61.917	-101.028	-37.594	-27.59
	± 70.08	± 29.93	± 22.359	± 34.58	± 16.114	± 13.071
[11] 05-05/02-02	-666.817	-58.407	-40.213	-66.743	-23.055	-19.028
	± 44.865	± 20.738	± 18.527	± 22.404	± 8.65	± 8.906
[12] 05-05/01-04	-709.643	-83.696	-62.698	-89.375	-37.395	-32.005
	± 43.351	± 26.063	± 20.97	± 24.793	± 12.875	± 12.701
[13] 50-100/00-00	-964.472	-409.135	-392.858	-416.173	-376.607	-348.054
	± 23.428	± 18.741	± 17.083	± 18.778	± 14.211	± 16.118
[14] 50-100/40-20	-16556.382	-6190.533	-5717.258	-6413.789	-5256.561	-4676.343
	± 456.712	± 475.713	± 436.027	± 474.25	± 403.486	± 372.17
[15] 50-100/20-40	-15268.126	-5253.522	-4603.061	-5323.489	-4228.631	-3452.247
	± 363.142	± 511.39	± 477.679	± 494.946	± 385.031	± 364.958
[16] 50-100/10-80	-16772.269	-5350.315	-4594.477	-5531.45	-4328.207	-3721.72
	± 381.317	± 523.759	± 494.45	± 555.158	± 419.017	± 336.593
[17] 100-50/00-00	-484.322	-193.48	-179.698	-197.41	-169.489	-150.633
	± 15.027	± 11.099	± 12.346	± 12.542	± 12.158	± 9.758
[18] 100-50/80-10	-8313.078	-2935.555	-2561.218	-2985.063	-2387.373	-1979.451
	± 209.22	± 240.608	± 273.197	± 287.177	± 225.58	± 222.514
[19] 100-50/40-20	-7686.645	-2707.263	-2365.559	-2769.724	-2136.108	-1814.414
	± 168.03	± 193.726	± 236.952	± 226.658	± 215.252	± 196.124
[20] 100-50/20-40	-8489.462	-2924.446	-2617.435	-3033.248	-2373.434	-2007.819
	± 154.39	± 246.651	± 232.283	± 278.757	± 221.848	± 182.329
[21] 50-50/00-00	-480.587	-189.262	-178.48	-195.074	-166.541	-149.285
	± 16.795	± 11.627	± 11.5	± 10.721	± 10.23	± 10.705
[22] 50-50/40-10	-8259.275	-2795.035	-2550.277	-2874.603	-2272.279	-1942.682
	± 261.97	± 277.273	± 302.686	± 302.541	± 227.109	± 234.778
[23] 50-50/20-20	-7653.261	-2470.082	-2205.013	-2602.122	-1897.438	-1534.605
	± 189.835	± 266.267	± 260.806	± 224.36	± 185.399	± 170.922
[24] 50-50/10-40	-8382.175	-2568.216	-2244.638	-2695.778	-2033.416	-1675.105
	± 188.574	± 285.537	± 253.49	± 252.431	± 196.029	± 193.883

the ECS method, however, both took twice as long to execute the procedures as the GA and BRKGA, and almost five times more if they are compared to the Kamila algorithm with all weights equal to one (KAMBSLN) and ILS (KAMILSF).

Table 35 – Results of experiment B - Average and standard deviation of the value of the objective function of the Kamila algorithm (the greater the better). Bold values indicate the best method with respect to performance (solution quality).

Dataset scenario	Baseline				Proposed algorithms	
	KAMBSLN	KAMGAFW	KAMBRKG	KAMILSF	KAMECSF	KAMECBR
[1] 05-10/00-00	-61040.613 ±82998.562	-4741.414 ±2387.55	-4262.231 ±1906.309	-5741.092 ±5312.833	-4023.747 ±3051.467	-2487.058 ±1470.577
[2] 05-10/04-02	-159694.377 ±357163.994	-7305.54 ±4904.866	-7357.5 ±5792.137	-7529.141 ±5362.489	-4896.436 ±3110.42	-3316.978 ±2656.234
[3] 05-10/02-04	-253954.745 ±1006828.736	-9780.672 ±8604.421	-8950.324 ±8734.769	-9985.305 ±7915.387	-6143.406 ±5257.348	-4875.592 ±3875.572
[4] 05-10/01-08	-100305.977 ±391307.233	-10549.268 ±10913.518	-10763.767 ±11055.04	-11345.241 ±11181.487	-6987.403 ±8357.809	-5303.858 ±5777.591
[5] 10-05/00-00	-29124.73 ±53827.737	-2823.717 ±2218.855	-2222.268 ±1552.986	-3020.291 ±2379.346	-1812.215 ±915.894	-1130.143 ±684.992
[6] 10-05/08-01	-28297.127 ±33360.203	-4063.215 ±2215.089	-3955.494 ±2768.999	-4666.85 ±3798.69	-2794.447 ±1514.688	-1965.99 ±2049.916
[7] 10-05/04-02	-32339.916 ±62686.611	-6062.095 ±3640.572	-5696.623 ±3447.326	-6513.429 ±4105.793	-4050.303 ±1935.591	-2694.396 ±1947.158
[8] 10-05/02-04	-20272.779 ±8324.498	-7793.173 ±4697.968	-7250.986 ±4384.419	-9417.417 ±5085.291	-5621.554 ±3034.293	-3427.091 ±2684.74
[9] 05-05/00-00	-19398.468 ±34606.47	-2305.757 ±1938.863	-1789.912 ±1758.392	-2055.925 ±1103.224	-1330.574 ±676.937	-700.004 ±349.403
[10] 05-05/04-01	-28823.419 ±113422.157	-3227.877 ±2454.126	-2519.844 ±1883.578	-2848.914 ±1525.574	-1751.345 ±1060.578	-1051.128 ±747.135
[11] 05-05/02-02	-28622.76 ±110354.623	-3438.425 ±2833.692	-3143.414 ±2351.135	-4041.9 ±3268.669	-1976.299 ±1228.007	-1443.55 ±1484.414
[12] 05-05/01-04	-11777.671 ±14085.244	-4685.229 ±4341.722	-4199.314 ±4013.71	-4780.083 ±4166.093	-2730.864 ±2445.418	-1736.92 ±1798.979
[13] 50-100/00-00	-1257331.825 ±308937.915	-355892.338 ±151672.582	-325222.016 ±130481.794	-364546.255 ±148664.96	-306582.66 ±129194.408	-261170.435 ±94171.132
[14] 50-100/40-20	-2255020.691 ±633561.414	-731954.537 ±284458.345	-704072.011 ±280534.657	-757155.657 ±282213.688	-639744.118 ±287614.209	-580154.595 ±262623.914
[15] 50-100/20-40	-2590867.418 ±1281171.743	-768202.866 ±301374.636	-734107.486 ±301752.962	-795506.24 ±309655.323	-707182.041 ±317691.224	-705998.004 ±282489.842
[16] 50-100/10-80	-3070735.514 ±2630685.684	-785830.284 ±292660.66	-753171.447 ±301435.763	-791814.008 ±296876.412	-700479.853 ±296072.681	-701478.072 ±278756.339
[17] 100-50/00-00	-551921.888 ±110904.374	-199322.363 ±38359.901	-185020.921 ±36480.042	-202130.626 ±39263.829	-179930.323 ±35489.188	-140730.818 ±28608.319
[18] 100-50/80-10	-1254543.741 ±2363360.229	-392014.266 ±59648.128	-370599.495 ±61253.793	-397243.669 ±64721.904	-362472.367 ±64245.071	-302444.002 ±64011.316
[19] 100-50/40-20	-1090197.158 ±289913.686	-414742.184 ±62881.094	-397612.661 ±63181.149	-418799.901 ±72250.379	-389616.609 ±68141.797	-347931.634 ±64194.595
[20] 100-50/20-40	-1172308.591 ±595559.971	-423341.755 ±71168.883	-405742.524 ±61693.184	-429953.581 ±71460.035	-394205.792 ±71360.477	-355645.054 ±63971.428
[21] 50-50/00-00	-632025.987 ±167774.571	-169874.611 ±75057.826	-159798.107 ±62601.958	-178473.78 ±69737.786	-152451.514 ±60593.095	-126512.66 ±41824.531
[22] 50-50/40-10	-1014168.804 ±298636.132	-310163.182 ±135344.243	-293205.167 ±131842.976	-339044.249 ±127089.706	-293213.942 ±128669.775	-243932.913 ±118322.477
[23] 50-50/20-20	-1118594.09 ±487601.964	-339604.812 ±139915.754	-337249.907 ±130127.655	-356240.275 ±135117.492	-299186.241 ±134296.581	-288145.72 ±121767.54
[24] 50-50/10-40	-1371539.492 ±1369523.232	-354574.11 ±129806.434	-352295.307 ±120280.536	-368372.308 ±122667	-308754.331 ±121410.208	-314135.159 ±110819.205

6.6 Final remarks

As highlighted in the Subsection 6.1, determining weights for each attribute of a mixed dataset in order to highlight relevant ones and minimizing the impact of those that are irrelevant is the main challenge of feature weighting approaches, especially because the choice of these weights relies on the characteristics of each dataset. As could be seen in this chapter, many scientific works have already addressed this problem, but for other branches of clustering algorithms, not for the Kamila algorithm, a model-based mixed

Table 36 – Results of experiment C - Average and standard deviation of the value of the objective function of the Kamila algorithm (the greater the better). Bold values indicate the best method with respect to performance (solution quality).

Dataset scenario	Baseline			Proposed algorithms		
	KAMBSLN	KAMGAFW	KAMBRKG	KAMILSF	KAMECSF	KAMECBR
[1] 05-05/02-02	-1326 ± 72.414	-90.22 ± 38.059	-90.559 ± 34.83	-110.478 ± 40.231	-54.751 ± 21.585	-41.949 ± 19.933
[2] 05-05/02-02	-2673.655 ± 102.386	-179.354 ± 67.861	-193.565 ± 77.487	-217.687 ± 92.243	-113.733 ± 44.538	-86.094 ± 44.084
[3] 05-05/02-02	-5361.298 ± 146.406	-389.687 ± 150.05	-386.957 ± 145.997	-448.336 ± 177.904	-233.529 ± 93.438	-178.065 ± 85.976
[4] 05-05/02-02	-8050.401 ± 184.663	-580.501 ± 243.631	-586.181 ± 240.705	-638.374 ± 241.615	-347.021 ± 137.872	-259.365 ± 119.337
[5] 05-05/02-02	-13395.583 ± 251.533	-948.97 ± 369.764	-999.09 ± 399.185	-1148.511 ± 411.612	-563.376 ± 221.693	-411.996 ± 184.015

Figure 48 – Experiment A - Normalized average values of the objective function of the Kamila algorithm for all dataset

data clustering method.

Hence, this chapter proposed three meta-heuristics to address the feature weighting problem, comparing them to three methods that have already been proposed in the literature. The methodology was composed of three experiments, which simulated 53

Figure 49 – Experiment B - Normalized average values of the objective function of the Kamila algorithm for all dataset

datasets with many characteristics in order to analyze the behavior of the algorithms in different scenarios. We defined all the main parameters of the algorithms using the irace package. The results showed that the hybridized meta-heuristic performed better than the other algorithms in almost all the datasets. Such hybridization took in place the BRKGA, ECS and ILS methods.

Part V

Achievements and paths to advance

7 CONCLUSION, RESULTS AND FUTURE RESEARCH LINES

“An expert is a person who has made all the mistakes that can be made in a very narrow field.” - Niels Bohr

*“Buscas ouro nativo entre a ganga da vida. Que esperança infinita no ilusório trabalho...
Para cada pepita, quanto cascalho” - Helena Kolody*

In the following sections, one brings the conclusion of this thesis, the scientific documents that have been published over the year of the development of this work, and the future research on mixed data clustering technique.

7.1 Conclusion

In this thesis, the mixed data clustering problem was addressed. Such a technique is widely demanded by many real-world problems in several areas, which makes it an interesting topic of study. As the goal of this thesis was to provide further contributions for the advancement of this thematic, many objectives were stated in section 1, which contemplated theoretical, methodology, and practical proposals. The statement the announced objectives and a brief conclusion of them are presented as follows.

Objective 1. To create a standard representation string that characterizes a scientific document in the knowledge-discovery field by the description of the main choices the authors made in their studies.

In Chapter 3, we could see that there was no proposal in the literature about a standard representation that would gather information from a scientific document so that would characterize it quickly. Therefore, we looked in other areas of knowledge for similar strings of standard representation. We found in the scheduling problem and in the queuing problem two proposals that served as an inspiration to make an initial idea proposal for a standard representation of documents published in the knowledge-discovery field. Thus, this objective was achieved and as a consequence, countless research opportunities arose, which are pointed out in a section ahead;

Objective 2. To provide a multi-view of the past of the mixed data clustering technique throughout a wide and in-depth systematic literature review, and bibliometric analysis.

In Chapter 4, we could see that the literature reviews proposed so far on the topic of mixed data clustering presented a very specific cut of the thematic, or did not present a systematic method to perform the search, selection, and analysis of documents. Thus, we combined these two gaps and proposed a systematic review of the literature, which

included a bibliometric and individual analysis of the 160 documents resulted from the search and selection steps defined in the methodological procedure. We used the standard representation defined in the previous chapter to characterize each document, in addition, we briefly described them, and tabulated the following information: name of the proposed method, the name of the approach the document inspired to make its contributions, which strategy to deal with mixed data the document used, if the document used a set of real and/or simulated data, and what were they, what performance measures did the document use, what programming language was used to implement the proposal (when applicable), and whether the proposed code was made available and, when it was, the address of that code.

Objective 3. To propose the modeling of an optimization method for the feature balancing problem, which requires the definition of the weight for each type of features, and provide a wide and in-depth analysis of the behavior of the considered methods facing different characteristics of datasets.

In Chapter 5, we could see that, until that moment, few studies had addressed the feature balancing in the literature, and none of them had proposed the application of a meta-heuristic for such a challenge. In this way, we gather studies and discussions on feature balancing in mixed data clustering algorithms and model the BRKGA meta-heuristic for the problem. We chose the k-prototype and Kamila algorithms as the clustering algorithms and simulated 423 datasets with different characteristics. Our results showed the proposed algorithms performed better than the others. In this way, this objective was achieved, opening many research opportunities, which are pointed out in a section ahead.

Objective 4. To propose a set of optimization methods for the feature weighting problem, which requires the definition of the weights of each attribute belonging to the mixed datasets, and provide a wide and in-depth analysis of the behavior of the considered methods facing different characteristics of datasets.

In Chapter 6, we could see that the problem of feature weighting was little addressed in the literature until that moment, with no publication addressing this problem in model-based data clustering algorithms. In this way, we modeled three meta-heuristics, one of which was a hybridization of the first two and BRKGA. Our results showed that the hybridized method achieved the best performance in almost all the 53 datasets, in only two of them such an algorithm did not achieve the best result. In this chapter, we also simulated datasets with different characteristics, overlapping of the numerical and categorical attributes, in addition to considering the normal and lognormal distribution. Hence, many opportunities for future research are opened, which are pointed out in a section ahead.

7.2 Results

The following documents were published in indexed journals and books throughout the development of this thesis.

- MARTARELLI, NÁDIA JUNQUEIRA; NAGANO, MARCELO SEIDO . A constructive evolutionary approach for feature selection in unsupervised learning. *Swarm and Evolutionary Computation*, v. 42, p. 125-137, 2018. (Impact factor: 6.912)
- MARTARELLI, NÁDIA JUNQUEIRA; NAGANO, MARCELO SEIDO . Unsupervised feature selection based on bio-inspired approaches. *Swarm and Evolutionary Computation*, v. 52, p. 100618, 2020. (Impact factor: 6.912)
- MARTARELLI, NÁDIA JUNQUEIRA; NAGANO, MARCELO SEIDO . Optimization of the Numeric and Categorical Attribute Weights in KAMILA Mixed Data Clustering Algorithm. In: Yin H., Camacho D., Tino P., Tallón-Ballesteros A., Menezes R., Allmendinger R.. (Org.). *Lecture Notes in Computer Science*. 1ed.: Springer International Publishing, 2019, v. 11871, p. 20-27.

7.3 Future research lines

In this section, we summarize the future research lines we described in the respected chapters.

- From the standard representation proposal - Many future works can base their motivation on the standard representation proposal. Some future research lines are: using machine learning algorithms to automatically recognize the information of the string and fill it with the appropriate acronyms; study more deeply a better solution to fill the string's components, instead of using acronyms; proposing a way to ensure each acronym will represent a unique word or a unique set of words; proposing a methodological procedure to use such a string in the review papers.
- From the systematic literature review - From the multi-views that the systematic literature review provided, the following topics can be explored in future research. Fuzzy mixed data clustering and the usage of optimization in mixed data clustering approaches are two themes classified as emerging by the bibliometric analysis. Therefore, addressing efforts toward these approaches can be a very important decision to the advance of this topic. Developing approaches belonging to the preprocessing and post-processing tasks of KDD process to support the mixed data clustering technique, such as noisy reduction, transformation methods, and visualization approaches, is another path for future research. In addition, proposing

benchmark datasets to be used exclusively by the mixed data clustering approaches is another good strategy to contribute in this area.

- From the feature balancing and weighting problem - Among many opportunities to go forward in this theme are the implementation of new meta-heuristics for both problems, the usage of different categories of performance measures, and dataset. In addition, the development of feature weighting methods for other types of data clustering algorithms, and the development of new performance measure specific to mixed datasets.

7.4 Acknowledgment

This thesis was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, under the grant number 88882.379080/2019-01.

REFERENCES

- ABIDIN, F. N. Z.; WESTHEAD, D. R. Flexible model-based clustering of mixed binary and continuous data: Application to genetic regulation and cancer. **Nucleic Acids Research**, v. 45, n. 7, 2017.
- AHMAD, A.; DEY, L. A k-mean clustering algorithm for mixed numeric and categorical data. **Data and Knowledge Engineering**, v. 63, n. 2, p. 503–527, 2007.
- AHMAD, A.; DEY, L. A k-means type clustering algorithm for subspace clustering of mixed numeric and categorical datasets. **Pattern Recognition Letters**, v. 32, n. 7, p. 1062–1069, 2011.
- AHMAD, A.; HASHMI, S. K-harmonic means type clustering algorithm for mixed datasets. **Applied Soft Computing Journal**, v. 48, p. 39–49, 2016.
- AHMAD, A.; KHAN, S. S. Survey of state-of-the-art mixed data clustering algorithms. **IEEE Access**, v. 7, p. 31883–31902, 2019.
- AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control**, v. 19, n. 6, p. 716–723, 1974.
- AL-SHAQSI, J.; WANG, W. A clustering ensemble method for clustering mixed data. *In: . [S.l.: s.n.]*, 2010.
- ALABI, G. Bradford's law and its application. **International Library Review**, Routledge, v. 11, n. 1, p. 151–158, 1979.
- ALQURASHI, T.; WANG, W. Clustering ensemble method. **International Journal of Machine Learning and Cybernetics**, v. 10, n. 6, p. 1227–1246, 2019.
- AMORIM, R. C. d.; MAKARENKOV, V. Applying subclustering and lp distance in weighted k-means with distributed centroids. **Neurocomputing**, v. 173, p. 700–707, 2016.
- ARIA, M.; CUCCURULLO, C. bibliometrix: An r-tool for comprehensive science mapping analysis. **Journal of Informetrics**, v. 11, n. 4, p. 959 – 975, 2017.
- AUDIGIER, V. *et al.* Multiple imputation for multilevel data with continuous and binary variables. **Statistical Science**, v. 33, n. 2, p. 160–183, 2018.
- BACHER, J. A probabilistic clustering model for variables of mixed type. **Quality and Quantity**, v. 34, n. 3, p. 223–235, 2000.
- BALAJI, K.; LAVANYA, K. Clustering algorithms for mixed datasets: A review. **International Journal of Pure and Applied Mathematics**, v. 118, n. 7, p. 547–556, 2018.
- BARCELÓ-RICO, F.; DIEZ, J.-L. Geometrical codification for clustering mixed categorical and numerical databases. **Journal of Intelligent Information Systems**, v. 39, n. 1, p. 167–185, 2012.

BARCELÓ-RICO, F.; DÍEZ, J. L.; BONDIA, J. A comparative study of codification techniques for clustering heart disease database. **Biomedical Signal Processing and Control**, v. 6, n. 1, p. 64–69, 2011.

BASFORD, K. E.; MCLACHLAN, G. J. The mixture method of clustering applied to three-way data. **Journal of Classification**, v. 2, n. 1, p. 109–125, 1985.

BATHLA, G.; AGGARWAL, H.; RANI, R. A novel approach for clustering big data based on mapreduce. **International Journal of Electrical and Computer Engineering**, v. 8, n. 3, p. 1711–1719, 2018.

BAUDRY, J.-P. *et al.* Enhancing the selection of a model-based clustering with external categorical variables. **Advances in Data Analysis and Classification**, v. 9, n. 2, p. 177–196, 2015.

BERNARD, J. *et al.* Visual-interactive exploration of interesting multivariate relations in mixed research data sets. **Computer Graphics Forum**, v. 33, n. 3, p. 291–300, 2014.

BHATNAGAR, V.; KAUR, S.; CHAKRAVARTHY, S. Clustering data streams using grid-based synopsis. **Knowledge and Information Systems**, v. 41, p. 127–152, 2013.

BHATNAGAR, V. *et al.* Dasc: data aware algorithm for scalable clustering. **Knowledge and Information Systems**, v. 50, n. 3, p. 851–881, 2017.

BIERNACKI, C.; LOURME, A. Unifying data units and models in (co-)clustering. **Advances in Data Analysis and Classification**, v. 13, n. 1, p. 7–31, 2019.

BISHNU, P. S.; BHATTACHERJEE, V. Software cost estimation based on modified k-modes clustering algorithm. **Natural Computing**, v. 15, n. 3, p. 415–422, 2016.

BISWAS, G.; WEINBERG, J.; LI, C. Iterate: A conceptual clustering method for knowledge discovery in databases. *In: In Braunschweig, B., and Day, R. Eds., Innovative Applications of Artificial Intelligence in the Oil and Gas Industry. [S.l.: s.n.]: Editions Technip, 1994. p. 111–139.*

BLOMSTEDT, P. *et al.* A bayesian predictive model for clustering data of mixed discrete and continuous type. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 37, n. 3, p. 489–498, 2015.

BOCK, H.-H. Origins and extensions of the k-means algorithm in cluster analysis. **Journal Électronique d'Histoire des Probabilités et de la Statistique [electronic only]**, v. 4, n. 2, 2008.

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. Recent advances and emerging challenges of feature selection in the context of big data. **Knowledge-Based Systems**, Elsevier Science Publishers B. V., v. 86, p. 33–45, 2015. ISSN 0950-7051.

BRACHMAN, R. J.; ANAND, T. The process of knowledge discovery in databases. *In: FAYYAD, U.; PIATETSKY-SHAPIO, P. S. G.; UTHURUSAMY, R. (ed.). Advances in Knowledge Discovery and Data Mining. USA: American Association for Artificial Intelligence, 1996. p. 37–57. ISBN 0262560976.*

BRAGATO, G.; GANIS, P.; FEOLI, E. Application of burnaby's and goodall's similarity indexes for local soil classification. **Catena**, v. 180, p. 169–182, 2019.

BRIDA, J. *et al.* Clustofvar and the segmentation of cruise passengers from mixed data: Some managerial implications. **Knowledge-Based Systems**, v. 70, p. 128–136, 2014.

BROUWER, R. Clustering feature vectors with mixed numerical and categorical attributes. **International Journal of Computational Intelligence Systems**, v. 1, n. 4, p. 285–298, 2008.

BUSHEL, P. R.; WOLFINGER, R. D.; GIBSON, G. Simultaneous clustering of gene expression data with clinical chemistry and pathological evaluations reveals phenotypic prototypes. **BMC Systems Biology**, v. 1, 2007.

CHAPELLE, O.; SCHLKOPF, B.; ZIEN, A. **Semi-Supervised Learning**. [*S.l.: s.n.*]: The MIT Press, 2010. ISBN 0262514125.

CHATZIS, S. P. A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. **Expert Systems with Applications**, v. 38, n. 7, p. 8684–8689, 2011.

CHEN, J.-Y.; HE, H.-H. A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data. **Information Sciences**, v. 345, p. 271–293, 2016.

CHENG, Y.-M.; LEU, S.-S. Constraint-based clustering model for determining contract packages of bridge maintenance inspection. **Automation in Construction**, v. 17, n. 6, p. 682–690, 2008.

CHENG, Y.-M.; LEU, S.-S. Constraint-based clustering and its applications in construction management. **Expert Systems with Applications**, v. 36, n. 3, p. 5761–5767, 2009.

CHENG, Y.-M.; LEU, S.-S. Integrating data mining with kj method to classify bridge construction defects. **Expert Systems with Applications**, v. 38, n. 6, p. 7143–7150, 2011.

CHEUNG, Y.-M.; JIA, H. Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. **Pattern Recognition**, v. 46, n. 8, p. 2228–2238, 2013.

CHITRADEVI, N. *et al.* A novel distance for clustering to support mixed data attributes and promote data reliability and network lifetime in large scale wireless sensor networks. *In: . [S.l.: s.n.]*, 2012. v. 30, p. 669–677.

Chung-Chian Hsu; Kuo-Min Wang; Sheng-Hsuan Wang. Gvisom for multivariate mixed data projection and structure visualization. *In: The 2006 IEEE International Joint Conference on Neural Network Proceedings*. [*S.l.: s.n.*], 2006. p. 3300–3305.

COBO, M. *et al.* An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. **Journal of Informetrics**, v. 5, n. 1, p. 146 – 166, 2011. ISSN 1751-1577.

COSO, C. D. *et al.* Mixing numerical and categorical data in a self-organizing map by means of frequency neurons. **Applied Soft Computing Journal**, v. 36, p. 246–254, 2015.

COX, T. F.; COX, M. A. A. A general weighted two-way dissimilarity coefficient. **Journal of Classification**, v. 17, n. 1, p. 101–121, 2000.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, PAMI-1, n. 2, p. 224–227, 1979. ISSN 0162-8828.

DAVOODI, R.; MORADI, M. H. Mortality prediction in intensive care units (icus) using a deep rule-based fuzzy classifier. **Journal of Biomedical Informatics**, v. 79, p. 48–59, 2018.

DENG, X. *et al.* Internal differentiation within the rural migrant population from the sustainable urban development perspective: Evidence from china. **Sustainability (Switzerland)**, v. 10, n. 12, 2018.

DING, S. *et al.* An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood. **Knowledge-Based Systems**, v. 133, p. 294–313, 2017.

DRAKOPOULOS, G. *et al.* A fuzzy graph framework for initializing k-means. **International Journal on Artificial Intelligence Tools**, v. 25, n. 6, 2016.

DU, M.; DING, S.; XUE, Y. A novel density peaks clustering algorithm for mixed data. **Pattern Recognition Letters**, v. 97, p. 46–53, 2017.

DUAN, B. *et al.* Clustering mixed data based on density peaks and stacked denoising autoencoders. **Symmetry**, v. 11, n. 2, 2019.

DUBES, R.; JAIN, A. K. Validity studies in clustering methodologies. **Pattern Recognition**, v. 11, n. 4, p. 235 – 254, 1979. ISSN 0031-3203.

DUNN, J. C. Well separated clusters and optimal fuzzy-partitions. **Journal of Cybernetics**, v. 4, p. 95–104, 1974.

DUTTA, D.; SIL, J.; DUTTA, P. Automatic clustering by multi-objective genetic algorithm with numeric and categorical features. **Expert Systems with Applications**, v. 137, p. 357–379, 2019.

ESTER, M. *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. *In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.]: AAAI Press, 1996. p. 226–231.

FAYYAD, U. M.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery: an overview. *In: FAYYAD, U. M. et al. (ed.). Advances in knowledge discovery and data mining*. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. p. 1–34. ISBN 0-262-56097-6.

FAYYAD, U. M.; PIATETSKY-SHAPIO, G.; SMYTH, P. Knowledge discovery and data mining : Towards a unifying framework. *In: Proceedings of the 2nd international conference on Knowledge Discovery and Data mining (KDD'96)*. [S.l.: s.n.]: AAAI Press, 1996. p. 82–88.

FAYYAD, U. M. *et al.* (ed.). **Advances in Knowledge Discovery and Data Mining**. USA: American Association for Artificial Intelligence, 1996. ISBN 0262560976.

FOGGIA, P. *et al.* A graph-based clustering method and its applications. *In: Advances in Brain, Vision, and Artificial Intelligence*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 277–287.

FONSECA, J. R. S. Clustering in the field of social sciences: that is your choice. **International Journal of Social Research Methodology**, v. 16, n. 5, p. 403–428, 2013.

FOSS, A. *et al.* A semiparametric method for clustering mixed data. **Machine Learning**, v. 105, n. 3, p. 419–458, 2016.

FOSS, A. H.; MARKATOU, M. kamila: Clustering mixed-type data in r and hadoop. **Journal of Statistical Software**, v. 83, 2018.

FOSS, A. H.; MARKATOU, M.; RAY, B. Distance metrics and clustering methods for mixed-type data. **International Statistical Review**, v. 87, n. 1, p. 80–109, 2019.

FOWLKES, E.; MALLOWS, C. A method for comparing two hierarchical clustering. **Journal of the American Statistical Association**, v. 78, p. 553–569, 1983.

FRAZER, K. A. *et al.* A second generation human haplotype map of over 3.1 million snps. **Nature**, v. 449, n. 7164, p. 851–861, 2007.

FRIEDMAN, J. H.; MEULMAN, J. J. Clustering objects on subsets of attributes. **Journal of the Royal Statistical Society. Series B: Statistical Methodology**, v. 66, n. 4, p. 815–839, 2004.

FRISVAD, J. C. Correspondence, principal coordinate, and redundancy analysis used on mixed chemotaxonomical qualitative and quantitative data. **Chemometrics and Intelligent Laboratory Systems**, v. 23, n. 1, p. 213–229, 1994.

GARFIELD, E. Keywords plus: Isi's breakthrough retrieval method. part 1. expanding your searching power on current contents on diskette. **Current Contents**, v. 32, p. 5–9, 1990.

GARFIELD, E. Keywords plus takes you beyond title words. part 2. expanded journal coverage for current contents on diskette, includes social and behavioral sciences. **Current Contents**, v. 33, p. 5–9, 1990.

GARFIELD, E.; SHER, I. H. Keywords plus - algorithmic derivative indexing. **JASIS**, v. 44, p. 298–299, 1993.

GOCSIK, . *et al.* Elicitation of preferences of dutch broiler and pig farmers to support decision making on animal welfare. **NJAS - Wageningen Journal of Life Sciences**, v. 76, p. 75–86, 2016.

Golinko, E.; Zhu, X. Gfel: Generalized feature embedding learning using weighted instance matching. *In: 2017 IEEE International Conference on Information Reuse and Integration (IRI)*. [S.l.: s.n.], 2017. p. 235–244.

GOLINKO, E.; ZHU, X. Generalized feature embedding for supervised, unsupervised, and online learning tasks. **Information Systems Frontiers**, v. 21, n. 1, p. 125–142, 2019.

GONDEK, D.; HOFMANN, T. Non-redundant data clustering. **Knowledge and Information Systems**, v. 12, n. 1, p. 1–24, 2007.

GORDON, A. Classification and assignment in soil science. **Soil Use and Management**, v. 3, n. 1, p. 3–8, 1987.

GORDON, A. D. Identifying genuine clusters in a classification. **Computational Statistics and Data Analysis**, v. 18, n. 5, p. 561 – 581, 1994. ISSN 0167-9473.

GOURGARIS, P.; MAKRIS, C. A density based k-means initialization scheme. *In: Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*. New York, NY, USA: [S.l.: s.n.], 2015. p. 19:1–19:9.

GOVAERT, G.; NADIF, M. **Co-clustering : models, algorithms and applications**. [S.l.: s.n.], 2013. ISBN 978-1-848-21473-6.

GOWER, J. C. A general coefficient of similarity and some of its properties. **Biometrics**, v. 27, n. 4, p. 857–871, 1971.

GRAHAM, R. L. *et al.* Optimization and approximation in deterministic sequencing and scheduling: a survey. **Annals of discrete mathematics**, v. 5, n. 2, p. 287–326, 1979.

GU, H. *et al.* Optimized scheme in coal-fired boiler combustion based on information entropy and modified k-prototypes algorithm. **Results in Physics**, v. 9, p. 1262–1274, 2018.

GUGLIELMI, A. *et al.* A semiparametric bayesian joint model for multiple mixed-type outcomes: An application to acute myocardial infarction. **Advances in Data Analysis and Classification**, v. 12, n. 2, p. 399–423, 2018.

HAI, N.; SUSUMU, H. Performances of parallel clustering algorithm for categorical and mixed data. *In: . [S.l.: s.n.]*, 2004. v. 3320, p. 252–256.

HAIKACEM, M. B.; N'CIR, C.-E.; ESSOUSSI, N. One-pass mapreduce-based clustering method for mixed large scale data. **Journal of Intelligent Information Systems**, v. 52, n. 3, p. 619–636, 2019.

HAN, J. *et al.* Mage: A semantics retaining k-anonymization method for mixed data. **Knowledge-Based Systems**, v. 55, p. 75–86, 2014.

HARCH, B. D. *et al.* Mixed data types and the use of pattern analysis on the australian groundnut germplasm data. **Genetic Resources and Crop Evolution**, v. 43, n. 4, p. 363–376, 1996.

HARCH, B. D. *et al.* The analysis of large scale data taken from the world groundnut (*arachis hypogaea* l.) germplasm collection. ii. two-way data with mixed data types. **Euphytica**, v. 105, n. 2, p. 73–82, 1999.

HARIKUMAR, S.; SURYA, P. V. K-medoid clustering for heterogeneous datasets. *In: Procedia Computer Science*. [S.l.: s.n.], 2015. v. 70, p. 226–237.

HARTIGAN, J. A.; WONG, M. A. A k-means clustering algorithm. **JSTOR: Applied Statistics**, v. 28, n. 1, p. 100–108, 1979.

HE, Z.; XU, X.; DENG, S. Squeezer: An efficient algorithm for clustering categorical data. **Journal of Computer Science and Technology**, v. 17, n. 5, p. 611–624, 2002.

HE, Z.; XU, X.; DENG, S. Scalable algorithms for clustering large datasets with mixed type attributes. **International Journal of Intelligent Systems**, v. 20, n. 10, p. 1077–1089, 2005.

HENNIG, C. What are the true clusters? **Pattern Recognition Letters**, v. 64, p. 53 – 62, 2015. ISSN 0167-8655. Philosophical Aspects of Pattern Recognition.

HENNIG, C.; LIAO, T. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. **Journal of the Royal Statistical Society. Series C: Applied Statistics**, v. 62, n. 3, p. 309–369, 2013.

HENNIG, C.; LIN, C.-J. Flexible parametric bootstrap for testing homogeneity against clustering and assessing the number of clusters. **Statistics and Computing**, v. 25, n. 4, p. 821–833, 2015.

HERR, A. Statistics for categorical surveys-a new strategy for multivariate classification and determining variable importance. **Sustainability**, v. 2, n. 2, p. 533–550, 2010.

HONG, J. *et al.* Development of ev charging templates: An improved k-prototypes method. **IET Generation, Transmission and Distribution**, v. 12, n. 20, p. 4361–4367, 2018.

HORTA, I.; CAMANHO, A. Competitive positioning and performance assessment in the construction industry. **Expert Systems with Applications**, v. 41, n. 4 PART 1, p. 974–983, 2014.

HOWARD, P.; HOWARD, D. Classification and dissection of environmental data using qualitative and mixed data types. **Journal of Environmental Management**, v. 26, n. 4, p. 313–319, 1988.

HSU, C.-C.; CHEN, C.-L.; SU, Y.-W. Hierarchical clustering of mixed data based on distance hierarchy. **Information Sciences**, v. 177, n. 20, p. 4474–4492, 2007.

HSU, C.-C.; CHEN, Y.-C. Mining of mixed data with application to catalog marketing. **Expert Systems with Applications**, v. 32, n. 1, p. 12–23, 2007.

HSU, C.-C.; HUANG, W.-H. Integrated dimensionality reduction technique for mixed-type data involving categorical values. **Applied Soft Computing Journal**, v. 43, p. 199–209, 2016.

HSU, C.-C.; HUANG, Y.-P. Incremental clustering of mixed data based on distance hierarchy. **Expert Systems with Applications**, v. 35, n. 3, p. 1177–1185, 2008.

HSU, C.-C.; LIN, S.-H. Visualized analysis of mixed numeric and categorical data via extended self-organizing map. **IEEE Transactions on Neural Networks and Learning Systems**, v. 23, n. 1, p. 72–86, 2012.

HSU, C.-C.; LIN, S.-H.; TAI, W.-S. Apply extended self-organizing map to cluster and classify mixed-type data. **Neurocomputing**, v. 74, n. 18, p. 3832–3842, 2011.

- HSU, C.-C.; WANG, S.-H. An integrated framework for visualized and exploratory pattern discovery in mixed data. **IEEE Transactions on Knowledge and Data Engineering**, v. 18, n. 2, p. 161–173, 2006.
- HSU, W.-H.; JAO, J.-A.; CHEN, Y.-L. Discovering conjecturable rules through tree-based clustering analysis. **Expert Systems with Applications**, v. 29, n. 3, p. 493–505, 2005.
- HUANG, J. Z. *et al.* Automated variable weighting in k-means type clustering. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 27, n. 5, p. 657–668, 2005.
- HUANG, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. **Data Mining and Knowledge Discovery**, v. 2, n. 3, p. 283–304, 1998.
- HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of classification**, Springer, v. 2, n. 1, p. 193–218, 1985.
- HUBERT, L.; LEVIN, J. **A General Statistical Framework for Assessing Categorical Clustering in Free Recall**. [*S.l.*: *s.n.*]: Wisconsin Research and Development Center for Cognitive Learning, 1975. (Theoretical papers).
- HUERTA-MUÑOZ, D. L.; RÍOS-MERCADO, R. Z.; RUIZ, R. An iterated greedy heuristic for a market segmentation problem with multiple attributes. **European Journal of Operational Research**, v. 261, n. 1, p. 75–87, 2017.
- HUMMEL, M.; EDELMANN, D.; KOPP-SCHNEIDER, A. Clustering of samples and variables with mixed-type data. **PLoS ONE**, v. 12, n. 11, 2017.
- HUNT, L.; BASFORD, K. Fitting a mixture model to three-mode three-way data with categorical and continuous variables. **Journal of Classification**, v. 16, n. 2, p. 283–296, 1999.
- HUNT, L.; JORGENSEN, M. Mixture model clustering for mixed data with missing information. **Computational Statistics and Data Analysis**, v. 41, n. 3-4, p. 429–440, 2003.
- HUNT, L.; JORGENSEN, M. Clustering mixed data. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 1, n. 4, p. 352–361, 2011.
- HUNT, L. A. **Finite mixture models**. 1996. Ph.D. Thesis — University of Waikato, 1996.
- HUNT, L. A.; BASFORD, K. E. Comparing classical criteria for selecting intra-class correlated features in multimix. **Computational Statistics and Data Analysis**, v. 103, p. 350–366, 2016.
- IAM-ON, N.; BOONGOEN, T. Improved student dropout prediction in thai university using ensemble of mixed-type data clusterings. **International Journal of Machine Learning and Cybernetics**, v. 8, n. 2, p. 497–510, 2017.
- ICHINO, M.; SKLANSKY, J. The relative neighborhood graph for mixed feature variables. **Pattern Recognition**, v. 18, n. 2, p. 161–167, 1985.

-
- INGRASSIA, S. *et al.* The generalized linear mixed cluster-weighted model. **Journal of Classification**, v. 32, n. 1, p. 85–113, 2015.
- IRIGOIEN, I.; ARENAS, C. Inca: New statistic for estimating the number of clusters and identifying atypical units. **Statistics in Medicine**, v. 27, n. 15, p. 2948–2973, 2008.
- IRIGOIEN, I.; SIERRA, B.; ARENAS, C. Icge: An r package for detecting relevant clusters and atypical units in gene expression. **BMC Bioinformatics**, v. 13, n. 1, 2012.
- JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. Englewood Cliffs, N.J., USA: Prentice Hall, 1988. 320 p. ISBN 013022278X.
- JANG, H.-J. *et al.* An efficient grid-based k-prototypes algorithm for sustainable decision-making on spatial objects. **Sustainability (Switzerland)**, v. 10, n. 8, 2018.
- JI, J. *et al.* An improved k-prototypes clustering algorithm for mixed numeric and categorical data. **Neurocomputing**, v. 120, p. 590–596, 2013.
- JI, J. *et al.* Clustering mixed numeric and categorical data with artificial bee colony strategy. **Journal of Intelligent and Fuzzy Systems**, v. 36, n. 2, p. 1521–1530, 2019.
- JI, J. *et al.* An initialization method for clustering mixed numeric and categorical data based on the density and distance. **International Journal of Pattern Recognition and Artificial Intelligence**, v. 29, n. 7, 2015.
- JI, J. *et al.* A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. **Knowledge-Based Systems**, v. 30, p. 129–135, 2012.
- JIA, H.; CHEUNG, Y.-M. Subspace clustering of categorical and numerical data with an unknown number of clusters. **IEEE Transactions on Neural Networks and Learning Systems**, v. 29, n. 8, p. 3308–3325, 2018.
- JING, L.; NG, M. K.; HUANG, J. Z. An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. **IEEE Transactions on Knowledge and Data Engineering**, v. 19, n. 8, p. 1026–1041, 2007. ISSN 2326-3865.
- JINYIN, C. *et al.* Fast density clustering algorithm for numerical data and categorical data. **Mathematical Problems in Engineering**, v. 2017, 2017.
- JORGENSEN, M.; HUNT, L. A. Mixture model clustering of data sets with categorical and continuous variables. *In: Proceedings of the Conference ISIS 1996*. [S.l.: s.n.], 1996. p. 375–384.
- KAMPERT, M.; MEULMAN, J.; FRIEDMAN, J. rcosa: A software package for clustering objects on subsets of attributes. **Journal of Classification**, v. 34, n. 3, p. 514–547, 2017.
- KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data: An Introduction to Cluster Analysis**. New Jersey, USA: John Wiley, 1990.
- KENDALL, D. G. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. **Ann. Math. Statist.**, The Institute of Mathematical Statistics, v. 24, n. 3, p. 338–354, 09 1953.
- KIM, B. A fast k-prototypes algorithm using partial distance computation. **Symmetry**, v. 9, n. 4, 2017.

KIM, D.-W.; LEE, K. H.; LEE, D. Fuzzy clustering of categorical data using fuzzy centroids. **Pattern Recognition Letters**, v. 25, n. 11, p. 1263–1271, 2004. ISSN 0167-8655.

KITCHENHAM, B. **Procedures for Performing Systematic Literature Reviews**. [S.l.], 2004. 33 p.

KUPPUSAMY, V.; PARAMASIVAM, I. Grey fuzzy neural network-based hybrid model for missing data imputation in mixed database. **International Journal of Intelligent Engineering and Systems**, v. 10, n. 2, p. 146–155, 2017.

LAM, D.; WEI, M.; WUNSCH, D. Clustering data of mixed categorical and numerical type with unsupervised feature learning. **IEEE Access**, v. 3, p. 1605–1616, 2015.

LAROSE, D. T. **Discovering Knowledge in Data: An Introduction to Data Mining**. USA: Wiley-Interscience, 2005.

LEE, M.; PEDRYCZ, W. The fuzzy c-means algorithm with fuzzy p-mode prototypes for clustering objects having mixed features. **Fuzzy Sets and Systems**, v. 160, n. 24, p. 3590–3600, 2009.

LEE, M.; PEDRYCZ, W. Adaptive learning of ordinal-numerical mappings through fuzzy clustering for the objects of mixed features. **Fuzzy Sets and Systems**, v. 161, n. 4, p. 564–577, 2010.

LI, N. *et al.* Exceptional events classification in warehousing based on an integrated clustering method for a dataset with mixed-valued attributes. **International Journal of Computer Integrated Manufacturing**, v. 31, n. 11, p. 1078–1096, 2018.

LIANG, J. *et al.* Determining the number of clusters using information entropy for mixed data. **Pattern Recognition**, v. 45, n. 6, p. 2251–2265, 2012.

LIANG, W.-B.; WANG, C.-D.; LAI, J.-H. Weighted numerical and categorical attribute clustering in data streams. *In: . [S.l.: s.n.]*, 2017. v. 2017-May, p. 3066–3072.

LIAO, W.-C.; HSU, C.-C. A self-organizing map for transactional data and the related categorical domain. **Applied Soft Computing Journal**, v. 12, n. 10, p. 3141–3157, 2012.

LIN, S.; AZARNOUSH, B.; RUNGER, G. Crafter: A tree-ensemble clustering algorithm for static datasets with mixed attributes and high dimensionality. **IEEE Transactions on Knowledge and Data Engineering**, v. 30, n. 9, p. 1686–1696, 2018.

LIU, S. *et al.* Clustering mixed data by fast search and find of density peaks. **Mathematical Problems in Engineering**, v. 2017, 2017.

LIU, X.; YANG, Q.; HE, L. A novel dbscan with entropy and probability for mixed data. **Cluster Computing**, v. 20, n. 2, p. 1313–1323, 2017.

LÓPEZ-IBÁÑEZ, M. *et al.* The irace package: Iterated racing for automatic algorithm configuration. **Operations Research Perspectives**, v. 3, p. 43 – 58, 2016. ISSN 2214-7160.

LOTKA, A. J. The frequency distribution of scientific productivity. **Journal of the Washington Academy of Sciences**, Washington Academy of Sciences, v. 16, n. 12, p. 317–323, 1926. ISSN 00430439.

LV, Z. *et al.* An optimizing and differentially private clustering algorithm for mixed data in sdn-based smart grid. **IEEE Access**, v. 7, p. 45773–45782, 2019.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. *In: In 5-th Berkeley Symposium on Mathematical Statistics and Probability*. [S.l.: s.n.], 1967. p. 281–297.

MAIONE, C.; NELSON, D.; BARBOSA, R. Research on social data by means of cluster analysis. **Applied Computing and Informatics**, v. 15, n. 2, p. 153–162, 2019.

MALONDKAR, A. *et al.* Spark-ghsom: Growing hierarchical self-organizing map for large scale mixed attribute datasets. **Information Sciences**, v. 496, p. 572–591, 2019.

MARBAC, M.; SEDKI, M. Variable selection for model-based clustering using the integrated complete-data likelihood. **Statistics and Computing**, v. 27, n. 4, p. 1049–1063, 2017.

MARBAC, M.; SEDKI, M. Varsellcm: An r/c++ package for variable selection in model-based clustering of mixed-data with missing values. **Bioinformatics**, v. 35, n. 7, p. 1255–1257, 2019.

MARBAC, M.; SEDKI, M.; PATIN, T. Variable selection for mixed data clustering: Application in human population genomics. **Journal of Classification**, 2019.

MARBAC, M.; VANDEWALLE, V. A tractable multi-partitions clustering. **Computational Statistics and Data Analysis**, v. 132, p. 167–179, 2019.

MARTARELLI, N. J.; NAGANO, M. S. Optimization of the numeric and categorical attribute weights in kamila mixed data clustering algorithm. *In: YIN, H. et al. (ed.). Intelligent Data Engineering and Automated Learning – IDEAL 2019*. [S.l.: s.n.]: Springer International Publishing, 2019. p. 20–27. ISBN 978-3-030-33607-3.

MCLACHLAN, G. J.; CHANG, S. U. Mixture modelling for cluster analysis. **Statistical Methods in Medical Research**, v. 13, n. 5, p. 347–361, 2004.

MCPARLAND, D.; GORMLEY, I. C. Model based clustering for mixed data: clustmd. **Advances in Data Analysis and Classification**, v. 10, n. 2, p. 155–169, 2016.

MCPARLAND, D. *et al.* Clustering high-dimensional mixed data to uncover sub-phenotypes: joint analysis of phenotypic and genotypic data. **Statistics in Medicine**, v. 36, n. 28, p. 4548–4569, 2017.

MICHALSKI, R. S.; STEPP, R. E. Learning from observation: conceptual clustering. *In: MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. (ed.). Machine Learning*. San Francisco (CA): Morgan Kaufmann, 1983. p. 331–363.

MIYAMOTO, S.; HUYNH, V.-N.; FUJIWARA, S. Methods for Clustering Categorical and Mixed Data: An Overview and New Algorithms. **Proceedings of the International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making (IUKM 2011)**, v. 1, p. 75–86, 2018.

MODHA, D. S.; SPANGLER, W. S. Feature weighting in k-means clustering. **Machine Learning**, v. 52, n. 3, p. 217–237, 2003.

MOISE, G.; SANDER, J.; ESTER, M. Robust projected clustering. **Knowledge and Information Systems**, v. 14, n. 3, p. 273–298, 2008.

MOUSTAKI, I.; PAPAGEORGIOU, I. Latent class models for mixed variables with applications in archaeometry. **Computational Statistics and Data Analysis**, v. 48, n. 3, p. 659 – 675, 2005. ISSN 0167-9473.

Najjar, A.; Gagné, C.; Reinharz, D. A novel mixed values k-prototypes algorithm with application to health care databases mining. *In: 2014 IEEE Symposium on Computational Intelligence in Healthcare and e-health (CICARE)*. [S.l.: s.n.], 2014. p. 159–166.

NAJJAR, A. *et al.* A two-step approach for mining patient treatment pathways in administrative healthcare databases. **Artificial Intelligence in Medicine**, v. 87, p. 34–48, 2018.

NG, A. Y.; JORDAN, M. I.; WEISS, Y. On spectral clustering: Analysis and an algorithm. *In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. Cambridge, MA, USA: MIT Press, 2001. p. 849–856.

NG, M. K.; WONG, J. C. Clustering categorical data sets using tabu search techniques. **Pattern Recognition**, v. 35, n. 12, p. 2783–2790, 2002.

NIU, K. *et al.* A coupled user clustering algorithm based on mixed data for web-based learning systems. **Mathematical Problems in Engineering**, v. 2015, 2015.

NOIA, A. D. *et al.* Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction. **Soft Computing**, 2019.

NOORBEHBAHANI, F. *et al.* An incremental intrusion detection system using a new semi-supervised stream classification method. **International Journal of Communication Systems**, v. 30, n. 4, 2017.

NOORBEHBAHANI, F.; MOUSAVI, S.; MIRZAEI, A. An incremental mixed data clustering method using a new distance measure. **Soft Computing**, v. 19, n. 3, p. 731–743, 2014.

NOWOSAD, J.; STEPINSKI, T. Spatial association between regionalizations using the information-theoretical v-measure. **International Journal of Geographical Information Science**, v. 32, n. 12, p. 2386–2401, 2018.

OLIVEIRA, A. C. M.; LORENA, L. A. N. Detecting promising areas by evolutionary clustering search. *In: BAZZAN, A. L. C.; LABIDI, S. (ed.). Advances in Artificial Intelligence – SBIA 2004*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. p. 385–394.

OLIVEIRA, A. C. M.; LORENA, L. A. N. Hybrid evolutionary algorithms and clustering search. *In: _____. Hybrid Evolutionary Algorithms*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 77–99. ISBN 978-3-540-73297-6.

- OVESNÁ, J. *et al.* Analysis of the genetic structure of a barley collection using dna diversity array technology (dart). **Plant Molecular Biology Reporter**, v. 31, n. 2, p. 280–288, 2013.
- ÖZBAKIR, L.; TURNA, F. Clustering performance comparison of new generation meta-heuristic algorithms. **Knowledge-Based Systems**, v. 130, p. 1–16, 2017.
- PAKHIRA, M. K.; BANDYOPADHYAY, S.; MAULIK, U. Validity index for crisp and fuzzy clusters. **Pattern Recognition**, v. 37, p. 487–501, 2004. ISSN 0031-3203.
- PASQUIER, P.-E. D.; NACIRI, Y.; JEANMONOD, D. Morphological analysis of the silene gigantea complex (caryophyllaceae) across the balkan peninsula, south-western turkey and cyprus. **Plant Systematics and Evolution**, v. 301, n. 8, p. 2025–2042, 2015.
- PATHAK, A.; PAL, N. Clustering of mixed data by integrating fuzzy, probabilistic, and collaborative clustering framework. **International Journal of Fuzzy Systems**, v. 18, n. 3, p. 339–348, 2016.
- PETERS, H. P. F.; RAAN, A. F. J. van. Structuring scientific activities by co-author analysis. **Scientometrics**, v. 20, p. 235–255, 1991.
- PHAM, D.-T.; SUAREZ-ALVAREZ, M. M.; PROSTOV, Y. I. Random search with k-prototypes algorithm for clustering mixed datasets. **Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 467, n. 2132, p. 2387–2403, 2011.
- PHILIP, G.; OTTAWAY, B. S. Mixed data cluster analysis: an illustration using cypriot hooked-tang weapons. **Archaeometry**, v. 25, n. 2, p. 119–133, 1983.
- PIATETSKY-SHAPIRO, G. Knowledge discovery in real databases: A report on the ijcai-89 workshop. **AI Magazine**, v. 11, n. 5, p. 68–70, 1991.
- PROSTOV, M.; SUAREZ-ALVAREZ, M.; PROSTOV, Y. Properties of the sample estimators used for statistical normalization of feature vectors. **Data Mining and Knowledge Discovery**, v. 29, n. 6, p. 1815–1837, 2015.
- PUTHPONGSIRIPORN, T. *et al.* Attribute-level neighbor hierarchy construction using evolved pattern-based knowledge induction. **IEEE Transactions on Knowledge and Data Engineering**, v. 18, n. 7, p. 917–929, 2006.
- QIN, B. *et al.* Emu: An expectation maximization based approach for clustering uncertain data. **Journal of Intelligent and Fuzzy Systems**, v. 25, n. 4, p. 1067–1083, 2013.
- RAHMAN, M. G.; ISLAM, M. Z. Missing value imputation using a fuzzy clustering-based em approach. **Knowledge and Information Systems**, v. 46, n. 2, p. 389–422, 2016.
- RAND, W. M. Objective criteria for the evaluation of clustering methods. **Journal of the American Statistical Association**, v. 66, n. 336, p. 846–850, 1971.
- ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, v. 20, p. 53–65, 1987. ISSN 0377-0427.

SANGAM, R.; OM, H. An equi-biased k-prototypes algorithm for clustering mixed-type data. **Sadhana - Academy Proceedings in Engineering Sciences**, v. 43, n. 3, 2018.

SANGAM, R. S.; OM, H. Hybrid data labeling algorithm for clustering large mixed type data. **Journal of Intelligent Information Systems**, v. 45, n. 2, p. 273–293, 2015.

SANGAM, R. S.; OM, H. Equi-clustream: A framework for clustering time evolving mixed data. **Advances in Data Analysis and Classification**, v. 12, n. 4, p. 973–995, 2018.

SAXENA, A. *et al.* A review of clustering techniques and developments. **Neurocomputing**, Elsevier B.V., v. 267, p. 664–681, 2017. Available at: <http://dx.doi.org/10.1016/j.neucom.2017.06.053>.

SCHWARZ, G. Estimating the dimension of a model. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. ISSN 00905364.

SHEU, J.-B. A hybrid fuzzy-optimization approach to customer grouping-based logistics distribution operations. **Applied Mathematical Modelling**, v. 31, n. 6, p. 1048–1066, 2007.

SHI, T.; HORVATH, S. Unsupervised learning with random forest predictors. **Journal of Computational and Graphical Statistics**, v. 15, n. 1, p. 118–138, 2006.

SHIN, H. W.; SOHN, S. Y. Multi-attribute scoring method for mobile telecommunication subscribers. **Expert Systems with Applications**, v. 26, n. 3, p. 363–368, 2004.

SHUKLA, U.; NANDA, S. Designing of a risk assessment model for issuing credit card using parallel social spider algorithm. **Applied Artificial Intelligence**, v. 33, n. 3, p. 191–207, 2019.

SHUKLA, U. P.; NANDA, S. J. Parallel social spider clustering algorithm for high dimensional datasets. **Engineering Applications of Artificial Intelligence**, v. 56, p. 75 – 90, 2016. ISSN 0952-1976.

SIEBER, T. N.; PETRINI, O.; GREENACRE, M. J. Correspondence analysis as a tool in fungal taxonomy. **Systematic and Applied Microbiology**, v. 21, n. 3, p. 433–441, 1998.

SOPAN, A. *et al.* Exploring data distributions: Visual design and evaluation. **International Journal of Human-Computer Interaction**, v. 29, n. 2, p. 77–95, 2013.

SPEROTTO, A. *et al.* A labeled data set for flow-based intrusion detection. *In*: NUNZI, G.; SCOGGIO, C.; LI, X. (ed.). **IP Operations and Management**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 39–50. ISBN 978-3-642-04968-2.

STORLIE, C. B. *et al.* Clustering and variable selection in the presence of mixed variable types and missing data. **Statistics in Medicine**, v. 37, n. 19, p. 2884–2899, 2018.

SUAREZ-ALVAREZ, M. *et al.* Statistical approach to normalization of feature vectors and clustering of mixed datasets. **Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 468, n. 2145, p. 2630–2651, 2012.

-
- SZEPANNEK, G. clustmixtype: User-friendly clustering of mixed-type data in r. **The R Journal**, v. 10, n. 2, p. 200–208, 2018.
- TAHA, A.; HADI, A. Pair-wise association measures for categorical and mixed data. **Information Sciences**, v. 346–347, p. 73–89, 2016.
- TAI, W.-S.; HSU, C.-C. Growing self-organizing map with cross insert for mixed-type data clustering. **Applied Soft Computing Journal**, v. 12, n. 9, p. 2856–2866, 2012.
- TAI, W.-S.; HSU, C.-C.; CHEN, J.-C. A mixed-type self-organizing map with a dynamic structure. *In: . [S.l.: s.n.]*, 2010.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367.
- TEKUMALLA, L. S.; RAJAN, V.; BHATTACHARYYA, C. Vine copulas for mixed data: multi-view clustering for mixed data beyond meta-gaussian dependencies. **Machine Learning**, v. 106, n. 9–10, p. 1331–1357, 2017.
- THONG, P. H.; SON, L. H. Picture fuzzy clustering for complex data. **Engineering Applications of Artificial Intelligence**, v. 56, p. 121–130, 2016.
- VALENZUELA, J. *et al.* Health and disease phenotyping in old age using a cluster network analysis. **Scientific Reports**, v. 7, n. 1, 2017.
- VEIGA, F. D. Structure discovery in medical databases: A conceptual clustering approach. **Artificial Intelligence in Medicine**, v. 8, n. 5, p. 473–491, 1996.
- VELDEN, M. van de; D'ENZA, A. I.; MARKOS, A. Distance-based clustering of mixed data. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 11, n. 3, 2019.
- VELDEN, M. van de; ENZA, A. I. D.; MARKOS, A. Distance-based clustering of mixed data. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 11, n. 3, 2019.
- VENDRAMIN, L.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R. Relative clustering validity criteria: A comparative overview. **Statistical Analysis and Data Mining: The ASA Data Science Journal**, v. 3, n. 4, p. 209–235, 2010.
- WANG, C.; MACHIRAJU, R.; HUANG, K. Breast cancer patient stratification using a molecular regularized consensus clustering method. **Methods**, v. 67, n. 3, p. 304–312, 2014.
- WANG, L.; ZHU, D.; DONG, M. Clustering over-dispersed data with mixed feature types. **Statistical Analysis and Data Mining**, v. 11, n. 2, p. 55–65, 2018.
- WANGCHAMHAN, T.; CHIEWCHANWATTANA, S.; SUNAT, K. Efficient algorithms based on the k-means and chaotic league championship algorithm for numeric, categorical, and mixed-type data clustering. **Expert Systems with Applications**, v. 90, p. 146–167, 2017.
- WEI, M.; CHOW, T.; CHAN, R. Clustering heterogeneous data with k-means by mutual information-based unsupervised feature transformation. **Entropy**, v. 17, n. 3, p. 1535–1548, 2015.

WONG, A.; CHIU, D. Synthesizing statistical knowledge from incomplete mixed-mode data. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 9, n. 6, p. 796–805, 1987.

XIONG, H.; LI, Z. Clustering validation measures. *In: _____*. **Data Clustering - Algorithms and Applications**. [S.l.: s.n.]: Chapman and Hall/CRC, 2014. cap. 13.

XU, R.; WUNSCH, D. C. **Clustering**. Hoboken, New Jersey: John Wiley and Sons, Inc., 2009. 358 p. ISBN 978-0-470-27680-8.

YANG, S.-B.; WU, Y.-G. Genetic algorithm for clustering mixed-type data. **Journal of Electronic Imaging**, v. 20, n. 1, 2011.

YAO, X. *et al.* An improved clustering algorithm and its application in wechat sports users analysis. *In: Procedia Computer Science*. [S.l.: s.n.], 2018. v. 129, p. 166–174.

YAO, X. *et al.* An improved clustering algorithm and its application in iot data analysis. **Computer Networks**, v. 159, p. 63–72, 2019.

YE, M.; LI, X.; ORLOWSKA, M. Projected outlier detection in high-dimensional mixed-attributes data set. **Expert Systems with Applications**, v. 36, n. 3 PART 2, p. 7104–7113, 2009.

YUAN, Z.; ZHANG, X.; FENG, S. Hybrid data-driven outlier detection based on neighborhood information entropy and its developmental measures. **Expert Systems with Applications**, v. 112, p. 243–257, 2018.

ZHANG, K.; GU, X. An affinity propagation clustering algorithm for mixed numeric and categorical datasets. **Mathematical Problems in Engineering**, v. 2014, 2014.

ZHANG, Q.; IP, E. H. Variable assessment in latent class models. **Computational Statistics and Data Analysis**, v. 77, p. 146–156, 2014.

ZHAO, P. *et al.* A multidimensional and multimembership clustering method for social networks and its application in customer relationship management. **Mathematical Problems in Engineering**, v. 2013, 2013.

ZHAO, X.; CAO, F.; LIANG, J. A sequential ensemble clusterings generation algorithm for mixed data. **Applied Mathematics and Computation**, v. 335, p. 264–277, 2018.

A APPENDIX A

In the following sections, one presents the results of the individual analysis, part 1, which brings to the basic description of the document collection.

Individual analysis - Part 1

In this section, one details the individual analysis, providing the information about each document in each branch. Tables from 37 to 42 bring the documents from partitioned to IAPRA branch, respectively. The tables inform the following information in each column: i) the described document; ii) the names of the approaches proposed by the described document; iii) the main algorithms, methods and/or concepts the document based its contributions, the strategy used by the described work to cope with mixed data (1: data transformation and 2: a specific function for each type of attribute), and iv) type of objective the described document announced.

Note that the meaning of the acronyms showed in the mentioned tables are presented in the next section. Besides, in each table, it may be possible that, before the name of the authors (first column), there is a letter inside in two brackets, indicating that document was also classified in another branch. Such a letter represents the first letter of the refereed branch. With respect to the objective columns, one considers the following scenarios: i) the described document proposes a pure method on some theoretical issue about mixed data clustering algorithms or related themes (pure proposal - PP); ii) the described document proposes an improvement based on, inspired by, and/or restricted to a solution of a specific problem (proposal inspired on application - PIA), and iii) the described document proposes a pure application of some existed method (pure application - PA). If the described document had as objective the last scenario, then one also informs the theme of the application/problem (application theme - AT), such as health, education, and finances. Table 13 shows the questions with respect to this block.

Table 37 – Individual analysis, Part 1 - Partitional branch

Article	Proposal's name	Based on	Strategy	Objective		
				PP	PA	AT
Huang (1998)	K-prototypes	K-means	2	x	-	-
Ng and Wong (2002)	Fuzzy K-prototypes	K-means	2	x	-	-
He, Xu and Deng (2005)	dSqueezer and usmSqueezer	Squeezer	Both	x	-	-
Ahmad and Dey (2007)	-	K-means	2	x	-	-
Bushel, Wolfinger and Gibson (2007)	modk-prototypes	K-prototypes	2	-	-	Health
Brouwer (2008)	-	Fuzzy c-means	2	x	-	-
Cheng and Leu (2008)	CPT	K-prototypes	2	-	-	Engineering
Cheng and Leu (2009)	CKP	K-prototypes	2	-	-	Engineering
Lee and Pedrycz (2009)	Fuzzy p-mode prototypes	Fuzzy c-means	2	x	-	-
Lee and Pedrycz (2010)	AFCM	Fuzzy c-means	1	x	-	-
Ahmad and Dey (2011)	-	K-means	2	x	-	-
Cheng and Leu (2011)	KJ-CKP	CKP and KJ method	2	-	-	Engineering
Pham, Suarez-Alvarez and Prostov (2011)	RANKPRO	K-prototypes	2	x	-	-
Irigoien, Sierra and Arenas (2012)	-	PAM and INCA index	2	-	x	Health
Ji <i>et al.</i> (2012)	-	K-prototypes	2	x	-	-
Liang <i>et al.</i> (2012)	-	K-prototypes	2	x	-	-
Cheung and Jia (2013)	OCIL	-	2	x	-	-
[M] Hennig and Liao (2013)	-	K-medoids	1	-	-	Social
Ji <i>et al.</i> (2013)	-	K-prototypes	2	x	-	-
Ovesná <i>et al.</i> (2013)	-	K-means	2	-	x	Agriculture
Han <i>et al.</i> (2014)	MAGE and TSCKA	K-anonymization	2	x	-	-
Pasquier, Naciri and Jeanmonod (2015)	-	PCA and K-means	2	-	x	Agriculture
Harikumar and Surya (2015)	-	K-modoid	2	x	-	-
Ji <i>et al.</i> (2015)	CCIDD	-	2	x	-	-
Sangam and Om (2015)	HDLA	-	2	x	-	-
Wei, Chow and Chan (2015)	UFT	K-means and MI	1	x	-	-
Ahmad and Hashmi (2016)	KMCMD	K-harmonic	2	x	-	-
Amorim and Makarenkov (2016)	MWK-DC	K-means	2	x	-	-
Bishnu and Bhattacharjee (2016)	-	K-mode	2	-	-	Cost estimation
Drakopoulos <i>et al.</i> (2016)	-	K-means	2	x	-	-
Gocsik <i>et al.</i> (2016)	-	-	not specified	-	x	Livestock
Thong and Son (2016)	PFCA-CD	FC-PFS	2	x	-	-

Table 37 Continued on next page

Table 37 continued from previous page

Article	Proposal's name	Based on	Strategy	Objective		
				PP	PA	AT
Huerta-Muñoz, Ríos-Mercado and Ruiz (2017)	MKM	K-means and IG	2	-	-	x
[H] Kampert, Meulman and Friedman (2017)	rCOSA	COSA	2	x	-	-
Kim (2017)	FKPT	K-prototypes	2	x	-	-
Liang, Wang and Lai (2017)	WKStream	K-means	2	x	-	-
Özbakir and Turna (2017)	M-IMO, M-WSA1, and M-WSA2	IMO and WSA	2	x	-	-
Wangchamhan, Chiewchanwattana and Sunat (2017)	C-LCA, SC-LCA, and KSC-LCA	K-means and LCA	2	x	-	-
Bathla, Aggarwal and Rani (2018)	-	K-prototypes and MapReduce	2	x	-	-
Davoodi and Moradi (2018)	DRBFS	Fuzzy K-prototypes	2	-	-	x
Deng <i>et al.</i> (2018)	-	PAM	2	-	x	-
Gu <i>et al.</i> (2018)	-	K-prototypes	2	-	-	x
Hong <i>et al.</i> (2018)	HKPDR	K-prototypes	2	-	-	x
Jia and Cheung (2018)	WOCIL	OCIL	2	x	-	-
Li <i>et al.</i> (2018)	SIC-PCC	-	2	-	-	x
[H] Najjar <i>et al.</i> (2018)	-	-	2	-	-	x
Sangam and Om (2018)	-	K-prototypes	2	-	-	x
Sangam and Om (2018)	Equi-Clustream	HDLA	2	x	-	-
Szepannek (2018)	-	-	-	-	-	x
Yao <i>et al.</i> (2018)	-	K-prototypes	2	-	-	x
Yuan, Zhang and Feng (2018)	NIEOD	Information entropy	2	x	-	-
HajKacem, N'cir and Essoussi (2019)	AMRKP	K-prototypes and MapReduce	2	x	-	-
Dutta, Sil and Dutta (2019)	MOGA-KP	K-prototypes and GA	2	x	-	-
Noia <i>et al.</i> (2019)	-	K-means and GA	2	-	-	x
Ji <i>et al.</i> (2019)	ABC K-prototypes	K-prototypes and ABC	2	x	-	-
Lv <i>et al.</i> (2019)	ODPCA	K-means and K-modes	2	x	-	-
Maione, Nelson and Barbosa (2019)	-	-	2	-	x	-
Shukla and Nanda (2019)	P-SSO	K-prototypes	2	-	-	x
Yao <i>et al.</i> (2019)	-	K-prototypes	2	-	-	x

Table 38 – Individual analysis, Part 1 - Hierarchical branch

Article	Proposal's name	Based on	Strategy	Objective			
				PP	PA	PIA	AT
Philip and Ottaway (1983)	-	-	2	-	x	-	War
Frisvad (1994)	-	-	2	-	x	-	Chemical
Harch <i>et al.</i> (1996)	-	-	2	-	x	-	Agriculture
Sieber, Petrini and Greenacre (1998)	-	-	1	-	x	-	Biology
Friedman and Meulman (2004)	COSA	-	2	x	-	-	-
Shin and Sohn (2004)	-	-	not specified	-	x	-	Marketing
Hsu, Chen and Su (2007)	-	DHC	2	x	-	-	-
Sheu (2007)	-	-	1	-	-	x	Logistics
Herr (2010)	-	-	2	-	-	x	Social
Chitradevi <i>et al.</i> (2012)	DACAD	KNN	2	-	-	x	IoT
[M] Fonseca (2013)	-	-	1	-	x	-	Social
Sopan <i>et al.</i> (2013)	-	ManyNets	2	-	x	-	Software
Brida <i>et al.</i> (2014)	-	ClustOFVar	2	-	x	-	Marketing
Horta and Camanho (2014)	-	BIRCH	2	-	x	-	Industry
[P] Kampert, Meulman and Friedman (2017)	rCOSA	COSA	2	x	-	-	-
[P] Najjar <i>et al.</i> (2018)	-	-	2	-	x	-	Health
Bragato, Ganis and Feoli (2019)	-	-	2	-	x	-	Agriculture

Table 39 – Individual analysis, Part 1 - Model-based branch

Article	Proposal's name	Based on	Strategy	Objective			
				PP	PA	PIA	AT
Wong and Chiu (1987)	-	-	1	x	-	-	-
Harch <i>et al.</i> (1999)	-	FASTCLUST and MULTIMIX	2	-	x	-	Agriculture
Hunt and Basford (1999)	-	MULTIMIX	2	x	-	-	-
Bacher (2000)	-	LCM	2	x	-	-	-
Hunt and Jorgensen (2003)	-	MULTIMIX	2	x	-	-	-
McLachlan and Chang (2004)	-	-	2	-	x	-	Discussion

Table 39 Continued on next page

Table 40 continued from previous page

Article	Proposal's name	Based on	Strategy	Objective			
				PP	PA	PIA	AT
Tai, Hsu and Chen (2010)	GMixSOM	DHC and GenSOM	2	x	-	-	-
Hsu, Lin and Tai (2011)	MixSOM	DHC and GVISOM	2	x	-	-	-
Hsu and Lin (2012)	MixSOM	DHC and SOM	2	x	-	-	-
Liao and Hsu (2012)	SOMCD	DHC and SOM	2	x	-	-	-
Tai and Hsu (2012)	GMixSOM	DHC and SOM	2	x	-	-	-
Noorbehbahani, Mousavi and Mirzaei (2014)	ASOINN	SOM	2	x	-	-	-
Coso <i>et al.</i> (2015)	FMSOM	NCSOM	2	x	-	-	-
Malondkar <i>et al.</i> (2019)	Spark-GHSOM	DHC and GHSOM	2	x	-	-	-

Table 41 – Individual analysis, Part 1 - Others branch

Article	Proposal's name	Based on	Strategy	Objective			
				PP	PA	PIA	AT
Ichino and Sklansky (1985)	IRIG and MNG	Graph-based (RNG and GG)	2	x	-	-	-
Veiga (1996)	ADAGIO	Conceptual-based (ITERATE)	1	-	-	x	Health
Hai and Susumu (2004)	PCACMD	Density-based (K-sets)	2	x	-	-	-
Hsu, Jao and Chen (2005)	Clus-Tree	Graph and Tree based (K-NNG and CTC)	2	x	-	-	-
Puthongsiriporn <i>et al.</i> (2006)	ePKI	Conceptual-based (PKI)	1	x	-	-	-
Hsu and Chen (2007)	CAVE	Entropy-based	2	-	-	x	Marketing
[M] Gondek and Hofmann (2007)	CCIB	Semi-supervised (IBF)	2	x	-	-	-
Ye, Li and Orlowska (2009)	PODM	Entropy-based	2	x	-	-	-
Al-Shaqsi and Wang (2010)	-	Ensemble	2	x	-	-	-
Yang and Wu (2011)	CLUSTERING	Graph-based	2	x	-	-	-
Zhao <i>et al.</i> (2013)	-	Graph-based	2	-	-	x	Marketing
Wang, Machiraju and Huang (2014)	MRCPS	Ensemble	2	-	-	x	Health

Table 41 Continued on next page

Table 41 continued from previous page

Article	Proposal's name	Based on	Strategy	Objective			
				PP	PA	PIA	AT
Zhang and Ip (2014)	-	Graph-based (APC and Ahmad and Dey (2007))	2	x	-	-	-
Niu <i>et al.</i> (2015)	CUCA-MD	Spectral clustering	2	-	-	x	Education
Chen and He (2016)	ACC-FSFDp	Density-based and Semi-supervised	2	x	-	-	-
Bhatnagar <i>et al.</i> (2017)	DASC	Grid-based (ExCC)	2	x	-	-	-
Ding <i>et al.</i> (2017)	DP-MD-FN	Density and entropy-based	2	x	-	-	-
Du, Ding and Xue (2017)	DPC-MD	Density-based (DPC)	2	x	-	-	-
Hummel, Edelmam and Kopp-Schneider (2017)	CluMix	Co-clustering	2	-	-	x	Health
Iam-On and Boongoen (2017)	WCT-T	Ensemble and	2	-	-	x	Education
Jinyin <i>et al.</i> (2017)	and WTQ-T FDCA	Graph-based (LCE)	2	x	-	-	-
Liu, Yang and He (2017)	EPDCA	Density-based	2	x	-	-	-
Liu <i>et al.</i> (2017)	DPC-M	Density and entropy-based (DBSCAN)	2	x	-	-	-
Noorbehbahani <i>et al.</i> (2017)	ISF-NIDS	Density-based (DPC)	2	-	-	x	Security
Valenzuela <i>et al.</i> (2017)	-	Semi-supervised	2	-	-	x	Health
Jang <i>et al.</i> (2018)	GK-prototype	Graph-based (LCDA and MST)	2	-	-	x	Health
Lin, Azarnoush and Runger (2018)	CRAFTER	Grid-based (K-prototype)	2	-	-	x	Environment
Zhao, Cao and Liang (2018)	SECG	Ensemble (RFC)	2	x	-	-	-
Duan <i>et al.</i> (2019)	DPC-SDAE	Entropy and ensemble-based	2	x	-	-	-
		Density-based (DPC)	1	x	-	-	-

Table 42 – Individual analysis, Part 1 - IAPRA branch

Article	Proposal's name	Based on	Strategy	Objective			
				PP	PA	PIA	AT
Gordon (1987)	-	-	-	-	x	-	Environment
Cox and Cox (2000)	-	Gower distance	2	x	-	-	-
Shi and Horvath (2006)	-	RFD	2	-	x	-	Theoretical
Irigoin and Arenas (2008)	INCA	-	2	x	-	-	-
Barceló-Rico, Díez and Bondia (2011)	-	Polar coordinate	1	-	-	x	Health
Barceló-Rico and Díez (2012)	-	Polar and spherical coordinate	1	x	-	-	-
Suarez-Alvarez <i>et al.</i> (2012)	-	-	2	x	-	-	-
Bernard <i>et al.</i> (2014)	-	-	1	x	-	-	-
Hennig and Lin (2015)	-	-	-	x	-	-	-
Lam, Wei and Wunsch (2015)	UFLA	FART	1	x	-	-	-
Prostov, Suarez-Alvarez and Prostov (2015)	-	Specific article ¹	2	x	-	-	-
Hsu and Huang (2016)	DRDH	DHC and t-SNE	2	x	-	-	-
Rahman and Islam (2016)	FEMI	-	2	x	-	-	-
Taha and Hadi (2016)	-	-	1	x	-	-	-
Kuppusamy and Paramasivam (2017)	GFNN	ANFIS and GWO	2	x	-	-	-
Audigier <i>et al.</i> (2018)	-	-	-	-	x	-	Theoretical
Nowosad and Stepinski (2018)	SABRE	V-measure	2	-	-	x	Geography
Ahmad and Khan (2019)	-	-	-	-	x	-	Theoretical
Foss, Markatou and Ray (2019)	-	-	-	-	x	-	Theoretical
Golinko and Zhu (2019)	GEL	GFEL	1	x	-	-	-
Velden, Enza and Markos (2019)	-	-	-	-	x	-	Theoretical

¹ Suarez-Alvarez *et al.* (2012)

Meaning of the acronyms

In the following sections, one presents the meaning of each acronym that was mentioned in the previous tables of this appendix.

Acronyms - Partitional branch

- CPT: Contract packaging technique
- CKP: Constrained k-prototypes algorithm
- AFCM: Adaptive fuzzy c-means
- RANKPRO: Random search with k-prototypes algorithm
- PAM: Partitioning around medoids algorithm
- INCA: Index number clusters atypical
- OCIL: Object-cluster similarity metric
- MAGE: Microaggregation and Generalization
- TSCKA: Two Steps Clustering K-anonymization Algorithm
- PCA: Principal component analysis
- CCIDD: Cluster centers initialization based on density and distance
- HDLA: Hybrid data labeling algorithm
- UFT: Unsupervised feature transformation
- MI: Mutual information
- KMCMD: K-means type clustering algorithm for mixed datasets
- MWK-DC: Minkowski weighted k-means with distributed centroids
- PFCA-CD: Picture fuzzy clustering algorithm for complex data
- FC-PFS: Fuzzy clustering algorithm on picture fuzzy set
- MKM: Modified k-means
- IG: Iterated greedy
- COSA: Clustering of objects on subsets of attributes
- rCOSA: R package of the COSA approach

- FKPT: Fast k-prototypes
- M-IMO: Modified ions motion optimization
- M-WSA1: Modified weighted superposition attraction 1
- M-WSA2: Modified weighted superposition attraction 2
- IMO: Ions motion optimization
- WSA: Weighted superposition attraction
- C-LCA: Chaotic league championship algorithm
- SC-LCA: Search clustering using chaotic league championship algorithm
- KSC-LCA: Hybrid k-means and chaotic league championship algorithm
- LCA: League championship algorithm
- DRBFS: Deep rule-based fuzzy system
- HKPDR: Hierarchical clustering and improved k-prototypes method based on average density and rough set theory
- SIC-PCC: Penalized competitive clustering algorithm based on the standardized initial clustering centers
- NIEOD: Neighborhood information entropy-based outlier detection algorithm
- AMRKP: Accelerated MapReduce-based k-prototypes algorithm
- MOGA-KP: Multi-objective genetic algorithm for k-prototypes
- GA: Genetic algorithm
- ABC: artificial bee colony
- ODPCA: Optimizing and differentially private clustering algorithm
- P-SSO: Parallel social spider optimization algorithm
- IoT: Internet of things

Acronyms - Hierarchical branch

- DHC: Distance hierarchy concept
- DACAD: Distributed agglomerative cluster-based anomaly detection algorithm
- KNN: K-nearest neighborhood
- BIRCH: Balanced iterative reducing and clustering using hierarchies
- COSA: Clustering of objects on subsets of attributes
- rCOSA: R package of the COSA approach

Acronyms - Model-based branch

- LCM: Latent class model
- CCIB: Coordinated conditional information bottleneck
- IBF: Information bottleneck framework
- KL-FCM-GM: Kullback-Leibler fuzzy c-means algorithm for Gauss-multinomial-distributed data
- FCM: Fuzzy c-means algorithm
- G-G: Gath-geva algorithm
- ICL: Integrated completed likelihood
- BPF: Bayesian predictive framework
- GLMCWM: Generalized linear mixed cluster-weighted model
- CWM: Cluster-weighted models
- KAMILA: Kaymeans form mixed large data
- FMM: Finite mixture models
- MFA-MD: Mixture of factor analyser for mixed data
- BFMM: Bayesian finite mixture model
- EDV: Ext-DVine
- EVMVC: Ext-Vine-MVC
- VC: Vine copula
- BSM: Bayesian statistic models
- PGMM: Poisson Gaussian multinomial mixture model

Acronyms - Neural network-based branch

- EAOI: Extended attribute-orientated induction
- SOM: Self-organized maps
- GSOM: Generalized self-organized maps
- AOI: Attribute-orientated induction
- DHC: Distance hierarchy concept
- M-ART: Modified adaptive resonance theory network
- ART1: Adaptive resonance theory network algorithm type I
- ART2: Adaptive resonance theory network algorithm type II
- GMixSOM: Growing mixed-type self-organized maps
- GViSOM: Generalized visualization-induced self-organizing map
- ASOINN: Adjusted self-organizing incremental neural network
- FMSOM: Frequency neuron mixed self-organized map
- GHSOM: Growing hierarchical self-organizing map algorithm

Acronyms - Other branches

- IRIG: Interclass rectangular influence graph
- MNG: Mutual neighborhood graph
- RNG: Relative neighborhood graph
- GG: Gabriel graph
- PCACMD: Parallel clustering algorithm for categorical and mixed data
- K-NNG: K-nearest neighborhood graph
- CTC: Clustering tree concept
- ePKI: Evolved pattern-based knowledge induction
- PKI: Pattern-based knowledge induction
- CCIB: Coordinated conditional information bottleneck
- IBF: Information bottleneck framework

- PODM: High dimensional mixed-attributes datasets
- MRCPS: Molecular regularized consensus patient stratification
- APC: Affinity propagation clustering algorithm
- ACC-FSFDP: Self-adaptive peak density clustering algorithm for data with mixed attributes
- DASC: Data aware scalable clustering
- DP-MD-FN: Entropy-based density peaks clustering algorithm for mixed data employing fuzzy neighborhood
- DPC-MD: Density peaks clustering for mixed-type data
- DPC: Density peaks clustering
- LCE: Link-based cluster ensembles
- FDCA: Fast density clustering algorithm
- DBSCAN: Density-based spatial clustering of applications with noise
- DPC-MD: Density peaks clustering for mixed-type data
- ISF-NIDS: incremental semi-supervised flow network-based intrusion detection system
- LCDA: Louvain community detection algorithm
- MST: Minimum spanning tree
- GK-prototype: Grid-based k-prototypes
- CRAFTER: Clustering algorithm using random forest based iterations
- RFC: Random forest clustering
- SECG: Sequential ensemble clustering generation algorithm for mixed data

Acronyms - IAPRA branch

- RFD: Random forest dissimilarities
- INCA: Index number clusters atypical
- UFLA: unsupervised feature learning with fuzzy adaptive resonance theory network
- FART: Fuzzy adaptive resonance theory

- DRDH: Dimensionality reduction with distance hierarchy
- DHC: Distance hierarchy concept
- t-SNE: t-distributed stochastic neighbor embedding
- FEMI: Fuzzy expectation maximization and fuzzy clustering-based missing value imputation framework for data pre-processing
- GFNN: Grey fuzzy neural network
- ANFIS: Adaptative neuro-fuzzy inference system
- GWO: Grey wolf optimizer
- SABRE: Spatial association between regionalizations
- GEL: Generalized feature embedding approach
- GFEL: Generalized feature embedding learning

B APPENDIX B

In the following sections, one presents the results of the individual analysis, part 2, which is respect to the standard representation string.

Individual analysis - Part 2

In this section, one provides the filled standard representation string for each document returned in this review, separated in one table per branch. Tables from 43 to 48 show the following branches, in this order, partitional, hierarchical, model-based, neural network-based, others, and IAPRA. Note that the meaning of the acronyms showed in the mentioned tables are presented in the next section. Besides, in each table, it may be possible that, before the name of the authors (first column), there is a letter inside in two brackets, indicating that document was also classified in another branch. Such a letter represents the first letter of the refereed branch.

Table 43 – Individual analysis, Part 2 - Partitional branch

Article	Standard representation			
	DTS	PRE	PRO	POS
Huang (1998)	-	-	DS-FNC	-
Ng and Wong (2002)	-	-	CN-DFN + FZZ	OPT
He, Xu and Deng (2005)	-	DT-TRS	DS-FNC	-
Ahmad and Dey (2007)	-	-	CN-DFN + DS-FNC	-
Bushel, Wolfinger and Gibson (2007)	-	-	DS-FNC + NB-GRP	-
Brouwer (2008)	-	-	FZZ	-
Cheng and Leu (2008)	-	-	CNS	-
Cheng and Leu (2009)	-	-	CNS	-
Lee and Pedrycz (2009)	-	-	CN-DFN + FZZ	-
Lee and Pedrycz (2010)	-	-	DS-FNC	-
Ahmad and Dey (2011)	-	-	AT-WGT + SBS	-
Cheng and Leu (2011)	-	-	CNS	IT-CLS
Pham, Suarez-Alvarez and Prostov (2011)	-	-	CN-DFN	OPT
Irigoin, Sierra and Arenas (2012)	-	-	-	PR-APP
Ji <i>et al.</i> (2012)	-	-	DS-FNC + NB-GRP + VR-IMP	-
Liang <i>et al.</i> (2012)	-	-	DS-FNC + NB-GRP	VL-MSR
Cheung and Jia (2013)	-	-	DS-FNC + NB-GRP + VR-IMP	CM-LRN
[M] Hennig and Liao (2013)	-	DT-TRS + FT-SCL	DS-FNC	IT-CLS
Ji <i>et al.</i> (2013)	-	-	AT-WGT + DS-FNC	-
Ovesná <i>et al.</i> (2013)	-	-	-	PR-APP
Han <i>et al.</i> (2014)	-	-	DS-FNC	-
Pasquier, Naciri and Jeanmonod (2015)	-	-	-	PR-APP
Harikumar and Surya (2015)	-	OT-DTC	DS-FNC	-
Ji <i>et al.</i> (2015)	-	-	CN-DFN	-
Sangam and Om (2015)	DYN + LRG	-	CL-UPD + DS-FNC	-
Wei, Chow and Chan (2015)	-	DT-TRS	-	-
Ahmad and Hashmi (2016)	-	-	CN-DFN + DS-FNC	-
Amorim and Makarenkov (2016)	-	-	AT-WGT + SBS	-
Bishnu and Bhattacharjee (2016)	-	-	DS-FNC	-
Drakopoulos <i>et al.</i> (2016)	-	-	CN-DFN + NB-GRP	VL-MSR

Table 43 Continued on next page

Table 43 continued from previous page

Article	Standard representation				
	DTS	PRE	PRO	POS	AUX
Gocsik <i>et al.</i> (2016)	-	-	-	-	PR-APP
Thong and Son (2016)	-	-	CN-DFN + DS-FNC + FZZ	-	OPT
Huerta-Muñoz, Ríos-Mercado and Ruiz (2017)	-	-	CN-DFN + DS-FNC	-	OPT
[H] Kampert, Meulman and Friedman (2017)	HDM	-	AT-WGT + CNS + SBS	VSL	-
Kim (2017)	-	-	DS-FNC	-	-
Liang, Wang and Lai (2017)	DYN	-	AT-WGT + DS-FNC	-	-
Özbakır and Turna (2017)	-	-	DS-FNC + NB-GRP	-	OPT
Wangchamhan, Chiewchanwattana and Sumat (2017)	-	-	CN-DFN	-	OPT
Bathia, Aggarwal and Rani (2018)	LRG	-	-	-	PRL
Davoodi and Moradi (2018)	LRG	-	DS-FNC + FZZ	-	-
Deng <i>et al.</i> (2018)	-	-	-	-	PR-APP
Gu <i>et al.</i> (2018)	-	-	CN-DFN + CNS + NB-GRP	-	OPT
Hong <i>et al.</i> (2018)	-	-	CN-DFN + NB-GRP + RG-STs	-	-
Jia and Cheung (2018)	-	-	AT-WGT + CN-DFN + NB-GRP + SBS	-	-
Li <i>et al.</i> (2018)	-	-	CN-DFN + NB-GRP	-	-
[H] Najjar <i>et al.</i> (2018)	DYN + LRG	-	-	-	FRM
Sangam and Om (2018)	-	-	DS-FNC	-	-
Sangam and Om (2018)	DYN	-	DS-FNC	VSL	-
Szepannek (2018)	-	-	AT-WGT	-	CD-AVL
Yao <i>et al.</i> (2018)	-	-	CN-DFN + NB-GRP	-	-
Yuan, Zhang and Feng (2018)	-	OT-DTC	-	-	-
HajKacem, N'cir and Essoussi (2019)	LRG	-	-	-	PRL
Dutta, Sil and Dutta (2019)	-	-	NB-GRP	-	OPT
Noia <i>et al.</i> (2019)	-	-	NB-GRP	-	OPT
Ji <i>et al.</i> (2019)	-	-	CN-DFN	-	OPT
Lv <i>et al.</i> (2019)	-	-	PR-PRT	-	-
Matone, Nelson and Barbosa (2019)	-	-	-	-	PR-APP
Shukla and Nanda (2019)	-	-	CN-DFN	-	PRL + OPT
Yao <i>et al.</i> (2019)	-	-	CN-DFN + NB-GRP	-	-

Table 44 – Individual analysis, Part 2 - Hierarchical branch

Article	Standard representation			
	DTS	PRE	PRO	POS
Philip and Ottaway (1983)	-	-	-	PR-APP
Frisvad (1994)	-	-	-	PR-APP
Harch <i>et al.</i> (1996)	-	-	-	PR-APP
Sieber, Petrini and Greenacre (1998)	-	-	-	FRM
Friedman and Meulman (2004)	-	MS-DTA	AT-WGT + SBS	IT-CLS
Shin and Sohn (2004)	LRG	-	-	PR-APP
Hsu, Chen and Su (2007)	-	-	DS-FNC	-
Sheu (2007)	-	-	FZZ	FRM + OPT
Herr (2010)	DYN	-	NB-GRP	VL-IDX
Chitradevi <i>et al.</i> (2012)	LRG	OT-DTC	DS-FNC	-
[M] Fonseca (2013)	-	-	-	CMP
Sopan <i>et al.</i> (2013)	-	-	-	PR-APP
Brida <i>et al.</i> (2014)	-	-	-	PR-APP
Horta and Camanho (2014)	-	-	-	PR-APP
[P] Kampert, Meulman and Friedman (2017)	HDM	-	AT-WGT + CNS + SBS	VSL
[P] Najjar <i>et al.</i> (2018)	DST + LRG	-	-	FRM
Bragato, Ganis and Feoli (2019)	-	-	-	CMP + PR-APP

Table 45 – Individual analysis, Part 2 - Model-based branch

Article	Standard representation			
	DTS	PRE	PRO	POS
Wong and Chiu (1987)	-	DT-TRS + MS-DTA	CL-UPD	-
Harch <i>et al.</i> (1999)	-	-	-	PR-APP
Hunt and Basford (1999)	TMW	-	ST-MDL	-
Bacher (2000)	-	-	NB-GRP + ST-MDL	-
Hunt and Jorgensen (2003)	-	MS-DTA	ST-MDL	-
McLachlan and Chang (2004)	-	-	-	CMP + DSC
[O] Gondek and Hofmann (2007)	-	-	NB-GRP + ST-MDL	-
Table 45 Continued on next page				

Table 45 continued from previous page

Article	Standard representation			
	DTS	PRE	PRO	AUX
Chatzis (2011)	-	-	ST-MDL	-
Hunt and Jorgensen (2011)	-	-	-	CMP + DSC
[H] Fonseca (2013)	-	-	-	PR-APP
[P] Hennig and Liao (2013)	-	DT-TRS	ST-MDL	FRM
Zhang and Ip (2014)	-	-	ST-MDL + VR-IMP	-
Baudry <i>et al.</i> (2015)	-	-	NB-GRP	VL-MSR
Blomstedt <i>et al.</i> (2015)	-	-	ST-MDL	OPT
Ingrassia <i>et al.</i> (2015)	-	-	ST-MDL	-
Foss <i>et al.</i> (2016)	LRG	-	ST-MDL	-
Hunt and Basford (2016)	-	-	-	CMP + DSC
McParland and Gormley (2016)	-	-	ST-MDL	-
McParland <i>et al.</i> (2017)	HDM	DM-RDC	ST-MDL	VL-MSR
Tekumalla, Rajan and Bhattacharyya (2017)	-	-	ST-MDL	-
Abidin and Westhead (2017)	-	-	ST-MDL	-
Foss and Markatou (2018)	LRG	-	ST-MDL	CD-AVL + PRL
Guglielmi <i>et al.</i> (2018)	-	-	ST-MDL	-
Storlie <i>et al.</i> (2018)	-	DM-RDC + MS-DTA	ST-MDL	-
Wang, Zhu and Dong (2018)	-	-	ST-MDL	-
Biernacki and Lourme (2019)	-	-	ST-MDL	FRM
Marbac and Sedki (2019)	-	DM-RDC + MS-DTA	ST-MDL	CD-AVL
Marbac, Sedki and Patin (2019)	LRG	DM-RDC + MS-DTA	ST-MDL	PRL
Marbac and Vandewalle (2019)	LRG	DM-RDC	VRB-CRR	FRM

Table 46 – Individual analysis, Part 2 - Neural network-based branch

Article	Standard representation			
	DTS	PRE	PRO	AUX
Hsu and Wang (2006)	HDM	-	DS-FNC	IT-CLS + VSL
Hsu and Huang (2008)	DYN + LRG	-	CL-UPD + DS-FNC	-
Tai, Hsu and Chen (2010)	DYN + HDM	-	CL-UPD + DS-FNC	VSL
Hsu, Lin and Tai (2011)	HDM	-	DS-FNC	VSL

Table 46 Continued on next page

Table 46 continued from previous page

Article	Standard representation			
	DTS	PRE	PRO	AUX
Hsu and Lin (2012)	HDM	-	DS-FNC	VSL
Liao and Hsu (2012)	HDM	-	DS-FNC	VSL
Tai and Hsu (2012)	DYN + HDM	-	CL-UPD + DS-FNC	VSL
Noorbehbahani, Mousavi and Mirzaei (2014)	DYN	NS-RDT	DS-FNC + NB-GRP + RL-UPD	-
Coso <i>et al.</i> (2015)	-	-	DS-FNC	-
Malondkar <i>et al.</i> (2019)	DST + LRG	-	DS-FNC	VSL
				PRL

Table 47 – Individual analysis, Part 2 - Others branch

Article	Standard representation			
	DTS	PRE	PRO	AUX
Ichino and Sklansky (1985)	-	-	DS-FNC + GRP-MDL	-
Veiga (1996)	-	MS-DTA	DS-FNC	-
Hai and Susumu (2004)	LRG	-	-	PRL
Hsu, Jao and Chen (2005)	-	-	DS-FNC + TRE-RLS	-
Puthongsiriporn <i>et al.</i> (2006)	-	-	CLS-CRR	-
Hsu and Chen (2007)	-	-	DS-FNC	VL-MSR
[M] Gondek and Hofmann (2007)	-	-	ST-MDL	-
Ye, Li and Orlowska (2009)	HDM	OT-DTC	DS-FNC + SBS	-
Al-Shaqsi and Wang (2010)	-	-	ENS-PRD	-
Yang and Wu (2011)	LRG	-	AT-WGT + NB-GRP	OPT
Zhao <i>et al.</i> (2013)	HDM	-	FZZ + GRP-MDL	-
Wang, Machiraju and Huang (2014)	DST	-	DS-FNC + ENS-PRD	-
Zhang and Ip (2014)	-	-	DS-FNC	-
Niu <i>et al.</i> (2015)	-	-	DS-FNC	-
Chen and He (2016)	DYN	OT-DTC	CN-DFN + DS-FNC	-
Bhatnagar <i>et al.</i> (2017)	DST + DYN + LRG	-	CL-UPD	FRM + PRL
Ding <i>et al.</i> (2017)	-	-	CN-DFN + DS-FNC + FZZ	-
Du, Ding and Xue (2017)	-	-	DS-FNC	-
Hummel, Edelmann and Kopp-Schneider (2017)	-	-	DS-FNC	VSL
Iam-On and Boongoen (2017)	-	-	DS-FNC	FRM

Table 47 Continued on next page

Table 47 continued from previous page

Article	Standard representation			
	DTS	PRE	PRO	AUX
Jinyin <i>et al.</i> (2017)	-	-	CN-DFN + DS-FNC	-
Liu, Yang and He (2017)	DYN + HDM	-	DS-FNC + CL-UPD + NB-GRP	PRL
Liu <i>et al.</i> (2017)	-	-	DS-FNC	-
Noorbehbahani <i>et al.</i> (2017)	DYN	-	CL-UPD + DS-FNC + NB-GRP	-
Valenzuela <i>et al.</i> (2017)	-	-	DS-FNC + GRP-MDL	-
Jang <i>et al.</i> (2018)	LRG	-	DS-FNC	-
Lin, Azarnoush and Runger (2018)	HDM	-	ENS-PRD	FRM
Zhao, Cao and Liang (2018)	-	-	CLS-CRR + ENS-PRD	-
Duan <i>et al.</i> (2019)	-	-	DS-FNC + NB-GRP	-

Table 48 – Individual analysis, Part 2 - IAPRA branch

Article	Standard representation			
	DTS	PRE	PRO	AUX
Gordon (1987)	-	-	-	DSC
Cox and Cox (2000)	-	-	DS-FNC	-
Shi and Horvath (2006)	-	-	-	PRL
Irigoien and Arenas (2008)	-	OT-DTC	NB-GRP	VL-MSR
Barceló-Rico, Díez and Bondia (2011)	-	DT-TRS	-	-
Barceló-Rico and Díez (2012)	-	DT-TRS	-	-
Suarez-Alvarez <i>et al.</i> (2012)	-	FT-SCL	-	-
Bernard <i>et al.</i> (2014)	LRG	-	-	VSL
Hennig and Lin (2015)	-	-	NB-GRP	VL-MSR
Lam, Wei and Wünsch (2015)	LRG + SPR	DT-TRS + MS-DTA	-	FRM
Prostov, Suarez-Alvarez and Prostov (2015)	-	FT-SCL	-	MTH-PRV
Hsu and Huang (2016)	HDM	DM-RDC	DS-FNC	VSL
Rahman and Islam (2016)	-	MS-DTA	FZZ	-
Taha and Hadi (2016)	-	-	VRB-CRR	-
Kuppusamy and Paramasivam (2017)	-	MS-DTA	CNS + FZZ	OPT
Audigier <i>et al.</i> (2018)	-	-	-	CMP + DSC
Nowosad and Stepinski (2018)	-	-	NB-GRP	VL-MSR

Table 48 Continued on next page

Table 48 continued from previous page

Article	Standard representation			
	DTS	PRE	PRO	POS
Ahmad and Khan (2019)	-	-	-	-
Foss, Markatou and Ray (2019)	-	-	-	-
Golinko and Zhu (2019)	DYN	DM-RDC	CL-UPD	-
Velden, Enza and Markos (2019)	-	-	-	-
				DSC + RVW
				DSC + RVW
				-
				DSC + RVW

Meaning of the acronyms

In this section, one presents the meaning of each acronym that was mentioned in the previous tables of this appendix in three Tables (from 49 to 51). where the first brings the meaning of the acronyms for the characteristics of the datasets, and the preprocessing tasks of KDD process, the second presents the meaning of the acronyms for processing tasks, and the last one shows the meaning of the acronyms for the post-processing tasks and auxiliary approaches.

Table 49 – Acronyms - Datasets' characteristics and preprocessing tasks

DTS - Dataset		PRE - Preprocessing tasks	
Acronym	Meaning	Acronym	Meaning
DST	Distributed	DT-TRS	Data transformation
DYN	Dynamic	DM-RDC	Dimensionality reduction
HDM	High-dimensional	FT-SCL	Feature scaling
LRG	Large	MS-DTA	Missing data
TMW	Three-mode and three-way	NS-RDT	Noise reduction
SPR	Sparse	OT-DTC	Outlier detection

Table 50 – Acronyms - Processing tasks

PRO - Processing tasks			
Branches		Activities	
Acronym	Name	Acronym	Meaning
H	Hierarchical	AT-WGT	Attributes weighting
I	IAPRA	CN-DFN	Centroids definition
M	Model-based	CLS-CRR	Clustering correlation
N	Neural network based	CL-INT	Clustering initialization
O	Others	CL-UPD	Clustering update
P	Partitional	CM-LRN	Competitive learning
		CNS	Constrained
		DS-FNC	Dissimilarity function
		ENS-PRD	Ensemble procedure
		FZZ	Fuzzy
		GRP-MDL	Graph model
		NB-GRP	Number of groups
		PR-PRT	Privacy protection
		RG-STs	Rough sets
		RL-UPD	Rules update
		ST-MDL	Statistical model
		SBS	Subspace
		TRE-RLS	Tree rules
		VR-IMP	Variable impact
		VRB-CRR	Variables correlation

Table 51 – Acronyms - Post-processing tasks and auxiliary approaches

POS - Post-processing tasks		AUX - Auxiliary approaches	
Acronym	Meaning	Acronym	Meaning
IT-CLS	Interpretation of the clusters	CD-AVL	Code available
VL-MSR	Validation measure	CMP	Comparison
VSL	Visualization	CMP-LRN	Competitive learning
		DSC	Discussion
		FRM	Framework
		MTH-PRV	Mathematical proof
		OPT	Optimization
		PRL	Parallelism
		PR-APP	Pure application
		RVW	Review
		SPC-KNW	Specialist knowledge

C APPENDIX C

In the following sections, one presents the results of the individual analysis, part 3, which brings to the datasets, performance measures, programming languages, and code availability.

Individual analysis - Part 3

In this section, one presents the information regarding the datasets, validation measures, programming languages and code availability from documents resulted in the systematic review, which are separated in the branches defined in this thesis. Concerning the validation measures shown in the following tables, one presents them into the three classical classes (internal, external, and relative) taking into consideration the classification the authors gave with this respect in their documents or the way they used these measures in their experiments. With respect to the programming language and the code availability, one did not register the name of commercial software (e.g., SPSS and SAS) that has been used just to run an already implemented algorithm on it. It happened in some documents in the hierarchical branch that used such a software to carry out a clustering algorithm to solve a specific problem. The information about programming language and the code availability is related to only those article that implements their proposed algorithm, and/or make the code of the proposed algorithms available.

Therefore, in the following, the 59 documents belonging to the partitionial branch are described in Tables 52 to 54, which show information about the datasets the authors used, the performance measures that were chosen, the programming language used, and code availability, respectively. The remaining tables bring the same information for the other branches, as follows. Tables 55 to 57 show the 17 documents from the hierarchical branch, Tables 58 to 60 show the 29 documents from the model-based branch, Tables 61 to 63 show the 10 documents from the neural network-based branch, Tables 64 to 66 show the 29 documents that addressed others approaches, and Tables 67 to 69 show the 21 documents from IAPRA branch.

Table 52 – Individual analysis, Part 3 – Partitional branch, Datasets

Document	Simulated		Real-world	
	Is it used?	Address	Is it used?	Acronym
Huang (1998)	no	-	yes	CRD-APP
Ng and Wong (2002)	no	-	yes	CRD-APP
He, Xu and Deng (2005)	no	-	yes	CRD-APP, HRD-CLV and STT-GRC
Ahmad and Dey (2007)	no	-	yes	HRD-CLV and STT-ACA
Bushel, Wolfinger and Gibson (2007)	yes	ASM ¹	yes	ACT-DTS and HRD-CLV
Brouwer (2008)	yes	not mentioned	yes	STT-ACA
Cheng and Leu (2008)	yes	not mentioned	yes	CRD-APP
Cheng and Leu (2009)	yes	not mentioned	yes	CRD-APP
Lee and Pedrycz (2009)	no	-	yes	CRD-APP and ZOO-DTS
Lee and Pedrycz (2010)	yes	not mentioned	yes	CNM-CHC, CRD-APP and STT-HRD
Ahmad and Dey (2011)	yes	not mentioned	yes	STT-ACA
Cheng and Leu (2011)	no	-	yes	PCC-TWN
Pham, Suarez-Alvarez and Prostov (2011)	no	-	yes	ADL-DTS and CVR-DTS
Irigoien, Sierra and Arenas (2012)	yes	R package (ICGE) ²	yes	CHW-DTS, GLB-DTS and LMP-CNC
Ji <i>et al.</i> (2012)	no	-	yes	ACT-INF, CRD-APP, HRD-DTS and ZOO-DTS
Liang <i>et al.</i> (2012)	yes	not mentioned	yes	ADL-DTS, CNM-CHC, CRD-APP, HRD-CLV and TCH-AEV
Cheung and Jia (2013)	no	-	yes	ADL-DTS, CRD-APP, DRM-DTS, HRD-DTS, STT-GRC and STT-HRD
[M] Hennig and Liao (2013)	no	-	yes	USC-FNC
Ji <i>et al.</i> (2013)	no	-	yes	CRD-APP and HRD-DTS
Ovesná <i>et al.</i> (2013)	no	-	yes	AGR-DTS
Han <i>et al.</i> (2014)	no	-	yes	ADL-DTS
Pasquier, Naciri and Jeanmonod (2015)	no	-	yes	PNT-DTS
Harikumar and Surya (2015)	no	-	yes	BNK-MRK, STT-ACA and ZOO-DTS
Ji <i>et al.</i> (2015)	no	-	yes	CRD-APP, HRD-DTS and ZOO-DTS

Table 52 Continued on next page

¹ <https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-1-15>
² <https://cran.r-project.org/web/packages/ICGE/index.html>

Table 52 continued from previous page

Document	Simulated		Real-world	
	Is it used?	Address	Is it used?	Acronym
Sangam and Om (2015)	yes	not mentioned	yes	ABL-DTS, ADL-DTS, CRD-APP, FAR-DTS, KDD-C99 and PKR-HND
Wei, Chow and Chan (2015)	yes	not mentioned	yes	DRM-DTS, HRD-CLV, STT-ACA, STT-GRC STT-HRD and ZOO-DTS
Ahmad and Hashmi (2016)	no	-	yes	CRD-APP, HRD-CLV and STT-HRD
Amorim and Makarenkov (2016)	no	-	yes	HRD-CLV, STT-ACA and TCH-AEV
Bishnu and Bhattacharjee (2016)	no	-	yes	CCM-NS1 and CCM-NS2
Drakopoulos <i>et al.</i> (2016)	no	-	yes	BMR-DTS
Gocsik <i>et al.</i> (2016)	no	-	yes	SBP-FRM
Thong and Son (2016)	no	-	yes	ABL-DTS, ATM-DTS, SRV-DTS and STT-GRC
Huerta-Muñoz, Ríos-Mercado and Ruiz (2017)	yes	not mentioned	no	-
[H] Kampert, Meulman and Friedman (2017)	yes	R package (rCOSA) ³	yes	AP3-DTS
Kim (2017)	yes	not mentioned	no	-
Liang, Wang and Lai (2017)	no	-	yes	DTO-RCR and KDD-C99
Özbakir and Turna (2017)	no	-	yes	CRD-APP and STT-HRD
Wangchamhan, Chiewchanwattana and Sumat (2017)	no	-	yes	CNM-CHC, LVR-DSR, STT-ACA and STT-HRD
Bathla, Aggarwal and Rani (2018)	no	-	yes	CHS-KRK
Davoodi and Moradi (2018)	no	-	yes	HRD-DTS, LVR-DSR and MIM-III
Deng <i>et al.</i> (2018)	no	-	yes	CMD-SRV
Gu <i>et al.</i> (2018)	no	-	yes	PCN-STR
Hong <i>et al.</i> (2018)	no	-	yes	ADL-DTS, CRD-APP, DRM-DTS, HRD-DTS and STT-GRC
Jia and Cheung (2018)	no	-	yes	EXE-WRH and STT-HRD
Li <i>et al.</i> (2018)	no	-	yes	ADM-HLT
[H] Najjar <i>et al.</i> (2018)	no	-	yes	ADL-DTS, CNM-CHC and CRD-APP
Sangam and Om (2018)	yes	not mentioned	yes	KDD-C99
Sangam and Om (2018)	yes	-	yes	-
Szepannek (2018)	yes	R package (clustMixType) ⁴	no	-
Yao <i>et al.</i> (2018)	no	-	yes	CNM-CHC, STT-HRD, TCH-AEV and WCH-SPR
Yuan, Zhang and Feng (2018)	no	-	yes	ANN-DTS

Table 52 Continued on next page

³ <https://github.com/mkampert/rCOSA/tree/master/R>⁴ <https://cran.r-project.org/web/packages/clustMixType/index.html>

Table 52 continued from previous page

Document	Simulated		Real-world	
	Is it used?	Address	Is it used?	Acronym
HajKacem, N'cir and Essoussi (2019)	yes	not mentioned	yes	KDD-C99 and PKR-HND
Dutta, Sil and Dutta (2019)	no	-	yes	ACT-INF, ATM-DTS, CNM-CHC, CRD-APP, DRM-DTS, LBR-RLT, LVR-DSR, SRV-DTS, STT-HRD and TCH-AEV
Noia <i>et al.</i> (2019)	no	-	yes	LMB-DTS and NTN-DTS
Ji <i>et al.</i> (2019)	no	-	yes	CRD-APP, HRD-DTS and ZOO-DTS
Lv <i>et al.</i> (2019)	no	-	yes	ADL-DTS and HRD-DTS
Maione, Nelson and Barbosa (2019)	no	-	yes	CER-HSH
Shukla and Nanda (2019)	no	-	yes	CRD-APP, JPC-SCR, STT-ACA and STT-GRC
Yao <i>et al.</i> (2019)	no	-	yes	CNM-CHC, STT-HRD, TCH-AEV and WCH-SPR

Table 53 – Individual analysis, Part 3 - Partitional branch, Performance measures

Document	Internal	External	Relative
Huang (1998)	EXC-TME	ACC-IND	-
Ng and Wong (2002)	-	ACC-IND	-
He, Xu and Deng (2005)	EXC-TME	CLS-ERR	-
Ahmad and Dey (2007)	-	MIP-IND	-
Bushel, Wolfinger and Gibson (2007)	-	AJR-IND	DVI-CTU
Brouwer (2008)	-	AJR-IND, JCC-IND, PRT-IND and RND-IND	RMS-ERR
Cheng and Leu (2008)	EXC-TME and OBJ-FNC	ENT-VLD	-
Cheng and Leu (2009)	EXC-TME and OBJ-FNC	ENT-VLD	-
Lee and Pedrycz (2009)	-	ACC-IND	CRS-CLS, MDP-CFF, PRB-IND, TNG-IND and XIB-IND
Lee and Pedrycz (2010)	CHI-IND, CRS-CLS, MDP-CFF, NMB-ITR, PRB-IND, TNG-IND and XIB-IND	-	-
Ahmad and Dey (2011)	-	CLS-ERR	-
Cheng and Leu (2011)	-	-	-
Pham, Suarez-Alvarez and Prostov (2011)	OBJ-FNC	-	-

Table 53 Continued on next page

Table 53 continued from previous page

Document	Internal	External	Relative
Irigioien, Sierra and Arenas (2012)	EXC-TME	-	INC-IND
Ji <i>et al.</i> (2012)	-	ACC-IND	-
Liang <i>et al.</i> (2012)	CTG-UTL	AJR-IND	CTG-UTL
Cheung and Jia (2013)	EXC-TME	ACC-IND	-
[M] Hennig and Liao (2013)	AVS-IND, CLH-IND, and PVH-IND	AJR-IND	AVS-IND
Ji <i>et al.</i> (2013)	-	ACC-IND	-
Ovesná <i>et al.</i> (2013)	-	-	-
Han <i>et al.</i> (2014)	EXC-TME	-	INF-LSS
Pasquier, Naciri and Jeanmonod (2015)	-	-	-
Harikumar and Surya (2015)	DVB-IND	PRT-IND	DVB-IND and PRT-IND
Ji <i>et al.</i> (2015)	-	ACC-IND	-
Sangam and Om (2015)	EXC-TME	ACC-IND	-
Wei, Chow and Chan (2015)	-	ACC-IND	-
Ahmad and Hashmi (2016)	-	AVA-IND	-
Amorim and Makarenkov (2016)	NMB-ITR and SLH-IND	AJR-IND	-
Bishnu and Bhattacharjee (2016)	EXC-TME	-	HIT-RTO and MMR-ERR
Drakopoulos <i>et al.</i> (2016)	EXC-TME and OBJ-FNC	TVA-IND	-
Gocsik <i>et al.</i> (2016)	-	JCC-IND	-
Thong and Son (2016)	DVB-IND and WGS-IND	AVA-IND and RND-IND	-
Huerta-Muñoz, Ríos-Mercado and Ruiz (2017)	EXC-TME	-	ARP-DVT and DVB-IND
[H] Kampert, Meulman and Friedman (2017)	-	-	-
Kim (2017)	EXC-TME	-	-
Liang, Wang and Lai (2017)	-	NRM-INF	-
Özbakir and Turna (2017)	EXC-TME and OBJ-FNC	ACC-IND and ERR-RTE	-
Wangchamhan, Chiewchanwattana and Sunat (2017)	OBJ-FNC	FMS-IND and NMF-EVL	-
Bathla, Aggarwal and Rani (2018)	EXC-TME	-	SPD-VLU
Davoodi and Moradi (2018)	-	FLS-NGT, FLS-PST, PSP-VLU, SNS-IND SPC-IND, TRU-NGT, and TRU-PST	-
Deng <i>et al.</i> (2018)	-	-	SLH-IND
Gu <i>et al.</i> (2018)	OBJ-FNC and specific others for the application	-	-

Table 53 Continued on next page

Table 53 continued from previous page

Document	Internal	External	Relative
Hong <i>et al.</i> (2018)	CLH-IND, DVB-IND and SMS-ERR	-	-
Jia and Cheung (2018)	EXC-TME	ACC-IND, NRM-INF and RND-IND	PRQ-IND
Li <i>et al.</i> (2018)	-	CLS-ERR	CRR-DGR
[H] Najjar <i>et al.</i> (2018)	Specific indices built for the application	-	-
Sangam and Om (2018)	EXC-TME	-	-
Sangam and Om (2018)	EXC-TME	ACC-IND	-
Szepannek (2018)	EXC-TME	RND-IND	-
Yao <i>et al.</i> (2018)	-	ACC-IND	CTG-UTL and INC-ENT
Yuan, Zhang and Feng (2018)	-	NRC-ICV	-
HajKacem, N'cir and Essoussi (2019)	EXC-TME and SMS-ERR	-	SPD-VLU
Dutta, Sil and Dutta (2019)	C-IND, CSM-CLS, DVB-IND, DNN-IND, SMS-ERR and XIB-IND	ENT-IND, FMS-IND and PRT-IND	-
Noia <i>et al.</i> (2019)	-	ACC-IND, CMP-MSR, INF-IND, NGP-VLU, PSP-VLU, SNS-IND and SPC-IND	-
Ji <i>et al.</i> (2019)	NMB-ITR	ACC-IND	-
Lv <i>et al.</i> (2019)	NIC-VRN	-	-
Maione, Nelson and Barbosa (2019)	-	-	AVS-IND
Shukla and Nanda (2019)	DNN-IND and SLH-IND	ACC-IND, FMS-IND, PRC-IND and RND-IND	-
Yao <i>et al.</i> (2019)	-	ACC-IND	CTG-UTL and INC-ENT

Table 54 – Individual analysis, Part 3 - Partitional branch, Programming languages and code availability

Document	Programming language	Availability of the code			
		Is it available?	Language	Name	Address
Huang (1998)	not mentioned	yes	R	clustMixType	CRAN ⁵
Ng and Wong (2002)	C++	no	-	-	-
He, Xu and Deng (2005)	C	no	-	-	-
Ahmad and Dey (2007)	not mentioned	yes	R and MATLAB	DisimForMixed and Mixed k-means	CRAN ⁶ and MFE ⁷
Bushel, Wolfinger and Gibson (2007)	MATLAB	yes	MATLAB	modk-prototypes	NIEHS ⁸
Brouwer (2008)	not mentioned	no	-	-	-
Cheng and Leu (2008)	MATLAB	no	-	-	-
Cheng and Leu (2009)	MATLAB	no	-	-	-
Lee and Pedrycz (2009)	not mentioned	no	-	-	-
Lee and Pedrycz (2010)	not mentioned	no	-	-	-
Ahmad and Dey (2011)	not mentioned	no	-	-	-
Cheng and Leu (2011)	not mentioned	no	-	-	-
Pham, Suarez-Alvarez and Prostov (2011)	not mentioned	no	-	-	-
Irigoien, Sierra and Arenas (2012)	R	yes	R	ICGE	CRAN ⁹
Ji <i>et al.</i> (2012)	not mentioned	no	-	-	-
Liang <i>et al.</i> (2012)	not mentioned	no	-	-	-
Cheung and Jia (2013)	MATLAB	no	-	-	-
[M] Hennig and Liao (2013)	R	yes	R	fcp	CRAN ¹⁰

Table 54 Continued on next page

- ⁵ <https://cran.r-project.org/web/packages/clustMixType/index.html>
- ⁶ <https://cran.r-project.org/web/packages/DisimForMixed/index.html>
- ⁷ https://www.mathworks.com/matlabcentral/fileexchange/53489-amjams-mixed_kmeans
- ⁸ <https://www.niehs.nih.gov/research/resources/software/biostatistics/modk/index.cfm>
- ⁹ <https://cran.r-project.org/web/packages/ICGE/index.html>
- ¹⁰ <https://www.rdocumentation.org/packages/SensoMineR/versions/1.23/topics/fcp>

Table 54 continued from previous page

Document	Programming language	Is it available?	Availability of the code		
			Language	Name	Address
Ji <i>et al.</i> (2013)	not mentioned	no	-	-	-
Ovesná <i>et al.</i> (2013)	not mentioned	no	-	-	-
Han <i>et al.</i> (2014)	not mentioned	no	-	-	-
Pasquier, Naciri and Jeanmonod (2015)	R	no	-	-	-
Harikumar and Surya (2015)	not mentioned	no	-	-	-
Ji <i>et al.</i> (2015)	not mentioned	no	-	-	-
Sangam and Om (2015)	C++	no	-	-	-
Wei, Chow and Chan (2015)	not mentioned	no	-	-	-
Ahmad and Hashmi (2016)	not mentioned	no	-	-	-
Amorim and Makarenkov (2016)	not mentioned	yes	MATLAB	-	AW ¹¹
Bishnu and Bhattacharjee (2016)	GNU Octave	no	-	-	-
Drakopoulos <i>et al.</i> (2016)	C# and MATLAB	no	-	-	-
Gocsik <i>et al.</i> (2016)	not mentioned	no	-	-	-
Thong and Son (2016)	C	no	-	-	-
Huerta-Muñoz, Ríos-Mercado and Ruiz (2017)	C++	no	-	-	-
[H] Kampert, Meulman and Friedman (2017)	R	yes	R	rCOSA	GitHub ¹²
Kim (2017)	Java	no	-	-	-
Liang, Wang and Lai (2017)	not mentioned	no	-	-	-

Table 54 Continued on next page

¹¹ <http://www.rcamorim.com/>

¹² <https://github.com/mkampert/rCOSA/tree/master/R>

Table 54 continued from previous page

Document	Programming language	Is it available?	Availability of the code		
			Language	Name	Address
Özbakir and Turna (2017)	MATLAB	no	-	-	-
Wangchamhan, Chiewchanwattana and Sunat (2017)	MATLAB	no	-	-	-
Bathla, Aggarwal and Rani (2018)	not mentioned	no	-	-	-
Davoodi and Moradi (2018)	MATLAB	no	-	-	-
Deng <i>et al.</i> (2018)	R	no	-	-	-
Gu <i>et al.</i> (2018)	not mentioned	no	-	-	-
Hong <i>et al.</i> (2018)	not mentioned	no	-	-	-
Jia and Cheung (2018)	MATLAB	no	-	-	-
Li <i>et al.</i> (2018)	C#	no	-	-	-
[H] Najjar <i>et al.</i> (2018)	Python	no	-	-	-
Sangam and Om (2018)	C++	no	-	-	-
Sangam and Om (2018)	Java	no	-	-	-
Szepannek (2018)	R	yes	R	clustMixType	CRAN ¹³
Yao <i>et al.</i> (2018)	Java	no	-	-	-
Yuan, Zhang and Feng (2018)	MATLAB	no	-	-	-
HajKacem, N'cir and Es-soussi (2019)	not mentioned	no	-	-	-
Dutta, Sil and Dutta (2019)	not mentioned	no	-	-	-
Noia <i>et al.</i> (2019)	not mentioned	no	-	-	-
Ji <i>et al.</i> (2019)	Java	no	-	-	-
Lv <i>et al.</i> (2019)	Python	no	-	-	-
Maione, Nelson and Barbosa (2019)	not mentioned	no	-	-	-

Table 54 Continued on next page

¹³ <https://cran.r-project.org/web/packages/clustMixType/index.html>

Table 54 continued from previous page

Document	Programming language	Availability of the code		
		Is it available?	Language	Name Address
Shukla and (2019)	Nanda MATLAB	no	-	-
Yao <i>et al.</i> (2019)	Java	no	-	-

Table 55 – Individual analysis, Part 3 - Hierarchical branch, Datasets

Document	Simulated			Real-world	
	Is it used?	Address	Is it used?	Acronym	
Philip and Ottaway (1983)	no	-	yes	CHT-WPN	
Frisvad (1994)	no	-	yes	PN1-DTS and PN2-DTS	
Harch <i>et al.</i> (1996)	no	-	yes	AGG-CL1	
Sieber, Petrini and Greenacre (1998)	no	-	yes	GLR-DTS	
Friedman and Meulman (2004)	yes	not mentioned	yes	MDC-DTS	
Shin and Sohn (2004)	no	-	yes	CST-DTS	
Hsu, Chen and Su (2007)	yes	not mentioned	yes	ADL-DTS	
Sheu (2007)	no	-	yes	TWN-MNF	
Herr (2010)	no	-	yes	IND-HSH	
Chitradevi <i>et al.</i> (2012)	no	-	yes	SNS-NTW	
[M] Fonseca (2013)	no	-	yes	NRC-WSC	
Sopan <i>et al.</i> (2013)	no	-	yes	USC-DTS	
Brida <i>et al.</i> (2014)	no	-	yes	UMT-SRV	
Horta and Camanho (2014)	no	-	yes	PRT-CNT	
[P] Kampert, Meulman and Friedman (2017)	yes	R package (rCOSA) ¹⁴	yes	AP3-DTS	
[P] Najjar <i>et al.</i> (2018)	no	-	yes	ADM-HLT	
Bragato, Ganis and Feoli (2019)	no	-	yes	SOL-DTS	

¹⁴ <https://github.com/mkampert/rCOSA/tree/master/R>

Table 56 – Individual analysis, Part 3 - Hierarchical branch, Performance measures

Document	Internal	External	Relative
Philip and Ottaway (1983)	-	-	-
Frisvad (1994)	-	-	-
Harch <i>et al.</i> (1996)	-	-	-
Sieber, Petrini and Greenacre (1998)	-	-	-
Friedman and Meulman (2004)	-	-	-
Shin and Sohn (2004)	-	-	-
Hsu, Chen and Su (2007)	CTG-UTL, SMS-ERR and STD-DVT	ENT-VLD	-
Sheu (2007)	Specific indices built for the application	-	-
Herr (2010)	-	-	GAP-STT
Chitradevi <i>et al.</i> (2012)	SLH-IND and SMS-ERR	DTC-RTE, FLS-NGT and FLS-PST	-
[M] Fonseca (2013)	-	-	-
Sopan <i>et al.</i> (2013)	-	-	-
Brida <i>et al.</i> (2014)	-	AJR-IND	-
Horta and Camanho (2014)	Specific indices built for the application	-	-
[P] Kampert, Meulman and Friedman (2017)	-	-	-
[P] Najjar <i>et al.</i> (2018)	Specific indices built for the application	-	-
Bragato, Ganis and Feoli (2019)	WTB-SML	-	-

Table 57 – Individual analysis, Part 3 - Hierarchical branch, Programming languages and code availability

Document	Programming language	Availability of the code		
		Is it available?	Language	Name Address
Philip and Ottaway (1983)	not mentioned	no	-	-
Frisvad (1994)	not mentioned	no	-	-
Harch <i>et al.</i> (1996)	not mentioned	no	-	-
Sieber, Petrini and Greenacre (1998)	not mentioned	no	-	-

Table 57 Continued on next page

Table 57 continued from previous page

Document	Programming language	Availability of the code		
		Is it available?	Language	Name Address
Friedman and Meulman (2004)	not mentioned	yes	R	rCOSA Github ¹⁵
Shin and Sohn (2004)	not mentioned	no	-	-
Hsu, Chen and Su (2007)	not mentioned	no	-	-
Sheu (2007)	not mentioned	no	-	-
Herr (2010)	not mentioned	no	-	-
Chitradevi <i>et al.</i> (2012)	not mentioned	no	-	-
[M] Fonseca (2013)	not mentioned	no	-	-
Sopan <i>et al.</i> (2013)	not mentioned	no	-	-
Brida <i>et al.</i> (2014)	not mentioned	no	-	-
Horta and Camanho (2014)	not mentioned	no	-	-
[P] Kampert, Meulman and Friedman (2017)	R	yes	R	rCOSA Github ¹⁶
[P] Najjar <i>et al.</i> (2018)	Python	no	-	-
Bragato, Ganis and Feoli (2019)	not mentioned	no	-	-

Table 58 – Individual analysis, Part 3 - Model-based branch, Datasets

Document	Simulated		Real-world	
	Is it used?	Address	Is it used?	Acronym
Wong and Chiu (1987)	yes	not mentioned	yes	HDR-NTW
Harch <i>et al.</i> (1999)	no	-	yes	AGG-CL2
Hunt and Basford (1999)	no	-	yes	THW-SBN
Bacher (2000)	yes	not mentioned	yes	SRC-AST
Hunt and Jorgensen (2003)	no	-	yes	PCC-TRL
McLachlan and Chang (2004)	no	-	yes	PCC-TRL
[O] Gondek and Hofmann (2007)	yes	not mentioned	no	-
Chatzis (2011)	no	-	yes	HRD-CLV and STT-ACA
Hunt and Jorgensen (2011)	no	-	yes	CLS-EXM and HRD-CLV

Table 58 Continued on next page

¹⁵ <https://github.com/mkampert/rCOSA/tree/master/R>
¹⁶ <https://github.com/mkampert/rCOSA/tree/master/R>

Table 58 continued from previous page

Document	Simulated		Real-world	
	Is it used?	Address	Is it used?	Acronym
[H] Fonseca (2013)	no	-	yes	NRC-WSC
[P] Hennig and Liao (2013)	no	-	yes	USC-FNN
Zhang and Ip (2014)	yes	ARC-DTS and HPM-DTS	no	-
Baudry <i>et al.</i> (2015)	no	-	yes	CSW-DST
Blomstedt <i>et al.</i> (2015)	no	-	yes	AMP-DTS and MNI-DTS
Ingrassia <i>et al.</i> (2015)	yes	not mentioned	yes	ENR-EFF
Foss <i>et al.</i> (2016)	yes	not mentioned	yes	CLL-CNT, CLN-BND INC-BNC and STT-ACA
Hunt and Basford (2016)	yes	not mentioned	yes	PCC-TRL
McParland and Gormley (2016)	yes	not mentioned	yes	PCC-TRL
McParland <i>et al.</i> (2017)	no	-	yes	LVM-DTS
Tekumalla, Rajan and Bhattacharyya (2017)	yes	not mentioned	yes	ABL-DTS and MIM-II
Abidin and Westhead (2017)	yes	not mentioned	yes	CLL-CCL
Foss and Markatou (2018)	yes	R package (Kamila) ¹⁷	yes	ARL-DTS
Guglielmi <i>et al.</i> (2018)	no	-	yes	STE-ARC
Storlie <i>et al.</i> (2018)	yes	not mentioned	yes	ATS-DSR
Wang, Zhu and Dong (2018)	yes	not mentioned	yes	HRD-DTS and WLM-RCR
Biernacki and Lourme (2019)	no	-	yes	PCC-TRL, RSG-EXP and RWM-DTS
Marbac and Sedki (2019)	no	-	no	-
Marbac, Sedki and Patin (2019)	yes	not mentioned	no	-
Marbac and Vandewalle (2019)	yes	not mentioned	no	-

Table 59 – Individual analysis, Part 3 – Model-based branch, Performance measures

Document	Internal		External		Relative
	Internal		External		
Wong and Chiu (1987)	-		ERR-RTE		-
Harch <i>et al.</i> (1999)	-		-		PSF-STT
Hunt and Basford (1999)	-		-		-

Table 59 Continued on next page

¹⁷ <https://cran.r-project.org/web/packages/kamila/index.html>

Table 59 continued from previous page

Document	Internal	External	Relative
Bacher (2000)	-	-	ALR-TST and LKR-STs
Hunt and Jorgensen (2003)	-	-	-
McLachlan and Chang (2004)	-	-	-
[O] Gondek and Hofmann (2007)	EXC-TME	PRC-IND	-
Chatzis (2011)	-	MIP-IND	-
Hunt and Jorgensen (2011)	-	-	-
[H] Fonseca (2013)	-	-	AIN-CRT
[P] Hennig and Liao (2013)	AVS-IND, CLH-IND, and PVH-IND	AJR-IND	AVS-IND and BSI-CRT
Zhang and Ip (2014)	-	ACC-IND, FMS-IND and LKR-STs	-
Baudry <i>et al.</i> (2015)	-	-	AIN-CRT, BSI-CRT, INC-LKL, and SIC-LKL
Blomstedt <i>et al.</i> (2015)	-	AJR-IND	-
Ingrassia <i>et al.</i> (2015)	-	-	AI3-CRT, AIC-CRT, AIN-CRT, AIU-CRT, APW-EVD, BSI-CRT, BZD-CRT and INC-LKL
Foss <i>et al.</i> (2016)	EXC-TME and NMB-ITR	AJR-IND, MAP-IND, MAR-IND and PRT-IND	PRD-STR
Hunt and Basford (2016)	-	-	AIN-CRT, BSI-CRT, CLI-CRT, CLI-BIC and INC-LKL
McParland and Gormley (2016)	-	-	ABI-CRT
McParland <i>et al.</i> (2017)	-	-	ABI-MCM
Tekumalla, Rajan and Bhattacharyya (2017)	EXC-TME	AJR-IND, NRM-INF, RMS-ERR and VRT-INF	RMS-ERR
Abidin and Westhead (2017)	-	-	AIN-CRT (and variations) and BSI-CRT
Foss and Markatou (2018)	-	AJR-IND	PRD-STR
Guglielmi <i>et al.</i> (2018)	-	-	LPL-STT, MNS-ERR and WAI-CRT
Storlie <i>et al.</i> (2018)	EXC-TME	ACC-IND, AJR-IND, FWM-IND, and PVC-IEM	BSI-CRT

Table 59 Continued on next page

Table 59 continued from previous page

Document	Internal	External	Relative
Wang, Zhu and Dong (2018)	-	AJR-IND	-
Biernacki and Lourme (2019)	-	-	BSI-CRT and INC-LKL
Marbac and Sedki (2019)	-	-	BSI-CRT and MIC-LKL
Marbac, Sedki and Patin (2019)	-	AJR-IND	BSI-CRT, INC-LKL and MIC-LKL
Marbac and Vandewalle (2019)	EXC-TME	-	BYI-CRT and MIC-LKL

Table 60 – Individual analysis, Part 3 - Model-based branch, Programming languages and code availability

Document	Programming language	Availability of the code		
		Is it available?	Name	Address
Wong and Chiu (1987)	not mentioned	no	-	-
Harch <i>et al.</i> (1999)	not mentioned	no	-	-
Hunt and Basford (1999)	not mentioned	no	-	-
Bacher (2000)	not mentioned	no	-	-
Hunt and Jorgensen (2003)	not mentioned	no	-	-
McLachlan and Chang (2004)	not mentioned	no	-	-
[O] Gondek and Hofmann (2007)	not mentioned	no	-	-
Chatzis (2011)	not mentioned	no	-	-
Hunt and Jorgensen (2011)	not mentioned	no	-	-
[H] Fonseca (2013)	not mentioned	no	-	-
[P] Hennig and Liao (2013)	not mentioned	yes	fcp	CRAN ¹⁸
Zhang and Ip (2014)	not mentioned	no	-	-
Baudry <i>et al.</i> (2015)	R	yes	Rmixmod	CRAN ¹⁹
Blomstedt <i>et al.</i> (2015)	not mentioned	no	-	-
Ingrassia <i>et al.</i> (2015)	not mentioned	no	-	-
Foss <i>et al.</i> (2016)	not mentioned	yes	Kamila	CRAN ²⁰
Hunt and Basford (2016)	not mentioned	no	-	-

Table 60 Continued on next page

¹⁸ <https://cran.r-project.org/web/packages/fpc/index.html>

¹⁹ <https://cran.r-project.org/web/packages/Rmixmod/index.html>

²⁰ <https://cran.r-project.org/web/packages/kamila/index.html>

Table 60 continued from previous page

Document	Programming language	Availability of the code			
		Is it available?	Language	Name	Address
McParland and Gormley (2016)	R	yes	R	ClustMD	CRAN ²¹
McParland <i>et al.</i> (2017)	not mentioned	no	-	-	-
Tekumalla, Rajan and Bhat-tacharyya (2017)	not mentioned	no	-	-	-
Abidin and Westhead (2017)	Java	yes	Java	FlexiCoClustering	GitHub ²²
Foss and Markatou (2018)	R	yes	R	Kamila	CRAN ²³
Guglielmi <i>et al.</i> (2018)	R	no	-	-	-
Storlie <i>et al.</i> (2018)	R	yes	R	DPM-vs	GitHub ²⁴
Wang, Zhu and Dong (2018)	R	no	-	-	-
Biernacki and Lourme (2019)	not mentioned	no	-	-	-
Marbac and Sedki (2019)	R	yes	R	VarSelLCM	CRAN ²⁵
Marbac, Sedki and Patin (2019)	R	yes	R	VarSelLCM	CRAN ²⁶
Marbac and Vandewalle (2019)	not mentioned	no	-	-	-

Table 61 – Individual analysis, Part 3 - Neural network-based branch, Datasets

Document	Simulated		Real-world	
	Is it used?	Address	Is it used?	Acronym
Hsu and Wang (2006)	yes	not mentioned	yes	ADL-DTS and SRS-UNV
Hsu and Huang (2008)	yes	not mentioned	yes	ADL-DTS
Tai, Hsu and Chen (2010)	yes	not mentioned	yes	ADL-DTS
Hsu, Lin and Tai (2011)	yes	not mentioned	yes	ADL-DTS and STT-ACA
Hsu and Lin (2012)	yes	not mentioned	yes	ADL-DTS and STT-ACA
Liao and Hsu (2012)	yes	not mentioned	yes	ADL-DTS

Table 61 Continued on next page

- ²¹ <https://cran.r-project.org/web/packages/clustMD/index.html>
- ²² <https://github.com/BioToolsLeeds/FlexiCoClusteringPackage>
- ²³ <https://cran.r-project.org/web/packages/kamila/index.html>
- ²⁴ <https://github.com/cbstorlie/DPM-vs.git>
- ²⁵ <https://cran.r-project.org/web/packages/VarSelLCM/index.html>
- ²⁶ <https://cran.r-project.org/web/packages/VarSelLCM/index.html>

Table 61 continued from previous page

Document	Simulated		Real-world	
	Is it used?	Address	Is it used?	Acronym
Tai and Hsu (2012)	yes	not mentioned	yes	ADL-DTS and CTL-MRK
Noorbehbahani, Mousavi and Mirzaei (2014)	no	-	yes	ADL-DTS, CNM-CHC, LMP-DTS, STT-GRC and ZOO-DTS
Coso <i>et al.</i> (2015)	no	-	yes	ADL-DTS and CRD-APP
Malondkar <i>et al.</i> (2019)	no	-	yes	ATM-DTS, CNS-INC, ENR-EFF, FCB-MTR, and FRS-DTS

Table 62 – Individual analysis, Part 3 - Neural network-based branch, Performance measures

Document	Internal	External	Relative
Hsu and Wang (2006)	-	ENT-VLD	AVC-UTL
Hsu and Huang (2008)	CFF-VRT	ENT-VLD	CTG-UTL
Tai, Hsu and Chen (2010)	MNS-ERR and WTB-RTI	-	-
Hsu, Lin and Tai (2011)	-	ACC-IND, ENT-VLD and GAN-RTO	-
Hsu and Lin (2012)	CFF-VRT	ACC-IND, ENT-VLD, and GAN-RTO	-
Liao and Hsu (2012)	EXC-TME	PRT-IND and WGE-VLD	-
Tai and Hsu (2012)	MNS-ERR and WTB-RTI	ENT-VLD	-
Noorbehbahani, Mousavi and Mirzaei (2014)	-	FMB-IND, FMS-IND, FWM-IND, JCC-IND and RND-IND	-
Coso <i>et al.</i> (2015)	-	ACC-IND and PRC-IND	-
Malondkar <i>et al.</i> (2019)	EXC-TME	ACC-IND	SCP-VLU and SPD-VLU

Table 63 – Individual analysis, Part 3 - Neural network-based branch, Programming languages and code availability

Document	Programming language	Availability of the code		
		Is it available?	Language	Name Address
Hsu and Wang (2006)	C++	no	-	-
Hsu and Huang (2008)	C++	no	-	-
Tai, Hsu and Chen (2010)	Java	no	-	-
Hsu, Lin and Tai (2011)	not mentioned	no	-	-
Hsu and Lin (2012)	not mentioned	no	-	-
Liao and Hsu (2012)	Java	no	-	-
Tai and Hsu (2012)	not mentioned	no	-	-
Noorbehbahani, Mousavi and Mirzaei (2014)	not mentioned	no	-	-
Coso <i>et al.</i> (2015)	not mentioned	no	-	-
Malondkar <i>et al.</i> (2019)	SPARK	yes	SPARK	Spark-GHSOM AW ²⁷

Table 64 – Individual analysis, Part 3 - Others branch, Datasets

Document	Mixed datasets			
	Simulated		Real-world	
	Is it used?	Address	Is it used?	Acronym
Ichino and Sklansky (1985)	no	-	no	-
Veiga (1996)	no	-	yes	HRD-DTS
Hai and Susumu (2004)	no	-	yes	CRD-APP and KDD-C99
Hsu, Jao and Chen (2005)	yes	not mentioned	yes	CRC-HLD
Puthongsiriporn <i>et al.</i> (2006)	no	-	no	-
Hsu and Chen (2007)	yes	not mentioned	yes	ADL-DTS
[M] Gondek and Hofmann (2007)	no	-	no	-
Ye, Li and Orlowska (2009)	yes	not mentioned	yes	CRD-APP
Al-Shaqsi and Wang (2010)	no	-	yes	CRD-APP, HRD-CLV and ZOO-DTS

Table 64 Continued on next page

²⁷ <http://www.di.uniba.it/~ceci/ghsom/>

Table 64 continued from previous page

Document	Mixed datasets			
	Simulated		Real-world	
	Is it used?	Address	Is it used?	Acronym
Yang and Wu (2011)	yes	not mentioned	no	-
Zhao <i>et al.</i> (2013)	no	-	no	-
Wang, Machiraju and Huang (2014)	no	-	yes	CNG-ATL
Zhang and Ip (2014)	no	-	yes	CRD-APP, HRD-DTS and ZOO-DTS
Niu <i>et al.</i> (2015)	no	-	yes	WBL-SST
Chen and He (2016)	yes	not mentioned	yes	ACT-INF, CRD-APP, KDD-C99 and STT-HRD
Bhatnagar <i>et al.</i> (2017)	yes	not mentioned	yes	CVR-DTS and KDD-C99
Ding <i>et al.</i> (2017)	yes	not mentioned	yes	BNK-MRK, CRD-APP, KDD-C99, STA-HRT
Du, Ding and Xue (2017)	no	-	yes	STT-ACA and STT-HRD
Hummel, Edelmann and Kopp-Schneider (2017)	yes	not mentioned	yes	CRD-APP, HRD-DTS, KDD-C99, LMP-DTS, and STT-ACA
Iam-On and Boongoen (2017)	no	-	yes	ACT-INF, BCT-RSP, BSC-SRV and CHM-PRC
Jinyin <i>et al.</i> (2017)	no	-	yes	SML-UNV
Liu, Yang and He (2017)	no	-	yes	ACT-INF, CRD-APP, KDD-C99, STT-HRD and ZOO-DTS
Liu <i>et al.</i> (2017)	no	-	yes	ADL-DTS, STT-ACA and STT-HRD
Noorbehbahani <i>et al.</i> (2017)	no	-	yes	ACT-INF, ADL-DTS, CRD-APP, HRD-CLV and STT-HRD
Valenzuela <i>et al.</i> (2017)	no	-	yes	INT-DTC, KDD-C99 and MRE-DTS
Jang <i>et al.</i> (2018)	yes	not mentioned	yes	SNL-AGN
			yes	BRT-PST

Table 64 Continued on next page

Table 64 continued from previous page

Document	Mixed datasets			
	Simulated		Real-world	
	Is it used?	Address	Is it used?	Acronym
Lin, Azarnoush and Runger (2018)	yes	Dataset generator website ²⁸	yes	CRD-APP and KDD-C99 ADL-DTS, CRD-APP, CNM-CHC, FLG-DTS, HRD-DTS, STT-GRC, STT-HRD and TCH-AEV ABL-DTS, ADL-DTS CRD-APP and DRM-DTS
Zhao, Cao and Liang (2018)	no	-	yes	
Duan <i>et al.</i> (2019)	no	-	yes	

Table 65 – Others - Performance measures

Document	Validation measure		
	Internal	External	Relative
Ichino and Sklansky (1985)	-	-	-
Veiga (1996)	OBJ-FNC	ACC-IND	-
Hai and Susumu (2004)	-	ACC-IND	SPD-VLU
Hsu, Jao and Chen (2005)	EXC-TME	ACC-IND	-
Puthongsiriporn <i>et al.</i> (2006)	-	-	-
Hsu and Chen (2007)	CFF-VRT and EXC-TME	-	-
[M] Gondek and Hofmann (2007)	-	-	-
Ye, Li and Orlowska (2009)	EXC-TME	DTC-RTE and FLS-PST	-
Al-Shaqsi and Wang (2010)	-	ACC-IND	-
Yang and Wu (2011)	OBJ-FNC	-	-
Zhao <i>et al.</i> (2013)	-	-	-
Wang, Machiraju and Huang (2014)	-	-	NRM-INF
Zhang and Ip (2014)	-	MIP-IND	-
Niu <i>et al.</i> (2015)	-	ACC-IND	-
Chen and He (2016)	-	ACC-IND and PRT-IND	-
Bhatnagar <i>et al.</i> (2017)	EXC-TME	NRM-INF and PRT-IND	SCP-VLU and SCT-VLU

Table 65 Continued on next page

²⁸ <http://www.datasetgenerator.com/>

Table 65 continued from previous page

Document	Validation measure		
	Internal	External	Relative
Ding <i>et al.</i> (2017)	-	ACC-IND, AVA-IND, F1M-IND and NRM-INF	-
Du, Ding and Xue (2017)	-	ACC-IND	-
Hummel, Edelmann and Kopp-Schneider (2017)	-	BLE-RTE and MSC-RTE	-
Iam-On and Boongoen (2017)	-	ERR-RTE	-
Jinyin <i>et al.</i> (2017)	-	ACC-IND and PRT-IND	-
Liu, Yang and He (2017)	-	FMS-IND, PRT-IND and RCL-IND	-
Liu <i>et al.</i> (2017)	-	ACC-IND	-
Noorbehhahani <i>et al.</i> (2017)	-	F-VLE, FLS-PST, KPP-STT, MTC-CFF and TRU-PST	-
Valenzuela <i>et al.</i> (2017)	-	-	-
Jang <i>et al.</i> (2018)	EXC-TME	-	-
Lin, Azarnoush and Runger (2018)	-	CLM-MSR, PRT-IND and RND-IND	-
Zhao, Cao and Liang (2018)	-	ACC-IND, AVA-IND and NRM-INF	-
Duan <i>et al.</i> (2019)	-	ACC-IND and RND-IND	-

Table 66 – Others - Programming language and code

Document	Programming language	Availability of the code		
		Is it available?	Language	Name Address
Ichino and Sklansky (1985)	not mentioned	no	-	-
Veiga (1996)	not mentioned	no	-	-
Hai and Susumu (2004)	not mentioned	no	-	-
Hsu, Jao and Chen (2005)	Java	no	-	-
Puthongsiriporn <i>et al.</i> (2006)	not mentioned	no	-	-
Hsu and Chen (2007)	not mentioned	no	-	-
[M] Gondek and Hofmann (2007)	not mentioned	no	-	-
Ye, Li and Orlowska (2009)	MATLAB	no	-	-
Al-Shaqsi and Wang (2010)	not mentioned	no	-	-
Yang and Wu (2011)	not mentioned	no	-	-
Zhao <i>et al.</i> (2013)	not mentioned	no	-	-

Table 66 Continued on next page

Table 66 continued from previous page

Document	Programming		Availability of the code		
	language	Is it available?	Language	Name	Address
Wang, Machiraju and Huang (2014)	MATLAB	yes	MATLAB	MRCPS	Github ²⁹
Zhang and Ip (2014)	C	no	-	-	-
Niu <i>et al.</i> (2015)	not mentioned	no	-	-	-
Chen and He (2016)	not mentioned	no	-	-	-
Bhatnagar <i>et al.</i> (2017)	not mentioned	no	-	-	-
Ding <i>et al.</i> (2017)	MATLAB	no	-	-	-
Du, Ding and Xue (2017)	MATLAB	no	-	-	-
Hummel, Edelman and Kopp-Schneider (2017)	R	yes	R	CluMix	CRAN ³⁰
Iam-On and Boongoen (2017)	not mentioned	no	-	-	-
Jinyin <i>et al.</i> (2017)	not mentioned	no	-	-	-
Liu, Yang and He (2017)	not mentioned	no	-	-	-
Liu <i>et al.</i> (2017)	MATLAB	no	-	-	-
Noorbehbahani <i>et al.</i> (2017)	Java	no	-	-	-
Valenzuela <i>et al.</i> (2017)	Python	no	-	-	-
Jang <i>et al.</i> (2018)	Java	no	-	-	-
Lin, Azarnoush and Runger (2018)	R	no	-	-	-
Zhao, Cao and Liang (2018)	MATLAB	no	-	-	-
Duan <i>et al.</i> (2019)	MATLAB	no	-	-	-

Table 67 – Individual analysis, Part 3 - IAPRA branch, Datasets

Document	Simulated		Real-world	
	Is it used?	Address	Is it used?	Acronym
Gordon (1987)	no	-	no	-
Cox and Cox (2000)	no	-	yes	ELH-DTS
Shi and Horvath (2006)	yes	not mentioned	no	-
Irigoien and Arenas (2008)	yes	not mentioned	yes	DRM-DTS and LYP-CNC
Barceló-Rico, Díez and Bondia (2011)	no	-	yes	HRD-DTS

Table 67 Continued on next page

²⁹ <https://github.com/chaowang1010/MorCPS>

³⁰ <https://cran.r-project.org/src/contrib/Archive/CluMix/>

Table 67 continued from previous page

Document	Simulated		Real-world	
	Is it used?	Address	Is it used?	Acronym
Barceló-Rico and Diez (2012)	no	-	yes	ADL-DTS, CLN-BND, CRD-APP and HRD-DTS
Suarez-Alvarez <i>et al.</i> (2012)	no	-	yes	ADL-DTS, CRD-APP and STT-HRD
Bernard <i>et al.</i> (2014)	no	-	yes	MTR-SNP and PRS-CNC
Hennig and Lin (2015)	no	-	yes	USC-FNN
Lam, Wei and Wunsch (2015)	no	-	yes	CRD-APP, HRD-CLV, PTR-DTS and TCH-AEV
Prostov, Suarez-Alvarez and Prostov (2015)	no	-	yes	ADL-DTS and HRD-DTS
Hsu and Huang (2016)	yes	not mentioned	yes	ADL-DTS and CRD-APP
Rahman and Islam (2016)	yes	not mentioned	yes	ADL-DTS, AUT-MP, CHS-KRK, CNM-CHC, HSN-DTS and STT-GRC.
Taha and Hadi (2016)	yes	not mentioned	no	-
Kuppusamy and Paramasivam (2017)	no	-	yes	HRD-DTS
Audigier <i>et al.</i> (2018)	yes	not mentioned	yes	GRE-DTS
Nowosad and Stepinski (2018)	no	-	yes	EE3-MAP
Ahmad and Khan (2019)	no	-	no	-
Foss, Markatou and Ray (2019)	yes	not mentioned	no	-
Golinko and Zhu (2019)	no	-	yes	CHK-DSS, MMM-MSS, STA-PRF and SSM-BMP
Velden, Enza and Markos (2019)	no	-	yes	F18-DTS, HRD-CLV and TPG-DTS

Table 68 – IAPRA - Performance measures

Document	Internal	External	Relative
Gordon (1987)	-	-	-
Cox and Cox (2000)	-	-	-
Shi and Horvath (2006)	-	ERR-RTE and RND-IND	-
Irigoien and Arenas (2008)	-	-	INC-IND and SLH-IND

Table 68 Continued on next page

Table 68 continued from previous page

Document	Internal	External	Relative
Barceló-Rico, Díez and Bondia (2011)	EXC-TME	ACC-IND	-
Barceló-Rico and Díez (2012)	EXC-TME	ACC-IND	-
Suarez-Alvarez <i>et al.</i> (2012)	-	ACC-IND and RND-IND	-
Bernard <i>et al.</i> (2014)	-	NRM-INF	-
Hennig and Lin (2015)	-	-	AVS-IND
Lam, Wei and Wunsch (2015)	-	ACC-IND and RND-IND	-
Prostov, Suarez-Alvarez and Prostov (2015)	-	ACC-IND and RND-IND	-
Hsu and Huang (2016)	-	ACC-IND	-
Rahman and Islam (2016)	-	MNA-ERR and RMS-ERR	RMS-ERR
Taha and Hadi (2016)	EXC-TME	-	BIS-VLU and RMS-ERR
Kuppusamy and Paramasivam (2017)	-	-	MNS-ERR and RMS-ERR
Audigier <i>et al.</i> (2018)	-	BIS-VLU, CAC-INT, EMC-SER, RMS-ERR RMS-ESE	BIS-VLU
Nowosad and Stepinski (2018)	-	-	CPL-MTR, HMG-MTR and VMS-MTR
Ahmad and Khan (2019)	-	-	-
Foss, Markatou and Ray (2019)	-	AJR-IND	-
Golinko and Zhu (2019)	EXC-TME	PRT-IND	-
Velden, Enza and Markos (2019)	-	AJR-IND	-

Table 69 – IAPRA - Programming language and code

Document	Programming language	Availability of the code			
		Is it available?	Language	Name Address	
Gordon (1987)	not mentioned	no	-	-	-
Cox and Cox (2000)	not mentioned	no	-	-	-
Shi and Horvath (2006)	R	yes	R	randomForest	CRAN ³¹

Table 69 Continued on next page

Table 69 Continued on next page

³¹ <https://cran.r-project.org/web/packages/randomForest/index.html>

Table 69 continued from previous page

Document	Programming language	Availability of the code			
		Is it available?	Language	Name	Address
Irigoien and Arenas (2008)	not mentioned	yes	R	ICGE	CRAN ³²
Barceló-Rico, Díez and Bondia (2011)	not mentioned	no	-	-	-
Barceló-Rico and Díez (2012)	not mentioned	no	-	-	-
Suarez-Alvarez <i>et al.</i> (2012)	not mentioned	no	-	-	-
Bernard <i>et al.</i> (2014)	not mentioned	no	-	-	-
Hennig and Lin (2015)	not mentioned	no	-	-	-
Lam, Wei and Wünsch (2015)	not mentioned	no	-	-	-
Prostov, Suarez-Alvarez and Prostov (2015)	not mentioned	no	-	-	-
Hsu and Huang (2016)	not mentioned	no	-	-	-
Rahman and Islam (2016)	Java	no	-	-	-
Taha and Hadi (2016)	not mentioned	no	-	-	-
Kuppusamy and Paramasivam (2017)	MATLAB	no	-	-	-
Audigier <i>et al.</i> (2018)	R	yes	R	-	ASM ³³
Nowosad and Stepinski (2018)	R	yes	R	SABRE	CRAN ³⁴
Ahmad and Khan (2019)	not mentioned	no	-	-	-
Foss, Markatou and Ray (2019)	C++ and R	no	-	-	-
Golinko and Zhu (2019)	R	yes	R	-	Github ³⁵
Velden, Enza and Markos (2019)	R	no	-	-	-

³² <https://cran.r-project.org/web/packages/ICGE/index.html>³³ <https://projecteuclid.org/euclid.ss/1525313140#supplemental>³⁴ <https://cran.r-project.org/web/packages/sabre/index.html>³⁵ <https://github.com/egolinko/GEL>

C.1 Meaning of the acronyms

This section brings the meaning of the acronyms used in the tables presented in the previous section. It is also provided some information about the real-world datasets, such as the address, number of instances, numerical and categorical attributes, when this information was available. Note the authors can have used a different number of instances, numerical and categorical attributes in their experiments. Therefore, the information showed in the following tables contemplates the information the authors written in their documents, or when they did not provide any information, one took it from the website of the datasets, when it was available. Thus, the described characteristics of the datasets might not match the original ones.

With respect to the acronyms that are not in the following tables, they are: i) MFE³⁶: MATLAB file exchange; ii) CRAN³⁷: Comprehensive R archive network; iii) NIEHS: National institute of environmental health sciences; iv) UCI³⁸: UC Irvine machine learning repository; v) KEEL³⁹: Knowledge extraction based on evolutionary learning dataset repository; vi) ASM: Article's supplementary material, and vii) AW: Author's website

³⁶ <https://www.mathworks.com/matlabcentral/fileexchange/>

³⁷ <https://cran.r-project.org/>

³⁸ <https://archive.ics.uci.edu/ml/index.php>

³⁹ <http://www.keel.es/>

Table 70 – Description of the datasets - Acronyms, address and composition

Real-world datasets					
Acronym	Name	Address	#Numerical	#Categorical	#Instances
ABL-DTS	Abalone dataset	UCI	8	1	4,177
ACT-DTS	Acetaminophen dataset	not specified	10	6	48
ACT-INF	Acute inflammations	UCI	1	4 (+1 label)	120
ADL-DTS	Adult dataset (also known as Salary and Census Income dataset)	UCI	6	8	48,842
ADM-HLT	Administrative healthcare	not specified	not specified	not specified	not specified
AGG-CL1	Australian groundnut germplasm collection 1	not specified	7	9	831
AGG-CL2	Australian groundnut germplasm collection 2	not specified	7	9	11,436
AGR-DTS	Agronomic dataset	not specified	11	4	41
AMP-DTS	Amphetamine dataset	not specified	not specified	not specified	not specified
ANN-DTS	Annealing dataset	UCI	7	30 (+1 label)	798
AP3-DTS	ApooE3 dataset	R package (rCOsa ⁴⁰)	not specified	not specified	not specified
ARC-DTS	Archeometry dataset	Moustaki and Papageorgiou (2005)	21	9	1,000
ARL-DTS	Airline dataset	Article's supplementary material	7	5	not specified
ATM-DTS	Automobile dataset	UCI	15	10 (+1 label)	205
ATS-DSR	Autism spectrum disorder	not specified	not specified	not specified	486
AUT-MPG	Auto MPG	UCI	5	4	398
BCT-RSP	Breast cancer treatment response	not specified	not specified	not specified	133
BMR-DTS	Biomarkers dataset	not specified	not specified	not specified	10,000
BNK-MRK	Bank marketing	UCI	10	10 (+1 label)	45,211
BRT-PST	British postcodes	Getthedata website ⁴¹	not specified	not specified	1,048,576
BSC-SRV	Breast cancer survival	not specified	not specified	not specified	144
CCM-NS1	COCOMO NASA 1	PROMISE website ⁴²	2	15	60

Table 70 Continued on next page

⁴¹ <https://www.getthedata.com/open-postcode-geo>⁴² <http://promise.site.uottawa.ca/SERepository>

Table 70 continued from previous page

Real-world datasets					
Acronym	Name	Address	#Numerical	#Categorical	#Instances
CCM-NS2	COCOMO NASA 2	PROMISE website ⁴³	2	19	93
CER-HSH	Ceará's households	not specified	38	38	476
CHK-DSS	Chronic kidney disease	UCI	11	13	400
CHM-PRC	Chemical manufacturing process dataset	R package (APM ⁴⁴)	not specified	not specified	176
CHS-KRK	Chess (king-rook vs. king)	UCI	6	10	28,056
CHT-WPN	Cypriot hooked-tang weapons	not specified	not specified	not specified	not specified
CHW-DTS	Chowdary's dataset	R package (ICGE ⁴⁵)	not specified	not specified	183
CLL-CCL	Cell cycle	not specified	not specified	not specified	not specified
CLL-CNT	Call center	not specified	2	7	44,518
CLN-BND	Cylinder bands	UCI	19	21	512
CLS-EXM	Class examination	not specified	not specified	not specified	66
CMD-SRV	China migrants dynamic survey	not specified	not specified	not specified	82,985
CNG-ATL	Cancer genome atlas	National C. Institute ⁴⁶	not specified	not specified	441
CNM-CHC	Contraceptive method choice	UCI	2	7 (+1 label)	1473
CNS-INC	Census-income (KDD)	UCI	6	10	48,842
CRC-HLD	Credit card holders	not specified	not specified	not specified	2,101,094
CRD-APP	Credit approval	UCI	6	9 (+1 label)	690
CST-DTS	Customer dataset	not specified	not specified	not specified	10,000
CSW-DST	Customers of a wholesale distributor	not specified	6	2	400
CTL-MRK	Catalog marketing	not specified	not specified	not specified	not specified
CVR-DTS	Coverttype dataset	UCI	10	43 (+1 label)	581,012
DRM-DTS	Dermatology dataset	UCI	1	32 (+1 label)	366
DTO-RCR	Database operation record	not specified	not specified	not specified	not specified
EE3-MAP	EPA evel III map	not specified	not specified	not specified	85
ENR-EFF	Energy efficiency	UCI	2	6	768
EXE-WRH	Exceptional events in warehouse	not specified	2	3	25
F18-PLR	FIFA 18 player	Kaggle	43	4	1,039
FAR-DTS	FARS dataset	KEEL	5	24	100,968

Table 70 Continued on next page

⁴³ <http://promise.site.uottawa.ca/SERepository>

⁴⁶ <https://portal.gdc.cancer.gov/>

Table 70 continued from previous page

Real-world datasets					
Acronym	Name	Address	#Numerical	#Categorical	#Instances
FCB-MTR	Facebook metrics	UCI	4	3	500
FLG-DTS	Flag dataset	UCI	10	18	194
FRS-DTS	Forest dataset	UCI	10	2	517
GLB-DTS	Golub's dataset	R package (ICGE ⁴⁷)	not specified	not specified	72
GLR-DTS	Galerina dataset	not specified	3	13	38
GRE-DTS	GREAT dataset	Authors' paper	7	2	28
HDR-NTW	Hydrometric network	not specified	3	4	131
HPM-DTS	Hapmap2 dataset	Frazer <i>et al.</i> (2007)	not specified	not specified	not specified
HRD-CLV	Heart disease (Cleveland)	UCI	6	7 (+1 label)	303
HRD-DTS	Heart disease dataset	UCI	6	7 (+1 label)	303
HSN-DTS	Housing dataset	UCI	11	3	506
INC-BNC	Insurance company benchmark (COIL 2000)	UCI	not specified	not specified	9,000
IND-HSH	Indonesian households	not specified	47	198	2,819
INT-DTC	Intrusion detection	Sperotto <i>et al.</i> (2009)	not specified	not specified	not specified
JPC-SCR	Japanese credit screening	UCI	9	6	125
KDD-C99	KDD cup 1999	UCI	34	7	4,94,020
LBR-RLT	Labor relations	UCI	8	9 (+1 label)	57
LMB-DTS	Lombardy dataset	not specified	3	3	9,676
LMP-CNC	Lymphatic cancer dataset	R package (ICGE ⁴⁸)	1	18	148
LMP-DTS	Lymphography dataset	UCI	3	16	148
LV/M-DTS	LIPGENE-SU.VI.MAX dataset	UCD website ⁴⁹	26	801	1,754
LV/R-DSR	Liver disorders	UCI	4	2	345
MDC-DTS	Medical dataset	not specified	4	7	242
MIM-II	MIMIC-II	MIT website ⁵⁰	not specified	not specified	800
MIM-III	MIMIC-III	MIT website ⁵¹	not specified	not specified	not specified
MMM-MSS	Mammographic mass	UCI	2	3	961
MNI-DTS	MNIST dataset	Yann website ⁵²	not specified	not specified	10,000

Table 70 Continued on next page

- ⁴⁹ www.ucd.ie/lipgene
⁵⁰ <https://mimic.physionet.org/>
⁵¹ <https://mimic.physionet.org/>
⁵² <http://yann.lecun.com/exdb/mnist/>

Table 70 continued from previous page

Real-world datasets					
Acronym	Name	Address	#Numerical	#Categorical	#Instances
MRE-DTS	Moore dataset	not specified	not specified	not specified	not specified
MTR-SNP	Meteorological synoptical	not specified	not specified	not specified	92,902
NRC-WSC	North central Wisconsin	not specified	not specified	not specified	not specified
NTN-DTS	National dataset	not specified	3	3	9,676
PCC-TRL	Prostate cancer clinical trial	Lib website ⁵³	8	4	506
PCC-TWVN	Public construction committee of Taiwan	not specified	not specified	not specified	94
PCN-STR	Pecan street	not specified	not specified	not specified	not specified
PKR-HND	Poker hand	UCI	5	5 (+1 label)	1,025,010
PN1-DTS	Penicillium 1 dataset	not specified	11	23	44
PN2C-DTS	Penicillium 2 dataset	not specified	11	7	44
PNT-DTS	Plant dataset	Article webpage ⁵⁴	43	41	92
PRS-CNC	Prostate cancer	not specified	not specified	not specified	13,517
PRT-CNT	Portuguese contracts	INCI website ⁵⁵	not specified	not specified	18,074
PTR-DTS	Petroleum dataset	Oil and Gas Journal ⁵⁶	7	1	460
RSG-EXP	RNA sequence gene expression	Bowtie website ⁵⁷	not specified	not specified	30,000
RWM-DTS	rwm1984COUNT dataset	not specified	5	6	3,874
SBP-FRM	Survey with broiler and fattening pig farmers	not specified	3	7	37
SML-UNV	Students from Mae Fah Luang university	not specified	17	4	811
SNL-AGN	Singapore longitudinal aging	not specified	638	898	1,581
SNS-NTW	Sensor networks	not specified	not specified	not specified	not specified
SOL-DTS	Soil dataset	not specified	4	7	105
SRC-AST	Survey of children in Austria (1960)	not specified	not specified	not specified	not specified

Table 70 Continued on next page

⁵³ <http://lib.stat.cmu.edu/datasets/Andrews/T46.1>.

⁵⁴ <https://link.springer.com/article/10.1007/s00606-015-1215-0>

⁵⁵ www.inci.pt

⁵⁶ <https://www.ogj.com/>

⁵⁷ <http://bowtie-bio.sourceforge.net/recount/>

Table 70 continued from previous page

Real-world datasets					
Acronym	Name	Address	#Numerical	#Categorical	#Instances
SRS-UNV	Sales records of a store at a university	not specified	2	2	6,863
SRV-DTS	Servo dataset	UCI	1	3	167
SSM-BMP	Seismic-bumps dataset	UCI	14	4	2,584
STA-HRT	South Africa hearth	UCI	8	1	462
STA-PRF	Students' academic performance	Kaggle	4	12	480
STE-ARC	STEMI archive	not specified	5	11	697
STT-ACA	Statlog (Australian credit approval)	UCI	6	7 (+1 label)	690
STT-GRC	Statlog (German credit)	UCI	6	13 (+1 label)	1,000
STT-HRD	Statlog (heart)	UCI	6	6 (+1 label)	270
TCH-AEV	Teaching assistant evaluation	UCI	1	3 (+1 label)	151
THW-SBN	Three-way soybean	not specified	5	1	58
TPG-DTS	TopGear dataset	R package (robustHD ⁵⁸)	13	16	242
TWN-MNF	Taiwan manufacture	not specified	not specified	not specified	not specified
UMT-SRV	Uruguayan ministry of tourism survey	not specified	2	7	3,537
USC-DTS	U.S. census dataset	not specified	not specified	not specified	3,018
USC-FNC	US consumer finances	BGFRS website ⁵⁹	5	3	20,090
USC-FNN	U.S. consumer finances dataset	B. W. Publishing website ⁶⁰	not specified	not specified	17,430
WBL-SST	Web-based learning system	not specified	not specified	not specified	not specified
WCH-SPR	WeChat sports	not specified	2	4	80
WLM-RCR	Walmart recruiting dataset	Kaggle	1	3	647,054
ZOO-DTS	Zoo dataset	UCI	1	16 (+1 label)	101

⁵⁹ https://www.federalreserve.gov/econres/scf_2007.htm⁶⁰ <http://www.blackwellpublishing.com/rss>

Table 71 – Acronyms for validation measures

Validation measures					
Internal			External		
Acronym	Name	Acronym	Name	Acronym	Name
C-IND	C index	ACC-IND	Accuracy index	ABI-CRT	Approximated Bayesian information criterion
CFF-VRT	Coefficient of variation	AJR-IND	Adjusted Rand index	ABI-MCM	Approximate Bayesian information and Markov chain Monte Carlo criterion
CHI-IND	χ^2 index	AVA-IND	Average accuracy index	ALR-TST	Adjusted likelihood ratio statistic
CLH-IND	Calinski-Harabasz index	BLE-RTE	Balanced error rate	AIC-CRT	Akaike information 3 criterion
CRS-CLS	Crispness of clusters	BIS-VLU	Bias value	AIC-CRT	Akaike information C criterion
CSM-CLS	Compactness and separation measure of clusters	CAC-INT	Coverage of the associated confidence interval	AIN-CRT	Akaike information criterion
CTG-UTL	Category utility	CLM-MSR	Cluster mapping measure	AIU-CRT	Akaike information U criterion
DNN-IND	Dunn index	CLS-ERR	Clustering error	APW-EVD	Approximate weight of evidence
DVB-IND	Davies-Bouldin index	CMP-MSR	Complexity measure	ARP-DVT	Average relative percentage deviation
EXC-TME	Execution time	DTC-RTE	Detection rate	AVS-IND	Average silhouette index
MDP-CFF	Modified partition coefficient	EMC-SER	Empirical Monte Carlo standard error	AVC-UTL	Average category utility
MNS-ERR	Mean squared error	ENT-VLD	Entropy validation	BIS-VLU	Bias value
NIC-VRN	Normalized intra-cluster variance	ERR-RTE	Error rate	BSI-CRT	Bayesian information criterion
NMB-ITR	Number of iterations	F1M-IND	F1-Score index	BZD-CRT	Bozdogan criterion
NMF-EVL	Number of function evaluations	FLS-NGT	False negatives	CLH-IND	Calinski-Harabasz index
OBJ-FNC	The objective function itself	FLS-PST	False positives	CLI-BIC	Classification likelihood information and Bayesian information criterion
PRB-IND	Proximity-based index	FMB-IND	F-Measure BCubed index	CLI-CRT	Classification likelihood information criterion
PVH-IND	Pearson version of Hubert's index	FMS-IND	F-Measure index	CPL-MTR	Completeness metric
SLH-IND	Silhouette index	FWM-IND	Fowlkes-Mallows index	CRR-DGR	Correlation degree
SMS-ERR	Sum squared error	GAN-RTO	Gain ratio	CRS-CLS	Crispness of clusters
STD-DVT	Standard deviation	INF-IND	Informedness index	CTG-UTL	Category utility
TNG-IND	Tang's index	JCC-IND	Jaccard index	DVB-IND	Davies-Bouldin index
WGS-IND	Weighted silhouette index	KPP-STT	Kappa statistic	DVL-CTU	Dynamic validity index with category utility
WTB-RTI	Within-between ratio	MAP-IND	Macro precision index	GAP-STT	Gap statistic
WTB-SML	Within-between similarity	MAR-IND	Macro recall index	HIT-RTO	Hit ratio
XIB-IND	Xie and Ben's index	MIP-IND	Micro precision index	HMG-MTR	Homogeneity metric
		MNA-ERR	Mean absolute error	INC-ENT	Intra-cluster entropy for mixed data
		MNS-ERR	Mean squared error	INC-IND	INCA index

Table 71 Continued on next page

Table 71 continued from previous page

Validation measures			
Internal		External	
Acronym	Name	Acronym	Name
MSC-RTE	Misclassification rate	INC-LKL	Integrated completed likelihood
MTC-CFF	Matthews correlation coefficient	INF-LSS	Information loss
NGP-VLU	Negative predictive value	LKR-STs	Likelihood ratio statistic
NRC-ICV	Number of rare classes included (coverage)	LPL-STT	Log pseudo-marginal likelihood statistic
NRC-ITR	Number of rare classes included (top ratio)	MDP-CFF	Modified partition coefficient
NRM-INF	Normalized mutual information	MIC-LKL	Maximum integrated complete likelihood
PRC-IND	Precision index	MMR-ERR	Mean magnitude of the relative error
PRT-IND	Purity index	MNS-ERR	Mean squared error
PSP-VLU	Purity index	PSF-STT	Pseudo F statistic
RCL-IND	Recall index	PRB-IND	Proximity based index
RMS-ERR	Root mean squared error	PRD-STR	Prediction strength
RMS-ESE	Root mean square of estimated standard error	PRQ-IND	Partition quality index
RND-IND	Rand index	PVC-IEM	Proportion of variables correctly included/excluded from the model
SNS-IND	Sensitivity index	RMS-ERR	Root mean squared error
SPC-IND	Specificity index	SCT-VLU	Scaleout value
TRU-NGT	True negatives	SCP-VLU	Scaleup value
TRU-PST	True positives	SIC-LKL	Supported integrated completed likelihood
TVA-IND	Tversky asymmetric index	SLH-IND	Silhouette index
VRT-INF	Variation of information	SPD-VLU	Speedup value
WGE-VLD	Weighted entropy validation	TNG-IND	Tang's index
		VMS-MTR	V-measure metric
		WAI-CRT	Watanabe Akaike information criterion
		WGS-IND	Weighted silhouette index
		XBN-IND	Xie and Ben's index

D APPENDIX D

Feature balancing

D.1 Parameters

Table 72 shows the parameters' values of the algorithms considered in Chapter 5, which were defined by the irace method.

D.2 Results - Tables

In the following subsections, we show the average of the weights assigned for the numerical and categorical attributes, and the execution time values.

Note that in all tables the bold values are statistically better than the others in the same group with respect to the ARI measure. When KMILA or FMMMD values are in bold, it means that the respective algorithm is the best algorithm of all methods of the four groups, and when one of both methods has a letter as an exponent, it means that the respective algorithm is better than the group represented by that letter, where group 1 is represented by letter the "K", group 2 by the "W", group 3 by the "P", and group 4 by the "Y". The terms better and best mentioned above refer to the ARI measure.

D.2.1 Experiment A

Tables from 73 to 76 present the results of experiment A1 to A2. In those tables, the following information is provided: the average percentage relative deviation (APRD) regarding the ARI measure, the average and the standard deviation of the weights assigned to the numerical and categorical attributes in each method and dataset, and the execution time each algorithm needed to carry out all the repetitions, respectively.

D.2.1.1 A1 - Normal distribution

D.2.1.2 A2 - LogNormal distribution

D.2.2 Experiment B

Tables from 81 to 84 present the results of experiment B1 to B2. In those tables, the following information is provided: the average percentage relative deviation (APRD) regarding the ARI measure, the average and the standard deviation of the weights assigned to the numerical and categorical attributes in each method and dataset, and the execution time each algorithm needed to carry out all the repetitions, respectively.

Table 72 – Parameters' values defined by the irace method

Experiment	Algorithms	Modha-Spangler algorithm	UFBMD-BRKGA algorithm			
		searchDensity	npop	p_e	p_m	ρ
A1	WKMMMS	5	-	-	-	-
A1	KMDBR	-	9	0.2	0.2	0.7
A1	WKMBR	-	9	0.2	0.2	0.7
A1	PMGBR	-	9	0.2	0.2	0.7
A1	KPTBR	-	9	0.2	0.2	0.7
A2	WKMMMS	2	-	-	-	-
A2	KMDBR	-	9	0.2	0.2	0.7
A2	WKMBR	-	9	0.2	0.2	0.7
A2	PMGBR	-	9	0.2	0.2	0.7
A2	KPTBR	-	9	0.2	0.2	0.7
B1	WKMMMS	5	-	-	-	-
B1	KMDBR	-	9	0.2	0.2	0.7
B1	WKMBR	-	11	0.5	0.4	0.1
B1	PMGBR	-	14	0.5	0.2	0.3
B1	KPTBR	-	9	0.2	0.2	0.7
B2	WKMMMS	2	-	-	-	-
B2	KMDBR	-	9	0.2	0.2	0.7
B2	WKMBR	-	9	0.2	0.2	0.7
B2	PMGBR	-	9	0.2	0.2	0.7
B2	KPTBR	-	9	0.2	0.2	0.7
C1	WKMMMS	2	-	-	-	-
C1	KMDBR	-	13	0.6	0.1	0.3
C1	WKMBR	-	11	0.6	0.2	0.7
C1	PMGBR	-	12	0.1	0.7	0.6
C1	KPTBR	-	9	0.2	0.2	0.7
C2	WKMMMS	6	-	-	-	-
C2	KMDBR	-	9	0.7	0.1	0.8
C2	WKMBR	-	12	0.4	0.3	0.4
C2	PMGBR	-	15	0.5	0.2	0.8
C2	KPTBR	-	13	0.2	0.6	0.2
D1	WKMMMS	5	-	-	-	-
D1	KMDBR	-	9	0.2	0.2	0.7
D1	WKMBR	-	9	0.2	0.2	0.7
D1	PMGBR	-	9	0.2	0.2	0.7
D1	KPTBR	-	9	0.2	0.2	0.7
D2	WKMMMS	-	-	-	-	-
D2	KMDBR	-	9	0.2	0.2	0.7
D2	WKMBR	-	9	0.2	0.2	0.7
D2	PMGBR	-	9	0.2	0.2	0.7
D2	KPTBR	-	9	0.2	0.2	0.7

D.2.2.1 B1 - Normal distribution

D.2.2.2 B2 - LogNormal distribution

D.2.3 Experiment C

Tables from 89 to 96 present the results of experiment C1 to C2. In those tables, the following information is provided: the APRD regarding the ARI measure, the average and the standard deviation of the weights assigned to the numerical and categorical attributes

Table 73 – Results of experiment A1 - ARPD values considering the ARI values of each repetition

Dataset		Group 1 (K)				Group 2 (W)				Group 3 (P)				Group 4 (Y)					
[ID]	Scenario	KAMLA	FMMMD	KMDBS	KMDBR	KAMLA	FMMMD	WKMBs	WKMBR	KAMLA	FMMMD	PMGBS	PMGBR	KAMLA	FMMMD	KPTBS	KPTNM	KPTBL	KPTBR
[1]	02-08/0.0-0.0	0.016	0	0	0	0.016	0	0	0	0.016	0	0	0	0.016	0	0	0	0	0
[2]	02-08/0.0-0.3	0.016	0	0.016	0	0.016	0	0.016	0	0.016	0	0	0	0.016	0	0	0	0	0
[3]	02-08/0.0-0.7	0.016	0	0.016	0	0.016	0	0.016	0	0.016	0	0.064	0	0.016	0	0.064	0.016	0	0
[4]	02-08/0.3-0.0	0.096	0	0	0	0.096	0	0	0	0.096	0	0	0	0.096	0	0	0	0	0
[5]	02-08/0.3-0.3	0.112	0	0.016	0.016	0.112	0	0.016	0.016	0.016	0.112	0	0	0.112	0	0.032	0	0.032	0.032
[6]	02-08/0.3-0.7	0.128	0	0.016	0.016	0.128	0	0.016	0.016	0.016	0.128	0	0.128	0.048	0.128	0	0.32	0.128	0.304
[7]	02-08/0.7-0.0	0.112	0	0	0	0.112	0	0	0	0.112	0	0	0	0.112	0	0	0	0	0
[8]	02-08/0.7-0.3	0.288	0	0.016	0.016	0.288	0	0.016	0.032	0.016	0.288	0	0.048	0.288	0	0.048	0.016	0.096	0.048
[9]	02-08/0.7-0.7	0.288	0	0.977	0.016	0.288	0	0.977	0.016	0.016	0.288	0	0.128	0.064	0.288	0	2.506	0.192	1.965
[10]	08-02/0.0-0.0	0	0	0.016	0	0	0	0.016	0.016	0	0	0	0	0	0	0.016	0	0	0
[11]	08-02/0.0-0.3	0	0	0.32	0	0	0	0.32	0	0	0	0	0	0	0	0.32	0	0	0
[12]	08-02/0.0-0.7	0	0	25.35	0	0	0	25.35	0	0	0	0	0	0	0	0.591	0	0	0
[13]	08-02/0.3-0.0	0.016	0	0	0	0.016	0	0	0	0.016	0	0	0	0.016	0	0	0	0	0
[14]	08-02/0.3-0.3	0.064	0	6.696	0	0.064	0	6.696	0.224	0	0.064	0	0.048	0	0.064	0	0.24	0	0
[15]	08-02/0.3-0.7	0.048	0	28.256	0	0.048	0	28.256	0.288	0	0.048	0	0.128	0	0.048	0	0.736	0	0.016
[16]	08-02/0.7-0.0	1.248	0	0.4	0.208	1.248	0	0.4	0.288	0.224	1.248	0	0	1.248	0	0.208	0	5.481	0.16
[17]	08-02/0.7-0.3	2.257	0	15.31	1.135	2.257	0	15.31	1.627	1.614	2.257	0	0.144	0	2.257	0	4.11	0	9.555
[18]	08-02/0.7-0.7	2.573	0	27.542	1.774	2.573	0	27.542	1.883	1.854	2.573	0	0.512	0	2.573	0	10.872	0.016	17.238
[19]	20-80/0.0-0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[20]	20-80/0.0-0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[21]	20-80/0.0-0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[22]	20-80/0.3-0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[23]	20-80/0.3-0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[24]	20-80/0.3-0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[25]	20-80/0.7-0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[26]	20-80/0.7-0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[27]	20-80/0.7-0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[28]	80-20/0.0-0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[29]	80-20/0.0-0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[30]	80-20/0.0-0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[31]	80-20/0.3-0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[32]	80-20/0.3-0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[33]	80-20/0.3-0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.999	0	0	0
[34]	80-20/0.7-0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[35]	80-20/0.7-0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[36]	80-20/0.7-0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.999	0	0	0

in each method and dataset, and the execution time each algorithm needed to carry out all the repetitions, respectively.

D.2.3.1 C1 - Normal distribution

D.2.3.2 C2 - LogNormal distribution

D.2.4 Experiment D

Tables from 97 to 106 present the results of experiment D1 to D2. In those tables, the following information is provided: the ARI values, the APRD regarding the ARI measure, the average and the standard deviation of the weights assigned to the numerical and categorical attributes in each method and dataset, and the execution time each algorithm needed to carry out all the repetitions, respectively.

D.2.4.1 D1 - Increasing the number of instances

D.2.4.2 D2 - Increasing the number of attributes

D.3 Results - Figures

In this section, we provide Figures from 50 to 50 to show graphically the average of the ARI value of each algorithm in each dataset, separated in groups. Besides, Figures from 54 and 55 show average of the execution time of each algorithm in each dataset.

Table 74 – Results of experiment A1 - Average and standard deviation of numerical attributes weights

Dataset		All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)		Group 4 (Y)			
[ID]	Scenario	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTel	KPTBR
[1]	02-08/0.0-0.0	1	1	1	1	0.5	0.188	0.874	0.5	0.874	1	1	1	1
		±0	±0	±0	±0	±0	±0.074	±0.015	±0	±0.015	±0	±0	±0	±0
		1	1	1	1	0.5	0.217	0.874	0.5	0.874	1	1	1	1
[2]	02-08/0.0-0.3	±0	±0	±0	±0	±0	±0.133	±0.015	±0	±0.015	±0	±0	±0	±0
		1	1	1	1	0.5	0.22	0.874	0.5	0.874	1	1	1	1
		±0	±0	±0	±0	±0	±0.127	±0.015	±0	±0.015	±0	±0	±0	±0
[3]	02-08/0.0-0.7	1	1	1	1	0.5	0.215	0.683	0.5	0.936	1	1	1	1
		±0	±0	±0	±0	±0	±0.121	±0.217	±0	±0.008	±0	±0	±0	±0
		1	1	1	1	0.5	0.233	0.616	0.5	0.923	1	1	1	1
[5]	02-08/0.3-0.3	±0	±0	±0	±0	±0	±0.157	±0.192	±0	±0.075	±0	±0	±0	±0
		1	1 ^{Py}	1	1	0.5	0.245	0.549	0.5	0.918	1	1	1	1
		±0	±0	±0	±0	±0	±0.161	±0.161	±0	±0.09	±0	±0	±0	±0
[6]	02-08/0.3-0.7	1	1	1	1	0.5	0.227	0.496	0.5	0.943	1	1	1	1
		±0	±0	±0	±0	±0	±0.141	±0.05	±0	±0.095	±0	±0	±0	±0
		1	1	1	1	0.5	0.252	0.496	0.5	0.925	1	1	1	1
[8]	02-08/0.7-0.3	±0	±0	±0	±0	±0	±0.181	±0.05	±0	±0.126	±0	±0	±0	±0
		1	1 ^{Py}	1	1	0.5	0.248	0.487	0.5	0.878	1	1	1	1
		±0	±0	±0	±0	±0	±0.163	±0.083	±0	±0.181	±0	±0	±0	±0
[10]	08-02/0.0-0.0	1	1	1	1	0.5	0.222	0.89	0.5	0.89	1	1	1	1
		±0	±0	±0	±0	±0	±0.109	±0.008	±0	±0.008	±0	±0	±0	±0
		1	1	1	1	0.5	0.498	0.89	0.5	0.89	1	1	1	1
[11]	08-02/0.0-0.3	±0	±0	±0	±0	±0	±0.197	±0.008	±0	±0.008	±0	±0	±0	±0
		1	1	1	1	0.5	0.58	0.89	0.5	0.89	1	1	1	1
		±0	±0	±0	±0	±0	±0.2	±0.008	±0	±0.008	±0	±0	±0	±0
[13]	08-02/0.3-0.0	1	1	1	1	0.5	0.293	0.947	0.5	0.951	1	1	1	1
		±0	±0	±0	±0	±0	±0.116	±0.04	±0	±0.006	±0	±0	±0	±0
		1	1	1	1	0.5	0.557	0.947	0.5	0.951	1	1	1	1
[14]	08-02/0.3-0.3	±0	±0	±0	±0	±0	±0.197	±0.04	±0	±0.006	±0	±0	±0	±0
		1	1	1	1	0.5	0.612	0.946	0.5	0.951	1	1	1	1
		±0	±0	±0	±0	±0	±0.247	±0.041	±0	±0.006	±0	±0	±0	±0
[16]	08-02/0.7-0.0	1	1 ^{KW}	1	1	0.5	0.352	0.511	0.5	0.973	1	1	1	1
		±0	±0	±0	±0	±0	±0.167	±0.106	±0	±0.004	±0	±0	±0	±0
		1	1 ^{KW}	1	1	0.5	0.578	0.458	0.5	0.973	1	1	1	1
[17]	08-02/0.7-0.3	±0	±0	±0	±0	±0	±0.279	±0.262	±0	±0.004	±0	±0	±0	±0
		1	1 ^{KW}	1	1	0.5	0.407	0.388	0.5	0.973	1	1	1	1
		±0	±0	±0	±0	±0	±0.274	±0.27	±0	±0.004	±0	±0	±0	±0
[19]	20-80/0.0-0.0	1	1	1	1	0.5	0.177	0.897	0.5	0.897	1	1	1	1
		±0	±0	±0	±0	±0	±0.046	±0.006	±0	±0.006	±0	±0	±0	±0
		1	1	1	1	0.5	0.193	0.897	0.5	0.897	1	1	1	1
[20]	20-80/0.0-0.3	±0	±0	±0	±0	±0	±0.12	±0.006	±0	±0.006	±0	±0	±0	±0
		1	1	1	1	0.5	0.207	0.897	0.5	0.897	1	1	1	1
		±0	±0	±0	±0	±0	±0.132	±0.006	±0	±0.006	±0	±0	±0	±0
[22]	20-80/0.3-0.0	1	1	1	1	0.5	0.183	0.956	0.5	0.956	1	1	1	1
		±0	±0	±0	±0	±0	±0.065	±0.005	±0	±0.005	±0	±0	±0	±0
		1	1	1	1	0.5	0.2	0.956	0.5	0.956	1	1	1	1
[23]	20-80/0.3-0.3	±0	±0	±0	±0	±0	±0.118	±0.005	±0	±0.005	±0	±0	±0	±0
		1	1	1	1	0.5	0.2	0.956	0.5	0.956	1	1	1	1
		±0	±0	±0	±0	±0	±0.128	±0.005	±0	±0.005	±0	±0	±0	±0
[25]	20-80/0.7-0.0	1	1	1	1	0.5	0.197	0.913	0.5	0.976	1	1	1	1
		±0	±0	±0	±0	±0	±0.112	±0.163	±0	±0.003	±0	±0	±0	±0
		1	1	1	1	0.5	0.208	0.862	0.5	0.976	1	1	1	1
[26]	20-80/0.7-0.3	±0	±0	±0	±0	±0	±0.145	±0.203	±0	±0.003	±0	±0	±0	±0
		1	1	1	1	0.5	0.213	0.806	0.5	0.976	1	1	1	1
		±0	±0	±0	±0	±0	±0.155	±0.228	±0	±0.003	±0	±0	±0	±0
[28]	80-20/0.0-0.0	1	1	1	1	0.5	0.207	0.906	0.5	0.906	1	1	1	1
		±0	±0	±0	±0	±0	±0.109	±0.006	±0	±0.006	±0	±0	±0	±0
		1	1	1	1	0.5	0.23	0.906	0.5	0.906	1	1	1	1
[29]	80-20/0.0-0.3	±0	±0	±0	±0	±0	±0.169	±0.006	±0	±0.006	±0	±0	±0	±0
		1	1	1	1	0.5	0.18	0.906	0.5	0.906	1	1	1	1
		±0	±0	±0	±0	±0	±0.056	±0.006	±0	±0.006	±0	±0	±0	±0
[31]	80-20/0.3-0.0	1	1	1	1	0.5	0.188	0.963	0.5	0.963	1	1	1	1
		±0	±0	±0	±0	±0	±0.077	±0.003	±0	±0.003	±0	±0	±0	±0
		1	1	1	1	0.5	0.212	0.963	0.5	0.963	1	1	1	1
[32]	80-20/0.3-0.3	±0	±0	±0	±0	±0	±0.138	±0.003	±0	±0.003	±0	±0	±0	±0
		1	1	1	1	0.5	0.182	0.963	0.5	0.963	1	1	1	1
		±0	±0	±0	±0	±0	±0.063	±0.003	±0	±0.003	±0	±0	±0	±0
[34]	80-20/0.7-0.0	1	1	1	1	0.5	0.182	0.98	0.5	0.98	1	1	1	1
		±0	±0	±0	±0	±0	±0.079	±0.002	±0	±0.002	±0	±0	±0	±0
		1	1	1	1	0.5	0.205	0.98	0.5	0.98	1	1	1	1
[35]	80-20/0.7-0.3	±0	±0	±0	±0	±0	±0.132	±0.002	±0	±0.002	±0	±0	±0	±0
		1	1	1	1	0.5	0.185	0.98	0.5	0.98	1	1	1	1
		±0	±0	±0	±0	±0	±0.085	±0.002	±0	±0.002	±0	±0	±0	±0
[36]	80-20/0.7-0.7	1	1	1	1	0.5	0.185	0.98	0.5	0.98	1	1	1	1
		±0	±0	±0	±0	±0	±0.085	±0.002	±0	±0.002	±0	±0	±0	±0

Table 75 – Results of experiment A1 - Average and standard deviation of categorical attributes weights

Dataset		All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
[ID]	Scenario	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTLE	KPTBR
[1]	02-08/0.0-0.0	1	1	1	0.126	0.5	0.812	0.126	0.5	0.126	1	1	0.13	0.126
		±0	±0	±0	±0.015	±0	±0.074	±0.015	±0	±0.015	±0	±0	±0.013	±0.015
[2]	02-08/0.0-0.3	1	1	1	0.126	0.5	0.783	0.126	0.5	0.126	1	1	0.131	0.126
		±0	±0	±0	±0.015	±0	±0.133	±0.015	±0	±0.015	±0	±0	±0.013	±0.015
[3]	02-08/0.0-0.7	1	1	1	0.126	0.5	0.78	0.126	0.5	0.126	1	1	0.13	0.126
		±0	±0	±0	±0.015	±0	±0.127	±0.015	±0	±0.015	±0	±0	±0.013	±0.015
[4]	02-08/0.3-0.0	1	1	1	0.344	0.5	0.785	0.317	0.5	0.064	1	1	0.065	0.064
		±0	±0	±0	±0.212	±0	±0.121	±0.217	±0	±0.008	±0	±0	±0.008	±0.008
[5]	02-08/0.3-0.3	1	1	1	0.427	0.5	0.767	0.384	0.5	0.077	1	1	0.065	0.064
		±0	±0	±0	±0.168	±0	±0.157	±0.192	±0	±0.075	±0	±0	±0.008	±0.008
[6]	02-08/0.3-0.7	1	1 ^{py}	1	0.467	0.5	0.755	0.451	0.5	0.082	1	1	0.065	0.073
		±0	±0	±0	±0.141	±0	±0.161	±0.161	±0	±0.09	±0	±0	±0.008	±0.063
[7]	02-08/0.7-0.0	1	1	1	0.503	0.5	0.773	0.503	0.5	0.057	1	1	0.038	0.038
		±0	±0	±0	±0.05	±0	±0.141	±0.05	±0	±0.095	±0	±0	±0.007	±0.006
[8]	02-08/0.7-0.3	1	1	1	0.503	0.5	0.748	0.503	0.5	0.075	1	1	0.038	0.047
		±0	±0	±0	±0.05	±0	±0.181	±0.05	±0	±0.126	±0	±0	±0.007	±0.068
[9]	02-08/0.7-0.7	1	1 ^{py}	1	0.508	0.5	0.752	0.513	0.5	0.122	1	1	0.038	0.142
		±0	±0	±0	±0.069	±0	±0.163	±0.083	±0	±0.181	±0	±0	±0.007	±0.194
[10]	08-02/0.0-0.0	1	1	1	0.11	0.5	0.778	0.11	0.5	0.11	1	1	0.113	0.11
		±0	±0	±0	±0.008	±0	±0.109	±0.008	±0	±0.008	±0	±0	±0.008	±0.008
[11]	08-02/0.0-0.3	1	1	1	0.11	0.5	0.502	0.11	0.5	0.11	1	1	0.113	0.11
		±0	±0	±0	±0.008	±0	±0.197	±0.008	±0	±0.008	±0	±0	±0.008	±0.008
[12]	08-02/0.0-0.7	1	1	1	0.11	0.5	0.42	0.11	0.5	0.11	1	1	0.113	0.11
		±0	±0	±0	±0.008	±0	±0.2	±0.008	±0	±0.008	±0	±0	±0.008	±0.008
[13]	08-02/0.3-0.0	1	1	1	0.053	0.5	0.707	0.053	0.5	0.049	1	1	0.049	0.049
		±0	±0	±0	±0.04	±0	±0.116	±0.04	±0	±0.006	±0	±0	±0.006	±0.006
[14]	08-02/0.3-0.3	1	1	1	0.057	0.5	0.443	0.053	0.5	0.049	1	1	0.049	0.049
		±0	±0	±0	±0.057	±0	±0.197	±0.04	±0	±0.006	±0	±0	±0.006	±0.006
[15]	08-02/0.3-0.7	1	1	1	0.057	0.5	0.388	0.054	0.5	0.049	1	1	0.049	0.049
		±0	±0	±0	±0.057	±0	±0.247	±0.041	±0	±0.006	±0	±0	±0.006	±0.006
[16]	08-02/0.7-0.0	1	1 ^{KW}	1	0.497	0.5	0.648	0.489	0.5	0.027	1	1	0.028	0.451
		±0	±0	±0	±0.038	±0	±0.167	±0.106	±0	±0.004	±0	±0	±0.004	±0.159
[17]	08-02/0.7-0.3	1	1 ^{KW}	1	0.489	0.5	0.422	0.542	0.5	0.027	1	1	0.028	0.471
		±0	±0	±0	±0.163	±0	±0.279	±0.262	±0	±0.004	±0	±0	±0.004	±0.217
[18]	08-02/0.7-0.7	1	1 ^{KW}	1	0.59	0.5	0.593	0.612	0.5	0.027	1	1	0.028	0.5
		±0	±0	±0	±0.253	±0	±0.274	±0.27	±0	±0.004	±0	±0	±0.004	±0.24
[19]	20-80/0.0-0.0	1	1	1	0.103	0.5	0.823	0.103	0.5	0.103	1	1	0.104	0.103
		±0	±0	±0	±0.006	±0	±0.046	±0.006	±0	±0.006	±0	±0	±0.007	±0.006
[20]	20-80/0.0-0.3	1	1	1	0.103	0.5	0.807	0.103	0.5	0.103	1	1	0.104	0.103
		±0	±0	±0	±0.006	±0	±0.12	±0.006	±0	±0.006	±0	±0	±0.007	±0.006
[21]	20-80/0.0-0.7	1	1	1	0.103	0.5	0.793	0.103	0.5	0.103	1	1	0.104	0.103
		±0	±0	±0	±0.006	±0	±0.132	±0.006	±0	±0.006	±0	±0	±0.007	±0.006
[22]	20-80/0.3-0.0	1	1	1	0.044	0.5	0.817	0.044	0.5	0.044	1	1	0.044	0.044
		±0	±0	±0	±0.005	±0	±0.065	±0.005	±0	±0.005	±0	±0	±0.005	±0.005
[23]	20-80/0.3-0.3	1	1	1	0.044	0.5	0.8	0.044	0.5	0.044	1	1	0.044	0.044
		±0	±0	±0	±0.005	±0	±0.118	±0.005	±0	±0.005	±0	±0	±0.005	±0.005
[24]	20-80/0.3-0.7	1	1	1	0.044	0.5	0.8	0.044	0.5	0.044	1	1	0.044	0.044
		±0	±0	±0	±0.005	±0	±0.128	±0.005	±0	±0.005	±0	±0	±0.005	±0.005
[25]	20-80/0.7-0.0	1	1	1	0.087	0.5	0.803	0.087	0.5	0.024	1	1	0.024	0.024
		±0	±0	±0	±0.163	±0	±0.112	±0.163	±0	±0.003	±0	±0	±0.003	±0.003
[26]	20-80/0.7-0.3	1	1	1	0.148	0.5	0.792	0.138	0.5	0.024	1	1	0.024	0.024
		±0	±0	±0	±0.21	±0	±0.145	±0.203	±0	±0.003	±0	±0	±0.003	±0.003
[27]	20-80/0.7-0.7	1	1	1	0.199	0.5	0.787	0.194	0.5	0.024	1	1	0.024	0.024
		±0	±0	±0	±0.23	±0	±0.155	±0.228	±0	±0.003	±0	±0	±0.003	±0.003
[28]	80-20/0.0-0.0	1	1	1	0.094	0.5	0.793	0.094	0.5	0.094	1	1	0.094	0.094
		±0	±0	±0	±0.006	±0	±0.109	±0.006	±0	±0.006	±0	±0	±0.006	±0.006
[29]	80-20/0.0-0.3	1	1	1	0.094	0.5	0.77	0.094	0.5	0.094	1	1	0.094	0.094
		±0	±0	±0	±0.006	±0	±0.169	±0.006	±0	±0.006	±0	±0	±0.006	±0.006
[30]	80-20/0.0-0.7	1	1	1	0.094	0.5	0.82	0.094	0.5	0.094	1	1	0.094	0.094
		±0	±0	±0	±0.006	±0	±0.056	±0.006	±0	±0.006	±0	±0	±0.006	±0.006
[31]	80-20/0.3-0.0	1	1	1	0.037	0.5	0.812	0.037	0.5	0.037	1	1	0.038	0.037
		±0	±0	±0	±0.003	±0	±0.077	±0.003	±0	±0.003	±0	±0	±0.003	±0.003
[32]	80-20/0.3-0.3	1	1	1	0.037	0.5	0.788	0.037	0.5	0.037	1	1	0.038	0.037
		±0	±0	±0	±0.003	±0	±0.138	±0.003	±0	±0.003	±0	±0	±0.003	±0.003
[33]	80-20/0.3-0.7	1	1	1	0.037	0.5	0.818	0.037	0.5	0.037	1	1	0.038	0.037
		±0	±0	±0	±0.003	±0	±0.063	±0.003	±0	±0.003	±0	±0	±0.003	±0.003
[34]	80-20/0.7-0.0	1	1	1	0.02	0.5	0.818	0.02	0.5	0.02	1	1	0.02	0.02
		±0	±0	±0	±0.002	±0	±0.079	±0.002	±0	±0.002	±0	±0	±0.002	±0.002
[35]	80-20/0.7-0.3	1	1	1	0.02	0.5	0.795	0.02	0.5	0.02	1	1	0.02	0.02
		±0	±0	±0	±0.002	±0	±0.132	±0.002	±0	±0.002	±0	±0	±0.002	±0.002
[36]	80-20/0.7-0.7	1	1	1	0.02	0.5	0.815	0.02	0.5	0.02	1	1	0.02	0.02
		±0	±0	±0	±0.002	±0	±0.085	±0.002	±0	±0.002	±0	±0	±0.002	±0.002

Table 76 – Results of experiment A1 - Average and standard deviation of execution time

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)		Group 4 (Y)			
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTel	KPTBR
[1] 02-08/0.0-0.0	0.03	0.044	0.005	0.289	0.005	2.001	0.308	0.011	0.621	0.025	0.026	0.027	1.34
	±0.01	±0.014	±0.001	±0.026	±0.001	±0.072	±0.019	±0.013	±0.045	±0.005	±0.011	±0.005	±0.083
	0.032	0.043	0.005	0.299	0.005	2.036	0.318	0.012	0.545	0.025	0.026	0.027	1.34
[2] 02-08/0.0-0.3	±0.011	±0.011	±0.001	±0.028	±0.001	±0.105	±0.032	±0.014	±0.047	±0.005	±0.006	±0.012	±0.07
	0.033	0.042	0.005	0.285	0.006	2.01	0.306	0.011	0.559	0.026	0.026	0.028	1.36
	±0.013	±0.007	±0.001	±0.024	±0.001	±0.096	±0.021	±0.01	±0.044	±0.006	±0.006	±0.007	±0.079
[4] 02-08/0.3-0.0	0.031	0.04	0.004	0.278	0.005	2.014	0.312	0.01	0.544	0.025	0.026	0.027	1.452
	±0.013	±0.006	±0.001	±0.021	±0.001	±0.103	±0.023	±0.01	±0.044	±0.005	±0.005	±0.013	±0.114
	0.032	0.043	0.005	0.283	0.005	1.989	0.307	0.011	0.556	0.026	0.025	0.027	1.519
[5] 02-08/0.3-0.3	±0.013	±0.007	±0.002	±0.024	±0.001	±0.087	±0.019	±0.01	±0.048	±0.012	±0.005	±0.005	±0.135
	0.034	0.043^{Py}	0.005	0.286	0.005	2.021	0.31	0.011	0.566	0.027	0.028	0.03	1.647
	±0.012	±0.008	±0.001	±0.024	±0.001	±0.101	±0.02	±0.01	±0.045	±0.006	±0.007	±0.008	±0.185
[7] 02-08/0.7-0.0	0.036	0.04	0.004	0.281	0.005	2.006	0.31	0.01	0.543	0.026	0.025	0.029	1.743
	±0.018	±0.006	±0.001	±0.022	±0.001	±0.094	±0.021	±0.01	±0.047	±0.012	±0.005	±0.007	±0.242
	0.036	0.043	0.005	0.291	0.006	2.018	0.311	0.01	0.555	0.025	0.026	0.033	1.821
[8] 02-08/0.7-0.3	±0.019	±0.008	±0.002	±0.028	±0.001	±0.095	±0.02	±0.002	±0.044	±0.006	±0.012	±0.009	±0.234
	0.037	0.046^{Py}	0.004	0.285	0.006	2.021	0.315	0.011	0.575	0.031	0.03	0.038	2.06
	±0.013	±0.008	±0.001	±0.02	±0.002	±0.104	±0.019	±0.01	±0.047	±0.007	±0.008	±0.015	±0.25
[10] 08-02/0.0-0.0	0.03	0.029	0.002	0.145	0.003	1.503	0.186	0.017	0.812	0.025	0.026	0.027	1.215
	±0.012	±0.005	±0.001	±0.013	±0.001	±0.061	±0.016	±0.011	±0.045	±0.005	±0.007	±0.007	±0.065
	0.032	0.03	0.002	0.147	0.003	1.513	0.186	0.017	0.807	0.025	0.026	0.027	1.204
[11] 08-02/0.0-0.3	±0.012	±0.004	±0.001	±0.013	±0.002	±0.067	±0.018	±0.011	±0.043	±0.005	±0.008	±0.008	±0.059
	0.034	0.029	0.002	0.149	0.003	1.523	0.183	0.017	0.802	0.025	0.026	0.026	1.235
	±0.013	±0.005	±0.001	±0.013	±0.001	±0.071	±0.018	±0.012	±0.042	±0.005	±0.014	±0.007	±0.074
[12] 08-02/0.0-0.7	0.03	0.031	0.002	0.148	0.003	1.511	0.183	0.018	0.811	0.024	0.026	0.028	1.264
	±0.01	±0.006	±0.001	±0.015	±0.001	±0.058	±0.012	±0.016	±0.044	±0.006	±0.008	±0.009	±0.061
	0.036	0.031	0.002	0.15	0.004	1.516	0.185	0.018	0.818	0.026	0.027	0.03	1.322
[14] 08-02/0.3-0.3	±0.022	±0.005	±0.001	±0.012	±0.012	±0.066	±0.017	±0.012	±0.04	±0.006	±0.008	±0.01	±0.082
	0.043	0.032	0.002	0.15	0.003	1.533	0.188	0.017	0.813	0.027	0.03	0.028	1.334
	±0.012	±0.011	±0.001	±0.015	±0.002	±0.073	±0.019	±0.011	±0.045	±0.006	±0.016	±0.008	±0.076
[15] 08-02/0.3-0.7	0.03	0.032^{KW}	0.002	0.165	0.003	1.517	0.202	0.018	0.84	0.029	0.03	0.05	1.948
	±0.009	±0.005	±0.001	±0.019	±0.001	±0.073	±0.019	±0.015	±0.048	±0.008	±0.008	±0.015	±0.305
	0.03	0.034^{KW}	0.002	0.171	0.003	1.57	0.227	0.02	0.869	0.033	0.033	0.05	2.135
[17] 08-02/0.7-0.3	±0.009	±0.007	±0.001	±0.02	±0.001	±0.126	±0.039	±0.012	±0.072	±0.007	±0.009	±0.018	±0.267
	0.034	0.04^{KW}	0.003	0.19	0.004	1.666	0.222	0.019	0.863	0.037	0.034	0.058	2.267
	±0.015	±0.008	±0.001	±0.031	±0.003	±0.164	±0.022	±0.016	±0.049	±0.009	±0.01	±0.017	±0.19
[18] 08-02/0.7-0.7	0.086	0.191	0.041	2.629	0.046	16.637	2.863	0.042	2.028	0.17	0.176	0.173	8.357
	±0.026	±0.032	±0.019	±0.101	±0.005	±0.953	±0.216	±0.02	±0.137	±0.042	±0.048	±0.04	±0.524
	0.085	0.181	0.051	2.816	0.053	16.497	2.801	0.035	2.207	0.158	0.164	0.178	8.799
[20] 20-80/0.0-0.3	±0.027	±0.033	±0.026	±0.21	±0.026	±0.853	±0.199	±0.005	±0.189	±0.041	±0.043	±0.034	±0.707
	0.099	0.176	0.04	2.835	0.049	16.367	2.689	0.038	2.092	0.15	0.153	0.154	7.926
	±0.049	±0.032	±0.021	±0.271	±0.022	±0.742	±0.122	±0.015	±0.055	±0.039	±0.038	±0.031	±0.337
[21] 20-80/0.0-0.7	0.08	0.154	0.038	2.528	0.045	15.891	2.687	0.035	1.872	0.144	0.146	0.153	7.763
	±0.022	±0.025	±0.016	±0.114	±0.015	±0.312	±0.117	±0.019	±0.06	±0.03	±0.029	±0.031	±0.272
	0.081	0.153	0.037	2.535	0.045	16.034	2.679	0.036	2.02	0.153	0.153	0.159	7.816
[23] 20-80/0.3-0.3	±0.024	±0.025	±0.016	±0.107	±0.016	±0.403	±0.107	±0.015	±0.06	±0.031	±0.033	±0.036	±0.283
	0.082	0.156	0.037	2.558	0.046	16.109	2.681	0.04	2.111	0.155	0.155	0.16	7.775
	±0.024	±0.025	±0.016	±0.109	±0.019	±0.423	±0.121	±0.019	±0.065	±0.046	±0.036	±0.036	±0.262
[24] 20-80/0.3-0.7	0.079	0.157	0.037	2.512	0.047	15.867	2.635	0.036	1.873	0.143	0.147	0.152	8.046
	±0.02	±0.031	±0.018	±0.113	±0.019	±0.429	±0.119	±0.02	±0.064	±0.037	±0.031	±0.039	±0.417
	0.08	0.151	0.038	2.504	0.048	15.985	2.658	0.037	2.001	0.148	0.154	0.163	8.237
[26] 20-80/0.7-0.3	±0.021	±0.02	±0.018	±0.113	±0.02	±0.382	±0.102	±0.016	±0.058	±0.029	±0.035	±0.037	±0.441
	0.084	0.158	0.039	2.589	0.046	16.197	2.984	0.037	2.201	0.161	0.186	0.196	9.103
	±0.019	±0.024	±0.018	±0.124	±0.013	±0.844	±0.267	±0.006	±0.106	±0.042	±0.057	±0.055	±0.91
[27] 20-80/0.7-0.7	0.067	0.079	0.01	0.8	0.016	10.722	1.02	0.088	4.388	0.143	0.147	0.145	6.836
	±0.019	±0.028	±0.002	±0.064	±0.003	±0.303	±0.072	±0.024	±0.103	±0.031	±0.035	±0.031	±0.234
	0.068	0.078	0.011	0.794	0.017	10.586	1.023	0.09	4.42	0.152	0.153	0.151	7.019
[29] 80-20/0.0-0.3	±0.024	±0.027	±0.01	±0.063	±0.01	±0.265	±0.077	±0.027	±0.106	±0.033	±0.037	±0.034	±0.216
	0.069	0.091	0.012	0.816	0.017	10.76	1.055	0.089	4.641	0.151	0.152	0.154	7.022
	±0.02	±0.031	±0.01	±0.055	±0.01	±0.262	±0.069	±0.023	±0.116	±0.037	±0.034	±0.034	±0.24
[30] 80-20/0.0-0.7	0.068	0.088	0.012	0.817	0.017	10.715	1.043	0.091	4.492	0.142	0.149	0.145	6.945
	±0.02	±0.028	±0.01	±0.048	±0.01	±0.312	±0.072	±0.029	±0.085	±0.033	±0.037	±0.03	±0.203
	0.069	0.085	0.011	0.812	0.015	10.743	1.037	0.09	4.559	0.143	0.148	0.145	6.894
[32] 80-20/0.3-0.3	±0.02	±0.03	±0.009	±0.057	±0.002	±0.346	±0.072	±0.027	±0.108	±0.033	±0.032	±0.032	±0.227
	0.067	0.087	0.011	0.827	0.017	10.603	1.025	0.091	4.588	0.147	0.151	0.146	7.003
	±0.014	±0.031	±0.01	±0.071	±0.01	±0.303	±0.059	±0.025	±0.126	±0.037	±0.035	±0.034	±0.234
[33] 80-20/0.3-0.7	0.071	0.082	0.012	0.838	0.017	10.601	1.041	0.09	4.507	0.148	0.148	0.15	7.225
	±0.022	±0.027	±0.01	±0.07	±0.01	±0.269	±0.065	±0.027	±0.094	±0.036	±0.033	±0.037	±0.265
	0.071	0.087	0.01	0.823	0.018	10.6	1.031	0.085	4.558	0.139	0.149	0.157	7.184
[34] 80-20/0.7-0.3	±0.021	±0.028	±0.003	±0.063	±0.014	±0.244	±0.054	±0.012	±0.112	±0.029	±0.038	±0.039	±0.219
	0.073	0.085	0.011	0.835	0.017	10.708	1.034	0.091	4.568	0.14	0.147	0.159	7.372
	±0.024	±0.028	±0.01	±0.064	±0.01	±0.327	±0.06	±0.028	±0.075	±0.031	±0.035	±0.041	±0.27

Table 77 – Results of experiment A2 - ARPD values considering the ARI values of each repetition

ID	Dataset [ID] Scenario	Group 1 (K)				Group 2 (W)				Group 3 (P)				Group 4 (Y)						
		KAMLA	FMMMD	KMDBS	KMDBR	KAMLA	FMMMD	WKMBs	WKMMs	WKMBR	KAMLA	FMMMD	PMGBs	PMGBR	KAMLA	FMMMD	KPTBS	KPTNM	KPTBL	KPTBR
[1]	02-08/0.0-0.0	29.048	0.574	0	0	29.048	0.574	0	0	0	29.048	0.574	0	0	29.048	0.574	12.045	13.051	12.045	0
[2]	02-08/0.0-0.3	36.066	1.621	0	0	36.066	1.621	0	0	0	36.066	1.606	0.066	0.016	36.066	1.613	1.145	3.165	3.168	0
[3]	02-08/0.0-0.7	48.763	3.87	3.009	0	48.763	3.87	3.009	0	0	48.71	2.956	2.647	1.886	48.713	3.034	3.811	3.936	4.882	2.648
[4]	02-08/0.3-0.0	27.006	0.734	0	0	27.006	0.734	0	0	0	27.006	0.734	0	0	27.006	0.734	9.051	10.045	12.05	0
[5]	02-08/0.3-0.3	37.048	2.848	0	0	37.048	2.848	0	0	0	37.064	2.863	0.144	0.048	37.048	2.816	3.154	1.145	3.152	0.144
[6]	02-08/0.3-0.7	44.047	6.234	0.961	0	44.047	6.234	0.961	0	0	44.015	5.487	2.825	2.301	44.031	5.336	4.607	4.595	6.581	2.535
[7]	02-08/0.7-0.0	29.039	0.734	0	0	29.039	0.734	0	0	0	29.039	0.734	0	0	29.039	0.734	8.048	10.047	11.047	0
[8]	02-08/0.7-0.3	36.067	2.865	0	0	36.067	2.865	0	0	0	36.083	2.88	0.128	0.048	36.067	2.833	1.147	1.145	2.148	0.144
[9]	02-08/0.7-0.7	43.082	6.526	0.961	0	43.082	6.526	0.961	0	0	43.018	5.733	2.679	2.125	43.034	5.584	4.605	3.558	5.624	2.455
[10]	08-02/0.0-0.0	16.11	99.759	0.948	0.113	16.11	99.76	0.996	1.044	0.033	15.839	99.759	0.758	0.242	15.855	99.757	8.531	16.634	25.596	0.338
[11]	08-02/0.0-0.3	47.051	99.974	25.948	0	47.051	99.974	25.941	26.396	0	38.623	99.955	7.96	6.505	39.711	99.97	15.25	21.53	32.377	6.395
[12]	08-02/0.0-0.7	59.252	100.028	46.375	0	59.252	100.028	46.375	52.397	0	52.27	100.073	21.552	19.073	52.991	100.067	32.985	37.764	55.804	18.484
[13]	08-02/0.3-0.0	15.85	99.853	1.023	0.081	15.866	99.853	1.039	1.009	0.032	15.546	99.854	0.851	0.53	15.756	99.854	10.743	18.713	31.06	0.451
[14]	08-02/0.3-0.3	44.273	99.904	26.163	0	44.273	99.904	26.163	24.725	0	35.802	99.889	9.223	7.79	36.99	99.888	15.416	29.598	39.568	7.67
[15]	08-02/0.3-0.7	58.589	99.927	52.014	0	58.589	99.927	52.014	53.617	0	51.01	99.802	21.875	19.426	52.21	99.839	34.691	46.849	65.021	18.3
[16]	08-02/0.7-0.0	17.818	99.828	1.007	0.064	17.802	99.828	1.023	1.087	0.032	17.659	99.829	0.788	0.482	17.707	99.828	10.747	18.722	30.903	0.435
[17]	08-02/0.7-0.3	44.157	99.921	26.041	0	44.157	99.921	26.041	24.88	0	35.579	99.903	8.781	7.645	37.115	99.903	16.706	30.237	39.602	7.776
[18]	08-02/0.7-0.7	58.346	99.97	51.023	0	58.346	99.97	51.023	54.902	0	51.736	99.893	22.47	19.853	52.728	99.942	34.623	46.642	64.815	18.833
[19]	20-80/0.0-0.0	21	0.208	0	0	21	0.208	0	0	0	21	0.208	0	0	21	0.208	3.015	9.04	11.036	0
[20]	20-80/0.0-0.3	51.993	0.368	0	0	51.993	0.368	0	0	0	51.993	0.368	0	0	51.993	0.368	0	3.002	3.002	0
[21]	20-80/0.0-0.7	49.991	0.48	0	0	49.991	0.48	0	0	0	49.991	0.48	0	0	49.991	0.48	0	1	3.001	0
[22]	20-80/0.3-0.0	25.997	0.08	0	0	25.997	0.08	0	0	0	25.997	0.08	0	0	25.997	0.08	2.013	6.017	8.024	0
[23]	20-80/0.3-0.3	51.993	0.272	0	0	51.993	0.272	0	0	0	51.993	0.272	0	0	51.993	0.272	0	2.001	3.002	0
[24]	20-80/0.3-0.7	53.99	0.416	0	0	53.99	0.416	0	0	0	53.99	0.416	0	0	53.99	0.416	1.003	1.001	2.001	0
[25]	20-80/0.7-0.0	28.997	0.08	0	0	28.997	0.08	0	0	0	28.997	0.08	0	0	28.997	0.08	2.013	4.018	8.024	0
[26]	20-80/0.7-0.3	52.993	0.256	0	0	52.993	0.256	0	0	0	52.993	0.256	0	0	52.993	0.256	0	2.001	3.002	0
[27]	20-80/0.7-0.7	55.99	0.432	0	0	55.99	0.432	0	0	0	55.99	0.432	0	0	55.99	0.432	1.003	1.001	2.001	0
[28]	80-20/0.0-0.0	3	24.936	0	0	3	24.936	0	0	0	3	24.936	0	0	3	24.936	10.018	21.033	31.028	0
[29]	80-20/0.0-0.3	88	70.849	0	0	88	70.849	0	0	0	88	70.849	0	0	88	70.849	0	20.016	30.086	0
[30]	80-20/0.0-0.7	89	94.583	0	0	89	94.583	0	0	0	89	94.583	0	0	89	94.583	0	20.03	33.216	0
[31]	80-20/0.3-0.0	3	24.391	0	0	3	24.391	0	0	0	3	24.391	0	0	3	24.391	9.018	19.029	30.035	0
[32]	80-20/0.3-0.3	86	71.967	0	0	86	71.967	0	0	0	86	71.967	0	0	86	71.967	0	14.993	28.996	0
[33]	80-20/0.3-0.7	87.002	95.288	0	0	87.002	95.288	0	0	0	87.002	95.288	0	0	87.002	95.288	2.001	16.994	35.172	0
[34]	80-20/0.7-0.0	3	24.938	0	0	3	24.938	0	0	0	3	24.938	0	0	3	24.938	12.027	19.027	33.044	0
[35]	80-20/0.7-0.3	86	71.919	0	0	86	71.919	0	0	0	86	71.919	0	0	86	71.919	0	13.993	28.094	0
[36]	80-20/0.7-0.7	87	95.609	0	0	87	95.609	0	0	0	87	95.609	0	0	87	95.609	2.001	18	33.18	0

Table 78 – Results of experiment A2 - Average and standard deviation of numerical attributes weights

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)		Group 4 (Y)			
	KAMLA	FMMMD	KMDBS	KMDBR	WKMB	WKMM	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-08/0.0-0.0	1	1	1	1	0.5	0.333	0.69	0.5	0.978	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.409	±0	±0.009	±0	±0	±0	±0
[2] 02-08/0.0-0.3	1	1	1	1	0.5	0.333	0.748	0.5	0.928	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.374	±0	±0.214	±0	±0	±0	±0
[3] 02-08/0.0-0.7	1	1	1	1	0.5	0.34	0.651	0.5	0.654	1	1	1	1
	±0	±0	±0	±0	±0	±0.047	±0.422	±0	±0.452	±0	±0	±0	±0
[4] 02-08/0.3-0.0	1	1	1	1	0.5	0.333	0.81	0.5	0.979	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.3	±0	±0.011	±0	±0	±0	±0
[5] 02-08/0.3-0.3	1	1	1	1	0.5	0.333	0.861	0.5	0.916	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.222	±0	±0.233	±0	±0	±0	±0
[6] 02-08/0.3-0.7	1	1	1	1	0.5	0.347	0.821	0.5	0.681	1	1	1	1
	±0	±0	±0	±0	±0	±0.066	±0.281	±0	±0.438	±0	±0	±0	±0
[7] 02-08/0.7-0.0	1	1	1	1	0.5	0.333	0.801	0.5	0.979	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.31	±0	±0.011	±0	±0	±0	±0
[8] 02-08/0.7-0.3	1	1	1	1	0.5	0.333	0.862	0.5	0.925	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.222	±0	±0.214	±0	±0	±0	±0
[9] 02-08/0.7-0.7	1	1	1	1	0.5	0.347	0.804	0.5	0.702	1	1	1	1
	±0	±0	±0	±0	±0	±0.066	±0.303	±0	±0.429	±0	±0	±0	±0
[10] 08-02/0.0-0.0	1	1	1	1	0.5	0.463	0.238	0.5	0.121	1	1	1	1
	±0	±0	±0	±0	±0	±0.163	±0.355	±0	±0.268	±0	±0	±0	±0
[11] 08-02/0.0-0.3	1	1	1	1	0.5	0.457	0.223	0.5	0.143	1	1	1	1
	±0	±0	±0	±0	±0	±0.162	±0.275	±0	±0.282	±0	±0	±0	±0
[12] 08-02/0.0-0.7	1	1	1	1	0.5	0.443	0.217	0.5	0.127	1	1	1	1
	±0	±0	±0	±0	±0	±0.158	±0.284	±0	±0.26	±0	±0	±0	±0
[13] 08-02/0.3-0.0	1	1	1	1	0.5	0.477	0.298	0.5	0.177	1	1	1	1
	±0	±0	±0	±0	±0	±0.166	±0.354	±0	±0.33	±0	±0	±0	±0
[14] 08-02/0.3-0.3	1	1	1	1	0.5	0.48	0.23	0.5	0.176	1	1	1	1
	±0	±0	±0	±0	±0	±0.166	±0.277	±0	±0.306	±0	±0	±0	±0
[15] 08-02/0.3-0.7	1	1	1	1	0.5	0.433	0.258	0.5	0.188	1	1	1	1
	±0	±0	±0	±0	±0	±0.154	±0.305	±0	±0.311	±0	±0	±0	±0
[16] 08-02/0.7-0.0	1	1	1	1	0.5	0.473	0.289	0.5	0.185	1	1	1	1
	±0	±0	±0	±0	±0	±0.165	±0.343	±0	±0.325	±0	±0	±0	±0
[17] 08-02/0.7-0.3	1	1	1	1	0.5	0.49	0.235	0.5	0.205	1	1	1	1
	±0	±0	±0	±0	±0	±0.167	±0.282	±0	±0.323	±0	±0	±0	±0
[18] 08-02/0.7-0.7	1	1	1	1	0.5	0.447	0.235	0.5	0.173	1	1	1	1
	±0	±0	±0	±0	±0	±0.159	±0.301	±0	±0.301	±0	±0	±0	±0
[19] 20-80/0.0-0.0	1	1	1	1	0.5	0.333	0.91	0.5	0.984	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.231	±0	±0.01	±0	±0	±0	±0
[20] 20-80/0.0-0.3	1	1	1	1	0.5	0.333	0.9	0.5	0.978	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.248	±0	±0.01	±0	±0	±0	±0
[21] 20-80/0.0-0.7	1	1	1	1	0.5	0.333	0.891	0.5	0.962	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.264	±0	±0.098	±0	±0	±0	±0
[22] 20-80/0.3-0.0	1	1	1	1	0.5	0.333	0.931	0.5	0.982	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.165	±0	±0.013	±0	±0	±0	±0
[23] 20-80/0.3-0.3	1	1	1	1	0.5	0.333	0.892	0.5	0.974	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.246	±0	±0.014	±0	±0	±0	±0
[24] 20-80/0.3-0.7	1	1	1	1	0.5	0.333	0.892	0.5	0.966	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.246	±0	±0.014	±0	±0	±0	±0
[25] 20-80/0.7-0.0	1	1	1	1	0.5	0.333	0.931	0.5	0.982	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.165	±0	±0.013	±0	±0	±0	±0
[26] 20-80/0.7-0.3	1	1	1	1	0.5	0.333	0.893	0.5	0.974	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.246	±0	±0.014	±0	±0	±0	±0
[27] 20-80/0.7-0.7	1	1	1	1	0.5	0.333	0.883	0.5	0.965	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.262	±0	±0.014	±0	±0	±0	±0
[28] 80-20/0.0-0.0	1	1	1	1	0.5	0.333	0.392	0.5	0.003	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.48	±0	±0.004	±0	±0	±0	±0
[29] 80-20/0.0-0.3	1	1	1	1	0.5	0.333	0.275	0.5	0.002	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.44	±0	±0.004	±0	±0	±0	±0
[30] 80-20/0.0-0.7	1	1	1	1	0.5	0.337	0.129	0.5	0.007	1	1	1	1
	±0	±0	±0	±0	±0	±0.033	±0.329	±0	±0.044	±0	±0	±0	±0
[31] 80-20/0.3-0.0	1	1	1	1	0.5	0.333	0.555	0.5	0.002	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.483	±0	±0.004	±0	±0	±0	±0
[32] 80-20/0.3-0.3	1	1	1	1	0.5	0.333	0.487	0.5	0.002	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.486	±0	±0.004	±0	±0	±0	±0
[33] 80-20/0.3-0.7	1	1	1	1	0.5	0.333	0.36	0.5	0.002	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.469	±0	±0.004	±0	±0	±0	±0
[34] 80-20/0.7-0.0	1	1	1	1	0.5	0.333	0.546	0.5	0.003	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.484	±0	±0.004	±0	±0	±0	±0
[35] 80-20/0.7-0.3	1	1	1	1	0.5	0.333	0.497	0.5	0.003	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.487	±0	±0.004	±0	±0	±0	±0
[36] 80-20/0.7-0.7	1	1	1	1	0.5	0.333	0.38	0.5	0.003	1	1	1	1
	±0	±0	±0	±0	±0	±0	±0.474	±0	±0.004	±0	±0	±0	±0

Table 79 – Results of experiment A2 - Average and standard deviation of categorical attributes weights

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)		Group 4 (Y)		
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBS	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTBL
[1] 02-08/0.0-0.0	1	1	1	0.338	0.5	0.667	0.31	0.5	0.022	1	1	0.02
	±0	±0	±0	±0.423	±0	±0	±0.409	±0	±0.009	±0	±0	±0.009
	1	1	1	0.28	0.5	0.667	0.252	0.5	0.072	1	1	0.02
[2] 02-08/0.0-0.3	±0	±0	±0	±0.394	±0	±0	±0.374	±0	±0.214	±0	±0	±0.009
	1	1	1	0.402	0.5	0.66	0.349	0.5	0.346	1	1	0.02
[3] 02-08/0.0-0.7	±0	±0	±0	±0.448	±0	±0.047	±0.422	±0	±0.452	±0	±0	±0.009
	1	1	1	0.218	0.5	0.667	0.19	0.5	0.021	1	1	0.019
[4] 02-08/0.3-0.0	±0	±0	±0	±0.329	±0	±0	±0.3	±0	±0.011	±0	±0	±0.009
	1	1	1	0.162	0.5	0.667	0.139	0.5	0.084	1	1	0.019
[5] 02-08/0.3-0.3	±0	±0	±0	±0.265	±0	±0	±0.222	±0	± 0.233	±0	±0	±0.009
	1	1	1	0.209	0.5	0.653	0.179	0.5	0.319	1	1	0.019
[6] 02-08/0.3-0.7	±0	±0	±0	±0.32	±0	±0.066	±0.281	±0	±0.438	±0	±0	±0.009
	1	1	1	0.226	0.5	0.667	0.199	0.5	0.021	1	1	0.019
[7] 02-08/0.7-0.0	±0	±0	±0	±0.338	±0	±0	±0.31	±0	±0.011	±0	±0	±0.009
	1	1	1	0.161	0.5	0.667	0.138	0.5	0.075	1	1	0.019
[8] 02-08/0.7-0.3	±0	±0	±0	±0.265	±0	±0	±0.222	±0	±0.214	±0	±0	±0.009
	1	1	1	0.217	0.5	0.653	0.196	0.5	0.298	1	1	0.019
[9] 02-08/0.7-0.7	±0	±0	±0	±0.329	±0	±0.066	±0.303	±0	±0.429	±0	±0	±0.009
	1	1	1	0.802	0.5	0.537	0.762	0.5	0.879	1	1	0.009
[10] 08-02/0.0-0.0	±0	±0	±0	± 0.318	±0	±0.163	± 0.355	±0	± 0.268	±0	±0	±0.004
	1	1	1	0.807	0.5	0.543	0.777	0.5	0.857	1	1	0.009
[11] 08-02/0.0-0.3	±0	±0	±0	± 0.273	±0	±0.162	± 0.275	±0	±0.282	±0	±0	±0.004
	1	1	1	0.816	0.5	0.557	0.783	0.5	0.873	1	1	0.009
[12] 08-02/0.0-0.7	±0	±0	±0	± 0.284	±0	±0.158	± 0.284	±0	± 0.26	±0	±0	±0.004
	1	1	1	0.79	0.5	0.523	0.702	0.5	0.823	1	1	0.009
[13] 08-02/0.3-0.0	±0	±0	±0	± 0.317	±0	±0.166	± 0.354	±0	±0.33	±0	±0	±0.004
	1	1	1	0.804	0.5	0.52	0.77	0.5	0.824	1	1	0.009
[14] 08-02/0.3-0.3	±0	±0	±0	± 0.252	±0	±0.166	± 0.277	±0	± 0.306	±0	±0	±0.004
	1	1	1	0.8	0.5	0.567	0.742	0.5	0.812	1	1	0.009
[15] 08-02/0.3-0.7	±0	±0	±0	± 0.275	±0	±0.154	± 0.305	±0	± 0.311	±0	±0	±0.004
	1	1	1	0.812	0.5	0.527	0.711	0.5	0.815	1	1	0.009
[16] 08-02/0.7-0.0	±0	±0	±0	± 0.308	±0	±0.165	± 0.343	±0	±0.325	±0	±0	±0.004
	1	1	1	0.796	0.5	0.51	0.765	0.5	0.795	1	1	0.009
[17] 08-02/0.7-0.3	±0	±0	±0	± 0.255	±0	±0.167	± 0.282	±0	±0.323	±0	±0	±0.004
	1	1	1	0.797	0.5	0.553	0.765	0.5	0.827	1	1	0.009
[18] 08-02/0.7-0.7	±0	±0	±0	± 0.276	±0	±0.159	± 0.301	±0	± 0.301	±0	±0	±0.004
	1	1	1	0.09	0.5	0.667	0.09	0.5	0.016	1	1	0.005
[19] 20-80/0.0-0.0	±0	±0	±0	±0.231	±0	±0	±0.231	±0	±0.01	±0	±0	±0.003
	1	1	1	0.109	0.5	0.667	0.1	0.5	0.022	1	1	0.005
[20] 20-80/0.0-0.3	±0	±0	±0	±0.264	±0	±0	±0.248	±0	±0.01	±0	±0	±0.003
	1	1	1	0.119	0.5	0.667	0.109	0.5	0.038	1	1	0.005
[21] 20-80/0.0-0.7	±0	±0	±0	±0.278	±0	±0	±0.264	±0	±0.098	±0	±0	±0.003
	1	1	1	0.069	0.5	0.667	0.069	0.5	0.018	1	1	0.005
[22] 20-80/0.3-0.0	±0	±0	±0	±0.165	±0	±0	±0.165	±0	±0.013	±0	±0	±0.003
	1	1	1	0.108	0.5	0.667	0.108	0.5	0.026	1	1	0.005
[23] 20-80/0.3-0.3	±0	±0	±0	±0.246	±0	±0	±0.246	±0	±0.014	±0	±0	±0.003
	1	1	1	0.108	0.5	0.667	0.108	0.5	0.034	1	1	0.005
[24] 20-80/0.3-0.7	±0	±0	±0	±0.246	±0	±0	±0.246	±0	±0.014	±0	±0	±0.003
	1	1	1	0.069	0.5	0.667	0.069	0.5	0.018	1	1	0.005
[25] 20-80/0.7-0.0	±0	±0	±0	±0.165	±0	±0	±0.165	±0	±0.013	±0	±0	±0.003
	1	1	1	0.107	0.5	0.667	0.107	0.5	0.026	1	1	0.005
[26] 20-80/0.7-0.3	±0	±0	±0	±0.246	±0	±0	±0.246	±0	±0.014	±0	±0	±0.003
	1	1	1	0.117	0.5	0.667	0.117	0.5	0.035	1	1	0.005
[27] 20-80/0.7-0.7	±0	±0	±0	±0.262	±0	±0	±0.262	±0	±0.014	±0	±0	±0.003
	1	1	1	0.617	0.5	0.667	0.608	0.5	0.998	1	1	0.003
[28] 80-20/0.0-0.0	±0	±0	±0	±0.478	±0	±0	±0.48	±0	±0.004	±0	±0	±0.001
	1	1	1	0.745	0.5	0.667	0.725	0.5	0.998	1	1	0.003
[29] 80-20/0.0-0.3	±0	±0	±0	±0.429	±0	±0	±0.44	±0	±0.004	±0	±0	±0.001
	1	1	1	0.9	0.5	0.663	0.871	0.5	0.993	1	1	0.003
[30] 80-20/0.0-0.7	±0	±0	±0	±0.293	±0	±0.033	±0.329	±0	±0.044	±0	±0	±0.001
	1	1	1	0.464	0.5	0.667	0.445	0.5	0.998	1	1	0.003
[31] 80-20/0.3-0.0	±0	±0	±0	±0.485	±0	±0	±0.483	±0	±0.004	±0	±0	±0.001
	1	1	1	0.533	0.5	0.667	0.513	0.5	0.998	1	1	0.003
[32] 80-20/0.3-0.3	±0	±0	±0	±0.486	±0	±0	±0.486	±0	±0.004	±0	±0	±0.001
	1	1	1	0.63	0.5	0.667	0.64	0.5	0.998	1	1	0.003
[33] 80-20/0.3-0.7	±0	±0	±0	±0.472	±0	±0	±0.469	±0	±0.004	±0	±0	±0.001
	1	1	1	0.464	0.5	0.667	0.454	0.5	0.997	1	1	0.003
[34] 80-20/0.7-0.0	±0	±0	±0	±0.485	±0	±0	±0.484	±0	±0.004	±0	±0	±0.001
	1	1	1	0.523	0.5	0.667	0.503	0.5	0.997	1	1	0.003
[35] 80-20/0.7-0.3	±0	±0	±0	±0.487	±0	±0	±0.487	±0	±0.004	±0	±0	±0.001
	1	1	1	0.649	0.5	0.667	0.62	0.5	0.997	1	1	0.003
[36] 80-20/0.7-0.7	±0	±0	±0	±0.467	±0	±0	±0.474	±0	±0.004	±0	±0	±0.001

Table 80 – Results of experiment A2 - Average and standard deviation of execution time

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTBL	KPTBR
[1] 02-08/0.0-0.0	0.039 ±0.011 0.045	0.059 ±0.017 0.073	0.005 ±0.002 0.005	0.305 ±0.035 0.306	0.006 ±0.002 0.007	1.186 ±0.232 1.008	0.348 ±0.04 0.326	0.011 ±0.014 0.012	0.62 ±0.05 0.555	0.031 ±0.02 0.027	0.032 ±0.018 0.029	0.037 ±0.024 0.032	1.918 ±0.461 1.843
[2] 02-08/0.0-0.3	±0.023 0.04 ±0.009	±0.025 0.08 ±0.029	±0.001 0.005 ±0.002	±0.033 0.292 ±0.024	±0.004 0.006 ±0.001	±0.045 1.03 ±0.101	±0.025 0.323 ±0.024	±0.013 0.011 ±0.01	±0.048 0.574 ±0.045	±0.008 0.032 ±0.007	±0.008 0.036 ±0.008	±0.009 0.038 ±0.014	±0.289 2.179 ±0.306
[3] 02-08/0.0-0.7	0.037 ±0.014 0.034	0.05 ±0.009 0.067	0.004 ±0.001 0.004	0.292 ±0.026 0.288	0.005 ±0.001 0.005	1.016 ±0.057 1.02	0.328 ±0.027 0.324	0.01 ±0.01 0.01	0.537 ±0.053 0.549	0.03 ±0.016 0.028	0.031 ±0.017 0.028	0.035 ±0.016 0.032	1.862 ±0.309 1.802
[4] 02-08/0.3-0.0	±0.009 0.039 ±0.015	±0.027 0.086 ±0.042	±0.001 0.004 ±0.001	±0.027 0.299 ±0.028	±0.001 0.005 ±0.001	±0.056 1.017 ±0.059	±0.03 0.324 ±0.022	±0.01 0.011 ±0.01	±0.051 0.575 ±0.048	±0.008 0.033 ±0.008	±0.008 0.036 ±0.01	±0.009 0.041 ±0.015	±0.299 2.163 ±0.276
[5] 02-08/0.3-0.3	0.038 ±0.019 0.038	0.052 ±0.009 0.065	0.004 ±0.001 0.005	0.289 ±0.031 0.292	0.006 ±0.001 0.005	1.014 ±0.051 1.017	0.344 ±0.033 0.32	0.01 ±0.01 0.01	0.54 ±0.048 0.555	0.03 ±0.015 0.027	0.031 ±0.019 0.028	0.034 ±0.021 0.032	1.843 ±0.323 1.778
[6] 02-08/0.7-0.0	±0.01 0.04 ±0.021	±0.031 0.09 ±0.042	±0.001 0.004 ±0.001	±0.03 0.294 ±0.025	±0.001 0.005 ±0.001	±0.059 1 ±0.046	±0.025 0.322 ±0.026	±0.01 0.01 ±0.002	±0.052 0.576 ±0.048	±0.007 0.033 ±0.008	±0.009 0.036 ±0.014	±0.013 0.041 ±0.013	±0.264 2.141 ±0.267
[7] 02-08/0.7-0.3	0.03 ±0.011 0.03	0.08 ±0.029 0.078	0.002 ±0.001 0.002	0.172 ±0.023 0.176	0.003 ±0.001 0.003	0.77 ±0.038 0.78	0.214 ±0.026 0.209	0.016 ±0.011 0.019	0.815 ±0.049 0.808	0.029 ±0.009 0.042	0.038 ±0.018 0.051	0.046 ±0.018 0.056	2.042 ±0.292 2.495
[8] 02-08/0.7-0.7	±0.007 0.032 ±0.013	±0.029 0.079 ±0.031	±0.001 0.002 ±0.001	±0.022 0.185 ±0.032	±0.002 0.003 ±0.002	±0.044 0.775 ±0.043	±0.024 0.216 ±0.031	±0.016 0.017 ±0.011	±0.047 0.831 ±0.056	±0.011 0.052 ±0.018	±0.011 0.068 ±0.026	±0.02 0.061 ±0.023	±0.019 2.851 ±0.285
[9] 08-02/0.0-0.0	0.03 ±0.006 0.03	0.078 ±0.024 0.076	0.002 ±0.001 0.002	0.176 ±0.028 0.172	0.003 ±0.001 0.003	0.78 ±0.044 0.778	0.209 ±0.028 0.217	0.019 ±0.015 0.017	0.808 ±0.05 0.817	0.042 ±0.019 0.034	0.051 ±0.018 0.042	0.056 ±0.024 0.047	2.495 ±0.286 2.044
[10] 08-02/0.3-0.0	±0.008 0.03 ±0.008	±0.023 0.076 ±0.023	±0.001 0.002 ±0.001	±0.025 0.181 ±0.025	±0.002 0.003 ±0.002	±0.044 0.785 ±0.044	±0.024 0.213 ±0.024	±0.015 0.017 ±0.015	±0.046 0.821 ±0.015	±0.016 0.044 ±0.015	±0.015 0.049 ±0.015	±0.026 0.06 ±0.026	±0.244 2.461 ±0.244
[11] 08-02/0.3-0.3	0.031 ±0.012 0.031	0.074 ±0.023 0.074	0.002 ±0.001 0.002	0.181 ±0.025 0.181	0.003 ±0.001 0.003	0.776 ±0.044 0.776	0.218 ±0.024 0.218	0.017 ±0.015 0.017	0.838 ±0.046 0.838	0.055 ±0.016 0.055	0.066 ±0.015 0.066	0.058 ±0.026 0.058	2.877 ±0.244 2.877
[12] 08-02/0.3-0.7	±0.012 0.03 ±0.008	±0.021 0.078 ±0.023	±0.001 0.002 ±0.001	±0.026 0.171 ±0.025	±0.002 0.003 ±0.001	±0.043 0.774 ±0.034	±0.028 0.215 ±0.025	±0.011 0.017 ±0.016	±0.053 0.838 ±0.016	±0.02 0.033 ±0.016	±0.028 0.039 ±0.015	±0.024 0.057 ±0.026	±0.286 2.048 ±0.279
[13] 08-02/0.7-0.0	±0.008 0.031 ±0.012	±0.023 0.076 ±0.023	±0.001 0.002 ±0.001	±0.025 0.183 ±0.027	±0.001 0.003 ±0.001	±0.034 0.766 ±0.04	±0.025 0.222 ±0.026	±0.016 0.016 ±0.011	±0.047 0.826 ±0.05	±0.019 0.043 ±0.013	±0.014 0.05 ±0.018	±0.02 0.057 ±0.024	±0.279 2.494 ±0.284
[14] 08-02/0.7-0.3	0.031 ±0.012 0.031	0.08 ±0.023 0.08	0.002 ±0.001 0.002	0.184 ±0.03 0.184	0.003 ±0.001 0.003	0.786 ±0.047 0.786	0.221 ±0.03 0.221	0.017 ±0.011 0.017	0.838 ±0.053 0.838	0.058 ±0.026 0.058	0.065 ±0.027 0.065	0.058 ±0.024 0.058	2.865 ±0.279 2.865
[15] 20-80/0.0-0.0	0.08 ±0.023 0.083	0.167 ±0.024 0.194	0.038 ±0.013 0.038	2.58 ±0.133 2.63	0.047 ±0.018 0.046	8.597 ±0.689 8.141	3.096 ±0.793 2.707	0.042 ±0.018 0.037	1.994 ±0.154 2.048	0.163 ±0.079 0.155	0.162 ±0.058 0.155	0.185 ±0.087 0.166	8.977 ±1.033 8.489
[16] 20-80/0.0-0.3	±0.025 0.083 ±0.022	±0.029 0.194 ±0.036	±0.019 0.038 ±0.009	±0.102 2.66 ±0.101	±0.02 0.049 ±0.02	±0.175 8.212 ±0.359	±0.123 2.694 ±0.113	±0.016 0.038 ±0.015	±0.053 2.147 ±0.069	±0.033 0.155 ±0.042	±0.037 0.163 ±0.043	±0.044 0.172 ±0.047	±0.69 8.787 ±0.719
[17] 20-80/0.0-0.7	0.081 ±0.024 0.081	0.201 ±0.034 0.177	0.038 ±0.013 0.039	2.64 ±0.108 2.525	0.049 ±0.018 0.048	7.981 ±0.332 7.954	2.77 ±0.153 2.661	0.034 ±0.015 0.036	1.906 ±0.092 2.051	0.16 ±0.075 0.151	0.157 ±0.055 0.155	0.171 ±0.078 0.165	8.67 ±1.139 8.212
[18] 20-80/0.3-0.0	±0.022 0.083 ±0.022	±0.029 0.167 ±0.027	±0.015 0.038 ±0.018	±0.102 2.563 ±0.123	±0.021 0.044 ±0.013	±0.023 7.923 ±0.256	±0.109 2.742 ±0.131	±0.011 0.038 ±0.011	±0.067 2.137 ±0.071	±0.032 0.155 ±0.038	±0.038 0.158 ±0.054	±0.039 0.171 ±0.044	±0.396 8.462 ±0.545
[19] 20-80/0.3-0.3	0.081 ±0.022 0.081	0.169 ±0.027 0.169	0.039 ±0.018 0.039	2.532 ±0.123 2.532	0.045 ±0.013 0.045	7.837 ±0.253 7.837	2.736 ±0.131 2.736	0.032 ±0.011 0.032	1.887 ±0.071 1.887	0.154 ±0.038 0.154	0.162 ±0.054 0.162	0.166 ±0.044 0.166	8.677 ±1.187 8.677
[20] 20-80/0.7-0.0	±0.02 0.078 ±0.016	±0.027 0.179 ±0.03	±0.02 0.039 ±0.015	±0.17 2.524 ±0.118	±0.018 0.046 ±0.019	±0.253 7.859 ±0.246	±0.153 2.598 ±0.073	±0.011 0.035 ±0.01	±0.115 2.017 ±0.062	±0.059 0.149 ±0.031	±0.054 0.153 ±0.041	±0.066 0.162 ±0.034	±1.187 8.052 ±0.338
[21] 20-80/0.7-0.3	0.08 ±0.021 0.062	0.181 ±0.023 0.267	0.038 ±0.017 0.011	2.526 ±0.125 0.835	0.046 ±0.018 0.016	7.747 ±0.192 5.334	2.635 ±0.102 1.133	0.038 ±0.014 0.094	2.124 ±0.068 4.535	0.155 ±0.044 0.155	0.153 ±0.036 0.157	0.169 ±0.045 0.174	8.445 ±0.566 8.023
[22] 80-20/0.0-0.0	±0.012 0.062 ±0.017	±0.08 0.261 ±0.079	±0.01 0.01 ±0.002	±0.071 0.829 ±0.074	±0.003 0.016 ±0.01	±0.189 5.314 ±0.207	±0.108 1.053 ±0.082	±0.028 0.091 ±0.022	±0.09 4.588 ±0.088	±0.09 0.153 ±0.034	±0.04 0.16 ±0.038	±0.047 0.184 ±0.049	±0.5 7.891 ±0.421
[23] 80-20/0.0-0.3	0.065 ±0.019 0.06	0.219 ±0.057 0.266	0.011 ±0.009 0.011	0.846 ±0.083 0.812	0.015 ±0.002 0.016	5.307 ±0.197 5.212	1.078 ±0.086 1.096	0.093 ±0.026 0.09	4.636 ±0.099 4.501	0.153 ±0.037 0.15	0.171 ±0.047 0.164	0.192 ±0.052 0.178	8.023 ±0.418 7.885
[24] 80-20/0.3-0.0	±0.015 0.062 ±0.017	±0.089 0.243 ±0.074	±0.009 0.011 ±0.009	±0.072 0.831 ±0.077	±0.009 0.015 ±0.002	±0.171 5.191 ±0.153	±0.09 1.034 ±0.069	±0.024 0.091 ±0.026	±0.103 4.547 ±0.089	±0.048 0.146 ±0.033	±0.052 0.157 ±0.033	±0.052 0.176 ±0.039	±0.589 7.67 ±0.387
[25] 80-20/0.3-0.3	0.064 ±0.018 0.06	0.212 ±0.068 0.259	0.01 ±0.002 0.009	0.848 ±0.086 0.803	0.015 ±0.002 0.017	5.247 ±0.18 5.193	1.05 ±0.085 1.093	0.096 ±0.029 0.091	4.578 ±0.113 4.49	0.147 ±0.039 0.156	0.164 ±0.041 0.159	0.189 ±0.054 0.179	7.918 ±0.467 7.973
[26] 80-20/0.7-0.0	±0.018 0.061 ±0.016	±0.087 0.263 ±0.087	±0.002 0.01 ±0.002	±0.064 0.834 ±0.07	±0.013 0.015 ±0.002	±0.147 5.281 ±0.213	±0.088 1.036 ±0.078	±0.026 0.089 ±0.022	±0.078 4.534 ±0.091	±0.06 0.145 ±0.03	±0.042 0.158 ±0.033	±0.055 0.18 ±0.044	±0.566 7.67 ±0.411
[27] 80-20/0.7-0.3	0.063 ±0.017 0.063	0.212 ±0.06 0.212	0.011 ±0.009 0.011	0.837 ±0.076 0.837	0.016 ±0.01 0.016	5.205 ±0.193 5.205	1.056 ±0.086 1.056	0.09 ±0.021 0.09	4.591 ±0.103 4.591	0.144 ±0.032 0.144	0.164 ±0.04 0.164	0.19 ±0.05 0.19	7.981 ±0.538 7.981
[28] 80-20/0.7-0.7	±0.017 0.063 ±0.017	±0.06 0.212 ±0.06	±0.009 0.011 ±0.009	±0.076 0.837 ±0.076	±0.01 0.016 ±0.01	±0.193 5.205 ±0.193	±0.086 1.056 ±0.086	±0.021 0.09 ±0.021	±0.103 4.591 ±0.103	±0.032 0.144 ±0.032	±0.04 0.164 ±0.04	±0.05 0.19 ±0.05	±0.538 7.981 ±0.538

Table 81 – Results of experiment B1 - ARPD values considering the ARI values of each repetition

Dataset [ID] Scenario	Group 1 (K)				Group 2 (W)				Group 3 (P)				Group 4 (Y)						
	KAMLA	FMMMD	KMDBS	KMDBR	KAMLA	FMMMD	WKMDBS	WKMDBR	KAMLA	FMMMD	PMGBS	PMGBR	KAMLA	FMMMD	KPTBS	KPTNM	KPTLE	KPTBR	
[1] 02-02/0.0-0.0	0.016	0	0.064	0	0.016	0	0.064	0.064	0	0.016	0	0.064	0	0.016	0	0.064	0.016	0	0
[2] 02-02/0.0-0.3	0.016	0.016	9.726	0	0.016	0.016	9.726	0.016	0	0	0	0.256	0.016	0	0	0.256	0.176	0.016	0.016
[3] 02-02/0.0-0.7	0.016	0	32.8	0	0.016	0	32.8	0	0	0.068	0.064	0.068	0.064	0.068	0.068	0.18	0.064	0.064	0.064
[4] 02-02/0.3-0.0	0.592	0	0.208	0.128	0.608	0.016	0.224	0.192	0.096	0.608	0.016	0.08	0.016	0.068	0.016	0.208	0.096	0.272	0.144
[5] 02-02/0.3-0.3	1.279	0	15.723	0.544	1.279	0	15.723	0.801	0.688	1.279	0	0.8	0.56	1.279	0	1.759	0.897	1.741	1.551
[6] 02-02/0.3-0.7	1.709	0	32.785	1.342	1.709	0	32.785	1.072	0.897	1.709	0	1.787	1.166	1.709	0	4.352	2.046	4.085	3.85
[7] 02-02/0.7-0.0	1.871	0	1.343	0.943	1.887	0.016	1.358	1.119	0.688	1.903	0.032	0.112	0.048	1.887	0.016	1.135	0.128	9.292	1.103
[8] 02-02/0.7-0.3	3.596	0.016	21.172	1.725	3.596	0.016	21.172	10.159	3.321	3.596	0.048	0.088	0.048	3.596	0.016	14.854	26.495	14.705	26.495
[9] 02-02/0.7-0.7	3.873	0	36.654	1.788	3.873	0	36.654	15.679	1.788	3.904	0.032	2.025	1.39	3.889	0.016	36.553	3.575	53.85	35.565
[10] 20-20/0.0-0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[11] 20-20/0.0-0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[12] 20-20/0.0-0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[13] 20-20/0.3-0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[14] 20-20/0.3-0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[15] 20-20/0.3-0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.996	0	0	0
[16] 20-20/0.7-0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[17] 20-20/0.7-0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[18] 20-20/0.7-0.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.996	0	0	0

Table 82 – Results of experiment B1 - Average and standard deviation of numerical attributes weights

Dataset	All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
[ID] Scenario	KAMLA	FMMMD	KMDBS	KMDBR	WKMBS	WKMS	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-02/0.0-0.0	1	1	1	1	0.5	0.282	0.874	0.5	0.874	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.118	± 0.015	± 0	± 0.015	± 0	± 0	± 0	± 0
[2] 02-02/0.0-0.3	1	1	1	1	0.5	0.605	0.871	0.5	0.874	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.218	± 0.037	± 0	± 0.015	± 0	± 0	± 0	± 0
[3] 02-02/0.0-0.7	1	1	1	1	0.5	0.628	0.871	0.5	0.874	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.273	± 0.037	± 0	± 0.015	± 0	± 0	± 0	± 0
[4] 02-02/0.3-0.0	1	1 ^K	1	1	0.5	0.338	0.512	0.5	0.798	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.176	± 0.114	± 0	± 0.206	± 0	± 0	± 0	± 0
[5] 02-02/0.3-0.3	1	1	1	1	0.5	0.63	0.576	0.5	0.824	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.235	± 0.276	± 0	± 0.194	± 0	± 0	± 0	± 0
[6] 02-02/0.3-0.7	1	1	1	1	0.5	0.615	0.561	0.5	0.851	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.292	± 0.284	± 0	± 0.177	± 0	± 0	± 0	± 0
[7] 02-02/0.7-0.0	1	1 ^{KWY}	1	1	0.5	0.38	0.472	0.5	0.699	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.217	± 0.222	± 0	± 0.233	± 0	± 0	± 0	± 0
[8] 02-02/0.7-0.3	1	1	1	1	0.5	0.815	0.345	0.5	0.745	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.097	± 0.259	± 0	± 0.232	± 0	± 0	± 0	± 0
[9] 02-02/0.7-0.7	1	1	1	1	0.5	0.74	0.373	0.5	0.792	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.22	± 0.234	± 0	± 0.223	± 0	± 0	± 0	± 0
[10] 20-20/0.0-0.0	1	1	1	1	0.5	0.193	0.897	0.5	0.897	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.097	± 0.006	± 0	± 0.006	± 0	± 0	± 0	± 0
[11] 20-20/0.0-0.3	1	1	1	1	0.5	0.235	0.897	0.5	0.897	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.19	± 0.006	± 0	± 0.006	± 0	± 0	± 0	± 0
[12] 20-20/0.0-0.7	1	1	1	1	0.5	0.215	0.897	0.5	0.897	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.154	± 0.006	± 0	± 0.006	± 0	± 0	± 0	± 0
[13] 20-20/0.3-0.0	1	1	1	1	0.5	0.19	0.956	0.5	0.956	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.079	± 0.005	± 0	± 0.005	± 0	± 0	± 0	± 0
[14] 20-20/0.3-0.3	1	1	1	1	0.5	0.208	0.956	0.5	0.956	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.133	± 0.005	± 0	± 0.005	± 0	± 0	± 0	± 0
[15] 20-20/0.3-0.7	1	1	1	1	0.5	0.197	0.956	0.5	0.956	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.112	± 0.005	± 0	± 0.005	± 0	± 0	± 0	± 0
[16] 20-20/0.7-0.0	1	1	1	1	0.5	0.202	0.68	0.5	0.976	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.119	± 0.229	± 0	± 0.003	± 0	± 0	± 0	± 0
[17] 20-20/0.7-0.3	1	1	1	1	0.5	0.225	0.638	0.5	0.976	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.161	± 0.213	± 0	± 0.003	± 0	± 0	± 0	± 0
[18] 20-20/0.7-0.7	1	1	1	1	0.5	0.21	0.614	0.5	0.976	1	1	1	1
	± 0	± 0	± 0	± 0	± 0	± 0.143	± 0.2	± 0	± 0.003	± 0	± 0	± 0	± 0

Table 83 – Results of experiment B1 - Average and standard deviation of categorical attributes weights

Dataset		All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
[ID]	Scenario	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBs	PMGBR	KPTBS	KPTNM	KPTEL	KPTBR
[1]	02-02/0.0-0.0	1	1	1	0.126	0.5	0.718	0.126	0.5	0.126	1	1	0.13	0.126
		±0	±0	±0	±0.015	±0	±0.118	±0.015	±0	±0.015	±0	±0	±0.013	±0.015
[2]	02-02/0.0-0.3	1	1	1	0.129	0.5	0.395	0.129	0.5	0.126	1	1	0.131	0.126
		±0	±0	±0	±0.037	±0	±0.218	±0.037	±0	±0.015	±0	±0	±0.013	±0.015
[3]	02-02/0.0-0.7	1	1	1	0.129	0.5	0.372	0.129	0.5	0.126	1	1	0.131	0.126
		±0	±0	±0	±0.037	±0	±0.273	±0.037	±0	±0.015	±0	±0	±0.013	±0.015
[4]	02-02/0.3-0.0	1	1 ^K	1	0.49	0.5	0.662	0.488	0.5	0.202	1	1	0.065	0.104
		±0	±0	±0	±0.073	±0	±0.176	±0.114	±0	±0.206	±0	±0	±0.008	±0.126
[5]	02-02/0.3-0.3	1	1	1	0.449	0.5	0.37	0.424	0.5	0.176	1	1	0.065	0.121
		±0	±0	±0	±0.143	±0	±0.235	±0.276	±0	±0.194	±0	±0	±0.008	±0.164
[6]	02-02/0.3-0.7	1	1	1	0.471	0.5	0.385	0.439	0.5	0.149	1	1	0.065	0.127
		±0	±0	±0	±0.239	±0	±0.292	±0.284	±0	±0.177	±0	±0	±0.008	±0.185
[7]	02-02/0.7-0.0	1	1 ^{KWY}	1	0.558	0.5	0.62	0.528	0.5	0.301	1	1	0.038	0.442
		±0	±0	±0	±0.174	±0	±0.217	±0.222	±0	±0.233	±0	±0	±0.007	±0.163
[8]	02-02/0.7-0.3	1	1	1	0.596	0.5	0.185	0.655	0.5	0.255	1	1	0.038	0.498
		±0	±0	±0	±0.216	±0	±0.097	±0.259	±0	±0.232	±0	±0	±0.007	±0.095
[9]	02-02/0.7-0.7	1	1	1	0.62	0.5	0.26	0.627	0.5	0.208	1	1	0.038	0.541
		±0	±0	±0	±0.217	±0	±0.22	±0.234	±0	±0.223	±0	±0	±0.007	±0.146
[10]	20-20/0.0-0.0	1	1	1	0.103	0.5	0.807	0.103	0.5	0.103	1	1	0.104	0.103
		±0	±0	±0	±0.006	±0	±0.097	±0.006	±0	±0.006	±0	±0	±0.007	±0.006
[11]	20-20/0.0-0.3	1	1	1	0.103	0.5	0.765	0.103	0.5	0.103	1	1	0.104	0.103
		±0	±0	±0	±0.006	±0	±0.19	±0.006	±0	±0.006	±0	±0	±0.007	±0.006
[12]	20-20/0.0-0.7	1	1	1	0.103	0.5	0.785	0.103	0.5	0.103	1	1	0.104	0.103
		±0	±0	±0	±0.006	±0	±0.154	±0.006	±0	±0.006	±0	±0	±0.007	±0.006
[13]	20-20/0.3-0.0	1	1	1	0.044	0.5	0.81	0.044	0.5	0.044	1	1	0.044	0.044
		±0	±0	±0	±0.005	±0	±0.079	±0.005	±0	±0.005	±0	±0	±0.005	±0.005
[14]	20-20/0.3-0.3	1	1	1	0.044	0.5	0.792	0.044	0.5	0.044	1	1	0.044	0.044
		±0	±0	±0	±0.005	±0	±0.133	±0.005	±0	±0.005	±0	±0	±0.005	±0.005
[15]	20-20/0.3-0.7	1	1	1	0.044	0.5	0.803	0.044	0.5	0.044	1	1	0.044	0.044
		±0	±0	±0	±0.005	±0	±0.112	±0.005	±0	±0.005	±0	±0	±0.005	±0.005
[16]	20-20/0.7-0.0	1	1	1	0.33	0.5	0.798	0.32	0.5	0.024	1	1	0.024	0.024
		±0	±0	±0	±0.227	±0	±0.119	±0.229	±0	±0.003	±0	±0	±0.003	±0.003
[17]	20-20/0.7-0.3	1	1	1	0.362	0.5	0.775	0.362	0.5	0.024	1	1	0.024	0.024
		±0	±0	±0	±0.213	±0	±0.161	±0.213	±0	±0.003	±0	±0	±0.003	±0.003
[18]	20-20/0.7-0.7	1	1	1	0.39	0.5	0.79	0.386	0.5	0.024	1	1	0.024	0.024
		±0	±0	±0	±0.197	±0	±0.143	±0.2	±0	±0.003	±0	±0	±0.003	±0.003

Table 84 – Results of experiment B1 - Average and standard deviation of execution time

Dataset	All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
[ID] Scenario	KAMLA	FMMMD	KMDBS	KMDBR	WKMB	WKMM	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-02/0.0-0.0	0.025 ±0.003	0.035 ±0.006	0.002 ±0.001	0.137 ±0.013	0.002 ±0.001	1 ±0.04	0.177 ±0.012	0.012 ±0.016	0.856 ±0.049	0.017 ±0.004	0.016 ±0.003	0.017 ±0.003	0.9 ±0.05
[2] 02-02/0.0-0.3	0.03 ±0.009	0.036 ±0.012	0.002 ±0.001	0.14 ±0.014	0.002 ±0.001	1.015 ±0.047	0.178 ±0.011	0.012 ±0.016	0.857 ±0.049	0.018 ±0.004	0.018 ±0.005	0.019 ±0.013	0.923 ±0.054
[3] 02-02/0.0-0.7	0.033 ±0.008	0.036 ±0.006	0.002 ±0.001	0.135 ±0.012	0.002 ±0.001	0.998 ±0.038	0.181 ±0.014	0.012 ±0.017	0.864 ±0.051	0.018 ±0.004	0.02 ±0.014	0.018 ±0.004	0.939 ±0.065
[4] 02-02/0.3-0.0	0.032 ±0.014	0.038^K ±0.008	0.002 ±0.001	0.14 ±0.012	0.003 ±0.001	1.006 ±0.044	0.183 ±0.013	0.011 ±0.011	0.867 ±0.052	0.018 ±0.004	0.018 ±0.004	0.02 ±0.005	1.075 ±0.086
[5] 02-02/0.3-0.3	0.037 ±0.01	0.042 ±0.01	0.002 ±0.001	0.142 ±0.016	0.003 ±0.001	0.991 ±0.039	0.182 ±0.012	0.012 ±0.016	0.872 ±0.053	0.02 ±0.004	0.02 ±0.004	0.022 ±0.005	1.187 ±0.09
[6] 02-02/0.3-0.7	0.04 ±0.019	0.044 ±0.01	0.002 ±0.001	0.142 ±0.013	0.003 ±0.001	0.99 ±0.036	0.185 ±0.015	0.011 ±0.011	0.879 ±0.053	0.022 ±0.004	0.02 ±0.004	0.024 ±0.005	1.239 ±0.109
[7] 02-02/0.7-0.0	0.034 ±0.01	0.039^{KWY} ±0.013	0.002 ±0.001	0.152 ±0.015	0.002 ±0.001	0.988 ±0.033	0.203 ±0.017	0.012 ±0.016	0.883 ±0.052	0.019 ±0.005	0.02 ±0.005	0.038 ±0.012	1.404 ±0.165
[8] 02-02/0.7-0.3	0.037 ±0.015	0.043 ±0.009	0.002 ±0.001	0.152 ±0.015	0.003 ±0.001	0.998 ±0.041	0.198 ±0.017	0.01 ±0.011	0.889 ±0.056	0.024 ±0.005	0.022 ±0.006	0.038 ±0.01	1.575 ±0.154
[9] 02-02/0.7-0.7	0.037 ±0.016	0.048 ±0.012	0.002 ±0.001	0.155 ±0.016	0.003 ±0.001	0.996 ±0.04	0.198 ±0.017	0.01 ±0.011	0.898 ±0.06	0.027 ±0.005	0.024 ±0.006	0.04 ±0.012	1.683 ±0.137
[10] 20-20/0.0-0.0	0.044 ±0.012	0.063 ±0.019	0.01 ±0.011	0.658 ±0.056	0.014 ±0.011	5.513 ±0.155	0.881 ±0.06	0.031 ±0.019	2.332 ±0.047	0.068 ±0.018	0.069 ±0.017	0.071 ±0.016	3.359 ±0.159
[11] 20-20/0.0-0.3	0.046 ±0.016	0.064 ±0.017	0.01 ±0.003	0.662 ±0.056	0.013 ±0.003	5.569 ±0.165	0.922 ±0.071	0.033 ±0.019	2.252 ±0.05	0.07 ±0.015	0.07 ±0.021	0.07 ±0.013	3.317 ±0.153
[12] 20-20/0.0-0.7	0.045 ±0.013	0.059 ±0.009	0.009 ±0.002	0.672 ±0.054	0.014 ±0.012	5.643 ±0.182	0.895 ±0.064	0.033 ±0.018	2.37 ±0.045	0.07 ±0.019	0.072 ±0.021	0.074 ±0.02	3.356 ±0.157
[13] 20-20/0.3-0.0	0.044 ±0.009	0.06 ±0.013	0.01 ±0.003	0.67 ±0.057	0.012 ±0.003	5.65 ±0.183	0.893 ±0.055	0.03 ±0.014	2.275 ±0.047	0.065 ±0.019	0.068 ±0.012	0.071 ±0.02	3.342 ±0.155
[14] 20-20/0.3-0.3	0.045 ±0.013	0.063 ±0.015	0.01 ±0.003	0.654 ±0.042	0.013 ±0.003	5.703 ±0.177	0.894 ±0.06	0.031 ±0.014	2.347 ±0.064	0.065 ±0.019	0.068 ±0.012	0.07 ±0.014	3.318 ±0.149
[15] 20-20/0.3-0.7	0.047 ±0.016	0.062 ±0.01	0.009 ±0.003	0.652 ±0.043	0.012 ±0.003	5.617 ±0.15	0.885 ±0.052	0.033 ±0.018	2.385 ±0.053	0.066 ±0.017	0.07 ±0.022	0.07 ±0.019	3.372 ±0.149
[16] 20-20/0.7-0.0	0.046 ±0.013	0.064 ±0.01	0.01 ±0.002	0.671 ±0.058	0.013 ±0.003	5.621 ±0.162	0.898 ±0.061	0.029 ±0.006	2.272 ±0.047	0.065 ±0.013	0.07 ±0.02	0.076 ±0.016	3.717 ±0.29
[17] 20-20/0.7-0.3	0.045 ±0.017	0.063 ±0.014	0.009 ±0.003	0.658 ±0.044	0.013 ±0.003	5.679 ±0.21	0.901 ±0.062	0.032 ±0.014	2.328 ±0.048	0.067 ±0.018	0.071 ±0.021	0.081 ±0.022	3.764 ±0.271
[18] 20-20/0.7-0.7	0.046 ±0.016	0.062 ±0.014	0.009 ±0.002	0.657 ±0.04	0.013 ±0.003	5.588 ±0.148	0.898 ±0.052	0.033 ±0.018	2.376 ±0.048	0.066 ±0.019	0.071 ±0.02	0.083 ±0.02	3.87 ±0.293

Table 85 – Results of experiment B2 - ARPD values considering the ARI values of each repetition

Dataset	Group 1 (K)				Group 2 (W)				Group 3 (P)				Group 4 (Y)						
	[ID] Scenario	KAMLA	FMMMD	KMDBS	KMDBR	KAMLA	FMMMD	WKMMB	WKMBR	KAMLA	FMMMD	PMGBS	PMGBR	KAMLA	FMMMD	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-02/0.0-0.0	43.924	84.694	0.835	0	43.941	84.705	0.916	1.188	0	43.717	84.551	0.403	0.065	43.685	84.604	7.374	14.323	17.815	0.275
[2] 02-02/0.0-0.3	74.763	98.25	23.744	0	74.763	98.25	23.744	26.395	0	70.551	97.782	5.038	3.151	70.473	97.835	8.772	15.116	21.157	3.602
[3] 02-02/0.0-0.7	82.185	99.857	47.592	0	82.185	99.857	47.592	41.009	0	77.524	99.721	12.451	8.296	77.709	99.729	22.494	27.06	32.87	7.646
[4] 02-02/0.3-0.0	56.519	86.423	0.868	0.032	56.6	86.434	1.045	0.837	0.032	56.473	86.322	0.531	0.177	56.489	86.328	7.326	12.305	16.569	0.21
[5] 02-02/0.3-0.3	68.988	96.301	28.284	0	68.989	96.3	28.05	24.595	0	64.903	95.361	7.687	5.969	65.489	95.535	14.835	21.244	29.651	5.864
[6] 02-02/0.3-0.7	81.885	99.936	53.888	0	81.885	99.936	53.888	49.524	0	78.129	99.849	12.023	8.701	78.499	99.852	19.111	27.408	34.464	8.698
[7] 02-02/0.7-0.0	54.407	87.569	0.98	0.113	54.488	87.581	1.141	0.853	0.08	54.394	87.496	0.562	0.209	54.393	87.471	6.421	13.326	15.655	0.338
[8] 02-02/0.7-0.3	71.388	96.903	27.143	0	71.388	96.903	27.143	25.143	0	66.695	96.269	6.181	4.449	67.387	96.274	12.489	18.964	26.973	4.615
[9] 02-02/0.7-0.7	77.981	99.938	54.025	0	77.981	99.938	54.025	51.482	0	73.677	99.865	14.359	10.969	74.516	99.871	21.732	28.878	34.281	10.87
[10] 20-20/0.0-0.0	31	2.212	0	0	31	2.212	0	0	0	31	2.212	0	0	31	2.212	6.032	11.055	13.05	0
[11] 20-20/0.0-0.3	50.991	5.207	0	0	50.991	5.207	0	0	0	50.991	5.207	0	0	50.991	5.207	0	3.009	12.015	0
[12] 20-20/0.0-0.7	54.993	8.315	0	0	54.993	8.315	0	0	0	54.993	8.315	0	0	54.993	8.315	0	4.991	13.031	0
[13] 20-20/0.3-0.0	31	2.075	0	0	31	2.075	0	0	0	31	2.075	0	0	31	2.075	2.006	9.008	9.006	0
[14] 20-20/0.3-0.3	51.002	5.29	0	0	51.002	5.29	0	0	0	51.002	5.29	0	0	51.002	5.29	0	6.904	14.002	0
[15] 20-20/0.3-0.7	54.997	8.664	0	0	54.997	8.664	0	0	0	54.997	8.649	0	0	54.997	8.634	1.996	7.003	16.016	0
[16] 20-20/0.7-0.0	32	2.074	0	0	32	2.074	0	0	0	32	2.074	0	0	32	2.074	1.003	9.015	10.088	0
[17] 20-20/0.7-0.3	52.002	5.212	0	0	52.002	5.212	0	0	0	52.002	5.212	0	0	52.002	5.212	0	5.003	14.002	0
[18] 20-20/0.7-0.7	54.997	8.665	0	0	54.997	8.665	0	0	0	54.997	8.65	0	0	54.997	8.635	1.996	7.003	15.016	0

Table 86 – Results of experiment B2 - Average and standard deviation of numerical attributes weights

Dataset	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)		Group 4 (Y)			
[ID] Scenario	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTEL	KPTBR
[1] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.447 ±0.159	0.225 ±0.337	0.5 ±0	0.696 ±0.415	1 ±0	1 ±0	1 ±0	1 ±0
[2] 02-02/0.0-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.443 ±0.158	0.17 ±0.251	0.5 ±0	0.584 ±0.45	1 ±0	1 ±0	1 ±0	1 ±0
[3] 02-02/0.0-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.443 ±0.158	0.21 ±0.287	0.5 ±0	0.555 ±0.451	1 ±0	1 ±0	1 ±0	1 ±0
[4] 02-02/0.3-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.45 ±0.16	0.292 ±0.393	0.5 ±0	0.742 ±0.374	1 ±0	1 ±0	1 ±0	1 ±0
[5] 02-02/0.3-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.447 ±0.159	0.209 ±0.283	0.5 ±0	0.625 ±0.432	1 ±0	1 ±0	1 ±0	1 ±0
[6] 02-02/0.3-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.447 ±0.159	0.222 ±0.291	0.5 ±0	0.65 ±0.418	1 ±0	1 ±0	1 ±0	1 ±0
[7] 02-02/0.7-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.457 ±0.162	0.284 ±0.387	0.5 ±0	0.733 ±0.38	1 ±0	1 ±0	1 ±0	1 ±0
[8] 02-02/0.7-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.447 ±0.159	0.223 ±0.29	0.5 ±0	0.625 ±0.432	1 ±0	1 ±0	1 ±0	1 ±0
[9] 02-02/0.7-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.443 ±0.158	0.253 ±0.301	0.5 ±0	0.584 ±0.44	1 ±0	1 ±0	1 ±0	1 ±0
[10] 20-20/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0.333	0.737 ±0.415	0.5 ±0	0.793 ±0.372	1 ±0	1 ±0	1 ±0	1 ±0
[11] 20-20/0.0-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0.333	0.611 ±0.469	0.5 ±0	0.589 ±0.471	1 ±0	1 ±0	1 ±0	1 ±0
[12] 20-20/0.0-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0.333	0.466 ±0.485	0.5 ±0	0.17 ±0.362	1 ±0	1 ±0	1 ±0	1 ±0
[13] 20-20/0.3-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0.333	0.796 ±0.362	0.5 ±0	0.901 ±0.229	1 ±0	1 ±0	1 ±0	1 ±0
[14] 20-20/0.3-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0.333	0.739 ±0.405	0.5 ±0	0.708 ±0.421	1 ±0	1 ±0	1 ±0	1 ±0
[15] 20-20/0.3-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0.333	0.671 ±0.441	0.5 ±0	0.354 ±0.46	1 ±0	1 ±0	1 ±0	1 ±0
[16] 20-20/0.7-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0.333	0.797 ±0.362	0.5 ±0	0.901 ±0.229	1 ±0	1 ±0	1 ±0	1 ±0
[17] 20-20/0.7-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0.333	0.748 ±0.399	0.5 ±0	0.708 ±0.421	1 ±0	1 ±0	1 ±0	1 ±0
[18] 20-20/0.7-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0.333	0.671 ±0.441	0.5 ±0	0.382 ±0.468	1 ±0	1 ±0	1 ±0	1 ±0

Table 87 – Results of experiment B2 - Average and standard deviation of categorical attributes weights

Dataset		All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)			
[ID]	Scenario	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBR	KPTBS	KPTNM	KPTBL	KPTBR
[1]	02-02/0.0-0.0	1	1	1	0.816	0.5	0.553	0.775	0.5	0.304	1	1	0.02
		±0	±0	±0	±0.29	±0	±0.159	±0.337	±0	±0.415	±0	±0	±0.009
[2]	02-02/0.0-0.3	1	1	1	0.798	0.5	0.557	0.83	0.5	0.416	1	1	0.02
		±0	±0	±0	±0.258	±0	±0.158	±0.251	±0	±0.45	±0	±0	±0.009
[3]	02-02/0.0-0.7	1	1	1	0.805	0.5	0.557	0.79	0.5	0.445	1	1	0.02
		±0	±0	±0	±0.265	±0	±0.158	±0.287	±0	±0.451	±0	±0	±0.009
[4]	02-02/0.3-0.0	1	1	1	0.696	0.5	0.55	0.708	0.5	0.258	1	1	0.019
		±0	±0	±0	±0.383	±0	±0.16	±0.393	±0	±0.374	±0	±0	±0.009
[5]	02-02/0.3-0.3	1	1	1	0.803	0.5	0.553	0.791	0.5	0.375	1	1	0.019
		±0	±0	±0	±0.269	±0	±0.159	±0.283	±0	±0.432	±0	±0	±0.009
[6]	02-02/0.3-0.7	1	1	1	0.779	0.5	0.553	0.778	0.5	0.35	1	1	0.019
		±0	±0	±0	±0.285	±0	±0.159	±0.291	±0	±0.418	±0	±0	±0.009
[7]	02-02/0.7-0.0	1	1	1	0.704	0.5	0.543	0.716	0.5	0.267	1	1	0.019
		±0	±0	±0	±0.388	±0	±0.162	±0.387	±0	±0.38	±0	±0	±0.009
[8]	02-02/0.7-0.3	1	1	1	0.767	0.5	0.553	0.777	0.5	0.375	1	1	0.019
		±0	±0	±0	±0.292	±0	±0.159	±0.29	±0	±0.432	±0	±0	±0.009
[9]	02-02/0.7-0.7	1	1	1	0.76	0.5	0.557	0.747	0.5	0.416	1	1	0.019
		±0	±0	±0	±0.3	±0	±0.158	±0.301	±0	±0.44	±0	±0	±0.009
[10]	20-20/0.0-0.0	1	1	1	0.292	0.5	0.667	0.263	0.5	0.207	1	1	0.005
		±0	±0	±0	±0.431	±0	±0	±0.415	±0	±0.372	±0	±0	±0.003
[11]	20-20/0.0-0.3	1	1	1	0.428	0.5	0.667	0.389	0.5	0.411	1	1	0.005
		±0	±0	±0	±0.478	±0	±0	±0.469	±0	±0.471	±0	±0	±0.003
[12]	20-20/0.0-0.7	1	1	1	0.602	0.5	0.667	0.534	0.5	0.83	1	1	0.005
		±0	±0	±0	±0.477	±0	±0	±0.485	±0	±0.362	±0	±0	±0.003
[13]	20-20/0.3-0.0	1	1	1	0.213	0.5	0.667	0.204	0.5	0.099	1	1	0.005
		±0	±0	±0	±0.37	±0	±0	±0.362	±0	±0.229	±0	±0	±0.003
[14]	20-20/0.3-0.3	1	1	1	0.261	0.5	0.667	0.261	0.5	0.292	1	1	0.005
		±0	±0	±0	±0.405	±0	±0	±0.405	±0	±0.421	±0	±0	±0.003
[15]	20-20/0.3-0.7	1	1	1	0.338	0.5	0.667	0.329	0.5	0.646	1	1	0.005
		±0	±0	±0	±0.445	±0	±0	±0.441	±0	±0.46	±0	±0	±0.003
[16]	20-20/0.7-0.0	1	1	1	0.213	0.5	0.667	0.203	0.5	0.099	1	1	0.005
		±0	±0	±0	±0.37	±0	±0	±0.362	±0	±0.229	±0	±0	±0.003
[17]	20-20/0.7-0.3	1	1	1	0.261	0.5	0.667	0.252	0.5	0.292	1	1	0.005
		±0	±0	±0	±0.405	±0	±0	±0.399	±0	±0.421	±0	±0	±0.003
[18]	20-20/0.7-0.7	1	1	1	0.329	0.5	0.667	0.329	0.5	0.618	1	1	0.005
		±0	±0	±0	±0.441	±0	±0	±0.441	±0	±0.468	±0	±0	±0.003

Table 88 – Results of experiment B2 - Average and standard deviation of execution time

Dataset		All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
[ID]	Scenario	KAMLA	FMMMD	KMDBS	KMDBR	WKMSBS	WKMMSS	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTBL	KPTBR
[1]	02-02/0.0-0.0	0.041 ±0.02	0.113 ±0.043	0.002 ±0.001	0.159 ±0.027	0.003 ±0.001	0.534 ±0.036	0.17 ±0.018	0.011 ±0.012	0.532 ±0.056	0.023 ±0.01	0.026 ±0.01	0.032 ±0.015	1.571 ±0.292
[2]	02-02/0.0-0.3	0.045 ±0.02	0.098 ±0.037	0.002 ±0.001	0.15 ±0.016	0.003 ±0.001	0.519 ±0.023	0.175 ±0.018	0.011 ±0.012	0.562 ±0.061	0.03 ±0.013	0.031 ±0.015	0.036 ±0.017	1.843 ±0.26
[3]	02-02/0.0-0.7	0.046 ±0.033	0.1 ±0.037	0.002 ±0.001	0.155 ±0.018	0.003 ±0.001	0.516 ±0.027	0.175 ±0.016	0.011 ±0.012	0.573 ±0.064	0.034 ±0.019	0.034 ±0.015	0.036 ±0.016	1.976 ±0.245
[4]	02-02/0.3-0.0	0.038 ±0.012	0.106 ±0.042	0.002 ±0.001	0.147 ±0.017	0.002 ±0.001	0.515 ±0.023	0.17 ±0.018	0.011 ±0.012	0.533 ±0.055	0.021 ±0.008	0.025 ±0.01	0.028 ±0.012	1.512 ±0.207
[5]	02-02/0.3-0.3	0.043 ±0.018	0.104 ±0.048	0.002 ±0.001	0.152 ±0.016	0.003 ±0.001	0.513 ±0.022	0.176 ±0.02	0.011 ±0.012	0.554 ±0.053	0.031 ±0.011	0.035 ±0.015	0.036 ±0.015	1.797 ±0.255
[6]	02-02/0.3-0.7	0.047 ±0.021	0.105 ±0.052	0.002 ±0.001	0.156 ±0.02	0.003 ±0.001	0.52 ±0.028	0.175 ±0.018	0.011 ±0.012	0.572 ±0.062	0.035 ±0.014	0.037 ±0.017	0.037 ±0.015	1.957 ±0.271
[7]	02-02/0.7-0.0	0.041 ±0.019	0.104 ±0.041	0.002 ±0.001	0.147 ±0.016	0.002 ±0.001	0.516 ±0.025	0.171 ±0.019	0.009 ±0.004	0.535 ±0.058	0.022 ±0.009	0.026 ±0.011	0.029 ±0.012	1.523 ±0.22
[8]	02-02/0.7-0.3	0.042 ±0.019	0.099 ±0.039	0.002 ±0.001	0.151 ±0.018	0.003 ±0.001	0.514 ±0.023	0.174 ±0.02	0.01 ±0.003	0.552 ±0.058	0.031 ±0.017	0.035 ±0.014	0.037 ±0.018	1.79 ±0.25
[9]	02-02/0.7-0.7	0.047 ±0.022	0.1 ±0.035	0.002 ±0.001	0.16 ±0.018	0.003 ±0.001	0.524 ±0.027	0.175 ±0.018	0.011 ±0.012	0.571 ±0.06	0.035 ±0.015	0.037 ±0.017	0.037 ±0.014	1.973 ±0.252
[10]	20-20/0.0-0.0	0.042 ±0.008	0.091 ±0.018	0.01 ±0.003	0.68 ±0.06	0.012 ±0.003	2.786 ±0.094	0.833 ±0.074	0.031 ±0.018	1.524 ±0.064	0.077 ±0.036	0.078 ±0.028	0.09 ±0.038	4.15 ±0.497
[11]	20-20/0.0-0.3	0.045 ±0.015	0.108 ±0.023	0.009 ±0.003	0.688 ±0.064	0.013 ±0.003	2.847 ±0.157	0.763 ±0.067	0.032 ±0.014	1.576 ±0.064	0.07 ±0.017	0.078 ±0.027	0.09 ±0.037	4.101 ±0.497
[12]	20-20/0.0-0.7	0.056 ±0.024	0.153 ±0.039	0.01 ±0.003	0.723 ±0.086	0.014 ±0.004	2.924 ±0.244	0.785 ±0.07	0.033 ±0.018	1.634 ±0.063	0.075 ±0.021	0.088 ±0.031	0.111 ±0.038	4.236 ±0.491
[13]	20-20/0.3-0.0	0.045 ±0.025	0.099 ±0.028	0.01 ±0.003	0.689 ±0.07	0.013 ±0.003	2.824 ±0.134	0.811 ±0.084	0.031 ±0.019	1.622 ±0.126	0.087 ±0.048	0.089 ±0.037	0.114 ±0.062	4.353 ±0.604
[14]	20-20/0.3-0.3	0.058 ±0.019	0.135 ±0.045	0.01 ±0.003	0.727 ±0.088	0.013 ±0.003	2.896 ±0.147	0.801 ±0.065	0.033 ±0.018	1.679 ±0.138	0.071 ±0.021	0.091 ±0.044	0.105 ±0.044	4.025 ±0.444
[15]	20-20/0.3-0.7	0.048 ±0.017	0.138 ±0.043	0.009 ±0.003	0.697 ±0.061	0.012 ±0.002	2.842 ±0.134	0.794 ±0.059	0.033 ±0.018	1.616 ±0.068	0.068 ±0.017	0.081 ±0.021	0.094 ±0.028	4.141 ±0.499
[16]	20-20/0.7-0.0	0.045 ±0.018	0.086 ±0.027	0.009 ±0.002	0.684 ±0.068	0.012 ±0.002	2.793 ±0.091	0.807 ±0.076	0.03 ±0.017	1.542 ±0.082	0.075 ±0.036	0.081 ±0.039	0.094 ±0.042	4.111 ±0.511
[17]	20-20/0.7-0.3	0.045 ±0.019	0.106 ±0.022	0.01 ±0.002	0.692 ±0.059	0.013 ±0.012	2.834 ±0.138	0.845 ±0.134	0.037 ±0.023	1.672 ±0.137	0.075 ±0.023	0.082 ±0.024	0.093 ±0.029	4.029 ±0.415
[18]	20-20/0.7-0.7	0.05 ±0.023	0.155 ±0.045	0.01 ±0.002	0.707 ±0.067	0.014 ±0.004	2.973 ±0.169	0.841 ±0.101	0.044 ±0.03	1.648 ±0.104	0.073 ±0.02	0.097 ±0.039	0.096 ±0.033	4.371 ±0.527

Table 89 – Results of experiment C1 - ARPD values considering the ARI values of each repetition

Dataset	Group 1 (K)					Group 2 (W)					Group 3 (P)					Group 4 (Y)				
	KAMLA	FMMMD	KMDBS	KMDBR	KAMLA	FMMMD	WKMBs	WKMBs	WKMBR	KAMLA	FMMMD	PMGBs	PMGBR	KAMLA	FMMMD	KPTBs	KPTNM	KPTBL	KPTBR	
1 02-08/0.0-0.0	0.016	0	0	0	0.016	0	0	0	0	0.016	0	0.544	0.016	0.016	0	0.752	0	0	0	
2 02-08/0.0-0.3	0.016	0	0.016	0	0.016	0	0.016	0	0	0.016	0	8.77	0	0.016	0	9.721	0.096	0.016	0.016	
3 02-08/0.0-0.7	0.016	0	0.016	0	0.016	0	0.016	0	0	0.016	0	25.222	0.016	0.016	0	29.423	0.56	0.096	0.016	
4 02-08/0.3-0.0	0.096	0	0	0	0.096	0	0	0	0.096	0	1.007	0.224	0.096	0	2.106	0.8	3.367	1.358		
5 02-08/0.3-0.3	0.112	0	0.016	0.016	0.112	0	0.016	0.016	0.112	0	11.364	0.656	0.112	0	11.426	1.837	4.716	2.66		
6 02-08/0.3-0.7	0.144	0	0.016	0.016	0.144	0	0.016	0.016	0.144	0	30.867	1.2	0.144	0	31.941	4.629	7.014	5.277		
7 02-08/0.7-0.0	0.272	0	0	0	0.272	0	0	0	0.272	0	1.343	0.208	0.272	0	12.301	1.341	60.107	9.379		
8 02-08/0.7-0.3	0.288	0	0.016	0.016	0.288	0	0.016	0.016	0.288	0	12.019	0.816	0.288	0	26.668	3.581	64.698	18.453		
9 02-08/0.7-0.7	0.32	0	0.016	0.016	0.32	0	0.016	0.016	0.32	0	32.924	1.568	0.32	0	54.293	9.572	72.948	37.224		
10 08-02/0.0-0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.643	0	0	0	
11 08-02/0.0-0.3	0	0	2.977	0	0	0	2.977	0	0	0	0	0	0	0	0	0.641	0	0	0	
12 08-02/0.0-0.7	0	0	1.982	0	0	0	1.982	0	0	0	0	0	0	0	0	1.609	0	0	0	
13 08-02/0.3-0.0	0	0	0.961	0	0	0	0.961	0	0	0	0	0	0	0	0	1.733	0	0.032	0.032	
14 08-02/0.3-0.3	0.064	0	2.982	0	0.064	0	2.982	0	0	0.064	0	0.032	0	0.064	0	13.754	0	0.032	0.032	
15 08-02/0.3-0.7	0.048	0	10.525	0	0.048	0	10.525	0	0	0.048	0	0.016	0	0.048	0	30.388	0	0.032	0.016	
16 08-02/0.7-0.0	7.904	0	8.493	0.208	7.904	0	8.493	0.336	0.432	7.904	0	0	0	7.904	0	22.943	0.016	34.885	18.402	
17 08-02/0.7-0.3	36.366	0	31.408	2.027	36.366	0	31.408	3.916	3.695	36.366	0	0.16	0	36.366	0	37.374	0.016	36.115	22.618	
18 08-02/0.7-0.7	71.483	0	50.948	1.373	71.483	0	50.948	8.7	2.454	71.483	0	0.304	0	71.483	0	53.74	0.016	37.499	28.548	
19 02-02/0.0-0.0	0.016	0	2.986	0	0.016	0	2.986	0	0	0.016	0	0.688	0.016	0.016	0	1.58	0.08	0.032	0.032	
20 02-02/0.0-0.3	0.016	0.016	3.15	0	0.016	0.016	3.15	0.377	0	0.016	0.016	7.786	0	0.016	0.016	14.99	0.064	0.016	0.016	
21 02-02/0.0-0.7	0.016	0.016	6.697	0	0.016	0.016	6.697	0	0	0.016	0.016	19.954	0	0.016	0.016	29.562	0.24	0.064	0.032	
22 02-02/0.3-0.0	0.592	0.032	5.24	0.128	0.592	0.032	5.24	0.224	0.24	0.577	0.016	2.881	1.566	0.577	0.016	9.336	3.245	9.409	6.593	
23 02-02/0.3-0.3	1.501	0.064	11.475	0.641	1.501	0.064	11.475	2.012	1.28	1.516	0.08	11.404	1.788	1.501	0.064	21.322	3.861	10.186	8.074	
24 02-02/0.3-0.7	1.709	0.064	20.965	0.656	1.709	0.064	20.965	2.33	1.247	1.692	0.048	25.455	2.041	1.677	0.032	37.777	5.358	11.286	9.271	
25 02-02/0.7-0.0	2.012	0	15.847	0.976	2.012	0	15.847	2.371	1.247	2.012	0	3.412	1.721	2.012	0	71.385	6.382	82.551	53.646	
26 02-02/0.7-0.3	14.699	0.016	43.237	7.745	14.699	0.016	43.237	16.8	10.338	14.697	0.016	13.022	2.248	14.684	0	79.364	7.832	83.303	59.566	
27 02-02/0.7-0.7	31.628	0.032	63.08	3.616	31.612	0.016	63.077	28.959	6.222	31.614	0.016	27.502	2.407	31.598	0	86.427	12.43	85.02	68.584	

Table 90 – Results of experiment C1 - Average and standard deviation of numerical attributes weights

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)		Group 4 (Y)			
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-08/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0	0.924 ±0.04	0.5 ±0	0.928 ±0.007	1 ±0	1 ±0	1 ±0	1 ±0
[2] 02-08/0.0-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.337 ±0.033	0.924 ±0.04	0.5 ±0	0.926 ±0.017	1 ±0	1 ±0	1 ±0	1 ±0
[3] 02-08/0.0-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.337 ±0.033	0.924 ±0.04	0.5 ±0	0.929 ±0.01	1 ±0	1 ±0	1 ±0	1 ±0
[4] 02-08/0.3-0.0	1 ±0	1 ^{PY} ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0	0.498 ±0.043	0.5 ±0	0.687 ±0.276	1 ±0	1 ±0	1 ±0	1 ±0
[5] 02-08/0.3-0.3	1 ±0	1 ^{PY} ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0	0.498 ±0.043	0.5 ±0	0.898 ±0.082	1 ±0	1 ±0	1 ±0	1 ±0
[6] 02-08/0.3-0.7	1 ±0	1 ^{PY} ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0	0.498 ±0.043	0.5 ±0	0.934 ±0.051	1 ±0	1 ±0	1 ±0	1 ±0
[7] 02-08/0.7-0.0	1 ±0	1 ^{PY} ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0	0.496 ±0.05	0.5 ±0	0.661 ±0.273	1 ±0	1 ±0	1 ±0	1 ±0
[8] 02-08/0.7-0.3	1 ±0	1 ^{PY} ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0	0.496 ±0.05	0.5 ±0	0.921 ±0.058	1 ±0	1 ±0	1 ±0	1 ±0
[9] 02-08/0.7-0.7	1 ±0	1 ^{PY} ±0	1 ±0	1 ±0	0.5 ±0	0.333 ±0	0.496 ±0.05	0.5 ±0	0.944 ±0.044	1 ±0	1 ±0	1 ±0	1 ±0
[10] 08-02/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.343 ±0.057	0.938 ±0.005	0.5 ±0	0.938 ±0.005	1 ±0	1 ±0	1 ±0	1 ±0
[11] 08-02/0.0-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.403 ±0.136	0.938 ±0.005	0.5 ±0	0.938 ±0.005	1 ±0	1 ±0	1 ±0	1 ±0
[12] 08-02/0.0-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.483 ±0.167	0.938 ±0.005	0.5 ±0	0.938 ±0.005	1 ±0	1 ±0	1 ±0	1 ±0
[13] 08-02/0.3-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.44 ±0.156	0.964 ±0.06	0.5 ±0	0.973 ±0.003	1 ±0	1 ±0	1 ±0	1 ±0
[14] 08-02/0.3-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.6 ±0.134	0.964 ±0.06	0.5 ±0	0.973 ±0.003	1 ±0	1 ±0	1 ±0	1 ±0
[15] 08-02/0.3-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.627 ±0.109	0.964 ±0.06	0.5 ±0	0.973 ±0.003	1 ±0	1 ±0	1 ±0	1 ±0
[16] 08-02/0.7-0.0	1 ±0	1 ^{KW} ±0	1 ±0	1 ±0	0.5 ±0	0.56 ±0.156	0.53 ±0.09	0.5 ±0	0.985 ±0.002	1 ±0	1 ±0	1 ±0	1 ±0
[17] 08-02/0.7-0.3	1 ±0	1 ^{KW} ±0	1 ±0	1 ±0	0.5 ±0	0.663 ±0.033	0.548 ±0.127	0.5 ±0	0.985 ±0.002	1 ±0	1 ±0	1 ±0	1 ±0
[18] 08-02/0.7-0.7	1 ±0	1 ^{KW} ±0	1 ±0	1 ±0	0.5 ±0	0.637 ±0.096	0.506 ±0.14	0.5 ±0	0.985 ±0.002	1 ±0	1 ±0	1 ±0	1 ±0
[19] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.47 ±0.165	0.915 ±0.073	0.5 ±0	0.926 ±0.016	1 ±0	1 ±0	1 ±0	1 ±0
[20] 02-02/0.0-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.6 ±0.134	0.918 ±0.061	0.5 ±0	0.926 ±0.031	1 ±0	1 ±0	1 ±0	1 ±0
[21] 02-02/0.0-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.603 ±0.131	0.921 ±0.046	0.5 ±0	0.929 ±0.015	1 ±0	1 ±0	1 ±0	1 ±0
[22] 02-02/0.3-0.0	1 ±0	1 ^{WPY} ±0	1 ±0	1 ±0	0.5 ±0	0.563 ±0.155	0.529 ±0.077	0.5 ±0	0.792 ±0.24	1 ±0	1 ±0	1 ±0	1 ±0
[23] 02-02/0.3-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.653 ±0.066	0.55 ±0.083	0.5 ±0	0.914 ±0.085	1 ±0	1 ±0	1 ±0	1 ±0
[24] 02-02/0.3-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.643 ±0.085	0.498 ±0.18	0.5 ±0	0.937 ±0.058	1 ±0	1 ±0	1 ±0	1 ±0
[25] 02-02/0.7-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.55 ±0.16	0.463 ±0.184	0.5 ±0	0.754 ±0.249	1 ±0	1 ±0	1 ±0	1 ±0
[26] 02-02/0.7-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.63 ±0.105	0.489 ±0.18	0.5 ±0	0.917 ±0.075	1 ±0	1 ±0	1 ±0	1 ±0
[27] 02-02/0.7-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.637 ±0.096	0.387 ±0.226	0.5 ±0	0.94 ±0.056	1 ±0	1 ±0	1 ±0	1 ±0

Table 91 – Results of experiment C1 - Average and standard deviation of categorical attributes weights

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)		Group 4 (Y)			
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTBL	KPTBR
[1] 02-08/0.0-0.0	1 ±0	1 ±0	1 ±0	0.076 ±0.04	0.5 ±0	0.667 ±0	0.076 ±0.04	0.5 ±0	0.072 ±0.007	1 ±0	1 ±0	0.072 ±0.007	0.072 ±0.007
[2] 02-08/0.0-0.3	1 ±0	1 ±0	1 ±0	0.076 ±0.04	0.5 ±0	0.663 ±0.033	0.076 ±0.04	0.5 ±0	0.074 ±0.017	1 ±0	1 ±0	0.072 ±0.007	0.072 ±0.007
[3] 02-08/0.0-0.7	1 ±0	1 ±0	1 ±0	0.085 ±0.073	0.5 ±0	0.663 ±0.033	0.076 ±0.04	0.5 ±0	0.071 ±0.01	1 ±0	1 ±0	0.072 ±0.007	0.069 ±0.016
[4] 02-08/0.3-0.0	1 ±0	1 ^{PY} ±0	1 ±0	0.502 ±0.043	0.5 ±0	0.667 ±0	0.502 ±0.043	0.5 ±0	0.313 ±0.276	1 ±0	1 ±0	0.036 ±0.004	0.337 ±0.286
[5] 02-08/0.3-0.3	1 ±0	1 ^{PY} ±0	1 ±0	0.502 ±0.043	0.5 ±0	0.667 ±0	0.502 ±0.043	0.5 ±0	0.102 ±0.082	1 ±0	1 ±0	0.036 ±0.004	0.071 ±0.057
[6] 02-08/0.3-0.7	1 ±0	1 ^{PY} ±0	1 ±0	0.502 ±0.043	0.5 ±0	0.667 ±0	0.502 ±0.043	0.5 ±0	0.066 ±0.051	1 ±0	1 ±0	0.036 ±0.004	0.049 ±0.051
[7] 02-08/0.7-0.0	1 ±0	1 ^{PY} ±0	1 ±0	0.503 ±0.05	0.5 ±0	0.667 ±0	0.503 ±0.05	0.5 ±0	0.339 ±0.273	1 ±0	1 ±0	0.021 ±0.004	0.632 ±0.288
[8] 02-08/0.7-0.3	1 ±0	1 ^{PY} ±0	1 ±0	0.503 ±0.05	0.5 ±0	0.667 ±0	0.503 ±0.05	0.5 ±0	0.079 ±0.058	1 ±0	1 ±0	0.021 ±0.004	0.577 ±0.338
[9] 02-08/0.7-0.7	1 ±0	1 ^{PY} ±0	1 ±0	0.503 ±0.05	0.5 ±0	0.667 ±0	0.503 ±0.05	0.5 ±0	0.056 ±0.044	1 ±0	1 ±0	0.021 ±0.004	0.527 ±0.34
[10] 08-02/0.0-0.0	1 ±0	1 ±0	1 ±0	0.062 ±0.005	0.5 ±0	0.657 ±0.057	0.062 ±0.005	0.5 ±0	0.062 ±0.005	1 ±0	1 ±0	0.063 ±0.005	0.062 ±0.005
[11] 08-02/0.0-0.3	1 ±0	1 ±0	1 ±0	0.062 ±0.005	0.5 ±0	0.597 ±0.136	0.062 ±0.005	0.5 ±0	0.062 ±0.005	1 ±0	1 ±0	0.063 ±0.005	0.062 ±0.005
[12] 08-02/0.0-0.7	1 ±0	1 ±0	1 ±0	0.062 ±0.005	0.5 ±0	0.517 ±0.167	0.062 ±0.005	0.5 ±0	0.062 ±0.005	1 ±0	1 ±0	0.063 ±0.005	0.062 ±0.005
[13] 08-02/0.3-0.0	1 ±0	1 ±0	1 ±0	0.036 ±0.06	0.5 ±0	0.56 ±0.156	0.036 ±0.06	0.5 ±0	0.027 ±0.003	1 ±0	1 ±0	0.027 ±0.003	0.027 ±0.003
[14] 08-02/0.3-0.3	1 ±0	1 ±0	1 ±0	0.036 ±0.06	0.5 ±0	0.4 ±0.134	0.036 ±0.06	0.5 ±0	0.027 ±0.003	1 ±0	1 ±0	0.027 ±0.003	0.027 ±0.003
[15] 08-02/0.3-0.7	1 ±0	1 ±0	1 ±0	0.036 ±0.06	0.5 ±0	0.373 ±0.109	0.036 ±0.06	0.5 ±0	0.027 ±0.003	1 ±0	1 ±0	0.027 ±0.003	0.028 ±0.012
[16] 08-02/0.7-0.0	1 ±0	1 ^{KW} ±0	1 ±0	0.519 ±0.117	0.5 ±0	0.44 ±0.156	0.47 ±0.09	0.5 ±0	0.015 ±0.002	1 ±0	1 ±0	0.015 ±0.002	0.59 ±0.312
[17] 08-02/0.7-0.3	1 ±0	1 ^{KW} ±0	1 ±0	0.573 ±0.172	0.5 ±0	0.337 ±0.033	0.452 ±0.127	0.5 ±0	0.015 ±0.002	1 ±0	1 ±0	0.015 ±0.002	0.27 ±0.31
[18] 08-02/0.7-0.7	1 ±0	1 ^{KW} ±0	1 ±0	0.66 ±0.211	0.5 ±0	0.363 ±0.096	0.494 ±0.14	0.5 ±0	0.015 ±0.002	1 ±0	1 ±0	0.015 ±0.002	0.129 ±0.241
[19] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	0.085 ±0.073	0.5 ±0	0.53 ±0.165	0.085 ±0.073	0.5 ±0	0.074 ±0.016	1 ±0	1 ±0	0.072 ±0.007	0.072 ±0.007
[20] 02-02/0.0-0.3	1 ±0	1 ±0	1 ±0	0.09 ±0.107	0.5 ±0	0.4 ±0.134	0.082 ±0.061	0.5 ±0	0.074 ±0.031	1 ±0	1 ±0	0.072 ±0.007	0.072 ±0.007
[21] 02-02/0.0-0.7	1 ±0	1 ±0	1 ±0	0.089 ±0.086	0.5 ±0	0.397 ±0.131	0.079 ±0.046	0.5 ±0	0.071 ±0.015	1 ±0	1 ±0	0.072 ±0.007	0.071 ±0.01
[22] 02-02/0.3-0.0	1 ±0	1 ^{WPY} ±0	1 ±0	0.519 ±0.098	0.5 ±0	0.437 ±0.155	0.471 ±0.077	0.5 ±0	0.208 ±0.24	1 ±0	1 ±0	0.036 ±0.004	0.403 ±0.305
[23] 02-02/0.3-0.3	1 ±0	1 ±0	1 ±0	0.561 ±0.172	0.5 ±0	0.347 ±0.066	0.45 ±0.083	0.5 ±0	0.086 ±0.085	1 ±0	1 ±0	0.036 ±0.004	0.08 ±0.16
[24] 02-02/0.3-0.7	1 ±0	1 ±0	1 ±0	0.622 ±0.202	0.5 ±0	0.357 ±0.085	0.502 ±0.18	0.5 ±0	0.063 ±0.058	1 ±0	1 ±0	0.036 ±0.004	0.04 ±0.03
[25] 02-02/0.7-0.0	1 ±0	1 ±0	1 ±0	0.552 ±0.153	0.5 ±0	0.45 ±0.16	0.537 ±0.184	0.5 ±0	0.246 ±0.249	1 ±0	1 ±0	0.021 ±0.004	0.627 ±0.285
[26] 02-02/0.7-0.3	1 ±0	1 ±0	1 ±0	0.665 ±0.188	0.5 ±0	0.37 ±0.105	0.511 ±0.18	0.5 ±0	0.083 ±0.075	1 ±0	1 ±0	0.021 ±0.004	0.637 ±0.296
[27] 02-02/0.7-0.7	1 ±0	1 ±0	1 ±0	0.832 ±0.192	0.5 ±0	0.363 ±0.096	0.613 ±0.226	0.5 ±0	0.06 ±0.056	1 ±0	1 ±0	0.021 ±0.004	0.635 ±0.316

Table 92 – Results of experiment C1 - Average and standard deviation of execution time

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBBS	WKMMBS	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-08/0.0-0.0	0.034 ±0.006	0.081 ±0.01	0.013 ±0.025	1.057 ±0.055	0.015 ±0.003	3.748 ±0.106	0.965 ±0.038	0.009 ±0.003	0.692 ±0.062	0.032 ±0.008	0.028 ±0.007	0.029 ±0.008	1.523 ±0.097
[2] 02-08/0.0-0.3	0.039 ±0.016	0.084 ±0.012	0.011 ±0.002	1.032 ±0.052	0.015 ±0.002	3.746 ±0.113	0.988 ±0.057	0.01 ±0.003	0.699 ±0.049	0.044 ±0.012	0.033 ±0.009	0.033 ±0.011	1.751 ±0.109
[3] 02-08/0.0-0.7	0.042 ±0.017	0.09 ±0.011	0.011 ±0.002	1.041 ±0.06	0.015 ±0.003	3.748 ±0.097	0.981 ±0.046	0.01 ±0.003	0.704 ±0.051	0.044 ±0.012	0.036 ±0.011	0.033 ±0.015	1.851 ±0.102
[4] 02-08/0.3-0.0	0.037 ±0.011	0.087^{PY} ±0.017	0.011 ±0.003	1.062 ±0.066	0.015 ±0.003	3.821 ±0.132	1.009 ±0.069	0.011 ±0.011	0.702 ±0.046	0.036 ±0.008	0.034 ±0.008	0.041 ±0.011	2.033 ±0.16
[5] 02-08/0.3-0.3	0.041 ±0.015	0.087^{PY} ±0.01	0.011 ±0.003	1.08 ±0.063	0.015 ±0.002	3.822 ±0.135	1.002 ±0.057	0.011 ±0.012	0.723 ±0.059	0.042 ±0.008	0.038 ±0.01	0.043 ±0.011	2.222 ±0.127
[6] 02-08/0.3-0.7	0.045 ±0.024	0.09^{PY} ±0.01	0.012 ±0.004	1.043 ±0.046	0.015 ±0.002	3.711 ±0.11	0.988 ±0.045	0.01 ±0.003	0.728 ±0.059	0.045 ±0.01	0.044 ±0.017	0.045 ±0.011	2.392 ±0.137
[7] 02-08/0.7-0.0	0.04 ±0.017	0.082^{PY} ±0.016	0.011 ±0.002	1.085 ±0.064	0.015 ±0.003	3.727 ±0.088	1.015 ±0.049	0.011 ±0.011	0.722 ±0.061	0.059 ±0.017	0.04 ±0.011	0.063 ±0.022	2.674 ±0.22
[8] 02-08/0.7-0.3	0.044 ±0.019	0.086^{PY} ±0.016	0.011 ±0.003	1.077 ±0.059	0.016 ±0.011	3.766 ±0.12	1.017 ±0.049	0.01 ±0.003	0.739 ±0.06	0.047 ±0.011	0.043 ±0.012	0.066 ±0.021	2.861 ±0.234
[9] 02-08/0.7-0.7	0.047 ±0.012	0.095^{PY} ±0.018	0.011 ±0.003	1.094 ±0.064	0.016 ±0.012	3.743 ±0.113	1.005 ±0.039	0.01 ±0.003	0.745 ±0.054	0.048 ±0.012	0.055 ±0.022	0.066 ±0.019	2.885 ±0.231
[10] 08-02/0.0-0.0	0.037 ±0.017	0.039 ±0.006	0.004 ±0.002	0.351 ±0.022	0.005 ±0.002	1.44 ±0.039	0.36 ±0.024	0.017 ±0.014	0.998 ±0.056	0.025 ±0.007	0.025 ±0.008	0.027 ±0.009	1.264 ±0.059
[11] 08-02/0.0-0.3	0.036 ±0.012	0.04 ±0.006	0.004 ±0.002	0.356 ±0.03	0.005 ±0.001	1.473 ±0.052	0.365 ±0.022	0.016 ±0.006	1.009 ±0.058	0.03 ±0.01	0.026 ±0.009	0.026 ±0.009	1.32 ±0.088
[12] 08-02/0.0-0.7	0.037 ±0.013	0.04 ±0.006	0.004 ±0.001	0.358 ±0.024	0.005 ±0.002	1.479 ±0.064	0.361 ±0.021	0.016 ±0.006	1.008 ±0.058	0.031 ±0.011	0.027 ±0.008	0.026 ±0.009	1.325 ±0.088
[13] 08-02/0.3-0.0	0.046 ±0.011	0.04 ±0.013	0.004 ±0.001	0.361 ±0.028	0.005 ±0.002	1.474 ±0.065	0.364 ±0.026	0.018 ±0.014	1.005 ±0.051	0.028 ±0.007	0.027 ±0.009	0.031 ±0.01	1.452 ±0.101
[14] 08-02/0.3-0.3	0.051 ±0.013	0.041 ±0.007	0.004 ±0.001	0.358 ±0.028	0.005 ±0.002	1.474 ±0.062	0.361 ±0.022	0.016 ±0.006	1 ±0.057	0.031 ±0.007	0.027 ±0.01	0.03 ±0.01	1.466 ±0.071
[15] 08-02/0.3-0.7	0.054 ±0.02	0.041 ±0.007	0.004 ±0.001	0.355 ±0.035	0.006 ±0.002	1.46 ±0.057	0.361 ±0.021	0.018 ±0.014	0.995 ±0.049	0.03 ±0.007	0.027 ±0.008	0.029 ±0.01	1.507 ±0.082
[16] 08-02/0.7-0.0	0.043 ±0.015	0.041^{KW} ±0.006	0.004 ±0.001	0.385 ±0.027	0.005 ±0.002	1.471 ±0.072	0.388 ±0.026	0.015 ±0.005	1.016 ±0.052	0.047 ±0.019	0.03 ±0.008	0.06 ±0.018	2.672 ±0.226
[17] 08-02/0.7-0.3	0.046 ±0.014	0.042^{KW} ±0.007	0.004 ±0.001	0.39 ±0.034	0.006 ±0.002	1.458 ±0.07	0.401 ±0.034	0.015 ±0.005	1.03 ±0.064	0.048 ±0.015	0.032 ±0.009	0.065 ±0.022	2.742 ±0.197
[18] 08-02/0.7-0.7	0.047 ±0.015	0.044^{KW} ±0.008	0.004 ±0.002	0.389 ±0.028	0.005 ±0.002	1.463 ±0.079	0.401 ±0.027	0.016 ±0.014	1.031 ±0.059	0.047 ±0.019	0.032 ±0.01	0.063 ±0.018	2.866 ±0.229
[19] 02-02/0.0-0.0	0.036 ±0.009	0.048 ±0.008	0.004 ±0.001	0.333 ±0.036	0.005 ±0.001	1.181 ±0.038	0.324 ±0.018	0.009 ±0.003	0.631 ±0.048	0.021 ±0.012	0.018 ±0.005	0.019 ±0.006	1.007 ±0.06
[20] 02-02/0.0-0.3	0.039 ±0.009	0.046 ±0.008	0.004 ±0.001	0.338 ±0.026	0.005 ±0.001	1.197 ±0.051	0.332 ±0.026	0.009 ±0.003	0.638 ±0.05	0.022 ±0.005	0.02 ±0.006	0.019 ±0.005	1.054 ±0.081
[21] 02-02/0.0-0.7	0.041 ±0.012	0.048 ±0.008	0.004 ±0.001	0.338 ±0.027	0.005 ±0.002	1.192 ±0.05	0.321 ±0.017	0.01 ±0.003	0.634 ±0.046	0.021 ±0.004	0.02 ±0.006	0.019 ±0.005	1.039 ±0.057
[22] 02-02/0.3-0.0	0.045 ±0.015	0.049^{W-PY} ±0.01	0.004 ±0.001	0.345 ±0.024	0.005 ±0.001	1.193 ±0.049	0.339 ±0.034	0.009 ±0.003	0.653 ±0.049	0.027 ±0.008	0.024 ±0.006	0.028 ±0.007	1.447 ±0.117
[23] 02-02/0.3-0.3	0.052 ±0.018	0.054 ±0.013	0.004 ±0.002	0.342 ±0.02	0.005 ±0.002	1.2 ±0.054	0.331 ±0.019	0.009 ±0.003	0.657 ±0.048	0.026 ±0.006	0.026 ±0.006	0.028 ±0.008	1.478 ±0.125
[24] 02-02/0.3-0.7	0.053 ±0.013	0.057 ±0.013	0.004 ±0.001	0.351 ±0.026	0.005 ±0.002	1.199 ±0.054	0.344 ±0.024	0.01 ±0.003	0.665 ±0.049	0.025 ±0.007	0.027 ±0.007	0.03 ±0.015	1.508 ±0.129
[25] 02-02/0.7-0.0	0.042 ±0.012	0.051 ±0.01	0.004 ±0.001	0.36 ±0.03	0.005 ±0.001	1.198 ±0.056	0.353 ±0.038	0.01 ±0.012	0.665 ±0.05	0.03 ±0.007	0.032 ±0.011	0.042 ±0.013	1.822 ±0.145
[26] 02-02/0.7-0.3	0.048 ±0.017	0.059 ±0.015	0.004 ±0.002	0.356 ±0.024	0.005 ±0.002	1.189 ±0.047	0.345 ±0.019	0.01 ±0.003	0.669 ±0.05	0.03 ±0.007	0.034 ±0.012	0.043 ±0.013	1.841 ±0.176
[27] 02-02/0.7-0.7	0.048 ±0.016	0.067 ±0.017	0.004 ±0.001	0.363 ±0.031	0.005 ±0.001	1.207 ±0.05	0.357 ±0.028	0.009 ±0.003	0.674 ±0.051	0.029 ±0.006	0.036 ±0.017	0.042 ±0.014	1.81 ±0.151

Table 93 – Results of experiment C2 - ARPD values considering the ARI values of each repetition

Dataset	Group 1 (K)				Group 2 (W)				Group 3 (P)				Group 4 (Y)						
	KAMLA	FMMMD	KMDBS	KMDBR	KAMLA	FMMMD	WKMBBS	WKMMBS	KAMLA	FMMMD	PMGBS	PMGBR	KAMLA	FMMMD	KPTBS	KPTNM	KPTPL	KPTBR	
[1] 02-08/0.0-0.0	49.065	1.147	0	0	49.065	1.147	0	4.001	0	49.05	0.43	24.524	20.906	49.05	0.43	27.566	33.362	48.21	16.468
[2] 02-08/0.0-0.3	47.062	2.572	0	0	47.062	2.572	0	10.051	0	47.047	1.082	39.686	35.581	47.047	1.082	43.286	47.577	63.428	28.699
[3] 02-08/0.0-0.7	50.114	5.942	0	0	50.114	5.942	0	15.007	0	50.022	2.985	64.63	60.536	50.021	2.983	70.913	74.993	83.159	50.991
[4] 02-08/0.3-0.0	42.998	1.117	0	0	42.998	1.117	0	6.994	0	42.998	0.495	23.653	21.023	42.998	0.495	26.58	32.746	45.028	16.416
[5] 02-08/0.3-0.3	47.008	2.666	0	0	47.008	2.666	0	16.913	0	47.009	1.112	39.891	36.304	47.009	1.112	44.376	49.778	60.246	28.779
[6] 02-08/0.3-0.7	50.097	6.379	0	0	50.097	6.379	0	22.052	0	50.005	3.901	65.42	60.369	50.004	3.899	71.97	73.653	81.16	50.487
[7] 02-08/0.7-0.0	41.016	1.132	0	0	41.016	1.132	0	7.011	0	41	0.494	23.906	20.842	41	0.494	28.7	33.584	43.313	16.469
[8] 02-08/0.7-0.3	48.005	2.653	0	0	48.005	2.653	0	15.908	0	48.006	1.081	39.71	36.338	48.006	1.081	46.607	51.177	60.847	28.596
[9] 02-08/0.7-0.7	50.08	6.537	0	0	50.08	6.537	0	21.791	0	49.986	3.906	65.264	60.281	49.985	3.91	71.646	74.505	81.277	50.914
[10] 08-02/0.0-0.0	44.125	99.741	51.527	0.306	44.118	99.741	51.512	57.476	0.258	22.492	100.502	52.565	47.065	25.587	99.316	74.76	93.628	99.889	43.384
[11] 08-02/0.0-0.3	55.99	99.969	72.468	0.701	56.077	99.962	72.157	79.099	0.739	26.893	100.271	56.502	43.776	28.951	100.085	80.635	94.954	99.627	36.84
[12] 08-02/0.0-0.7	77.337	99.965	80.482	3.203	80.085	99.877	82.637	90.024	2.401	53.92	102.226	66.827	22.138	62.173	99.924	77.825	91.549	99.178	8.994
[13] 08-02/0.3-0.0	39.796	99.829	49.49	0.258	39.795	99.829	49.501	60.066	0.258	20.48	99.757	59.296	51.974	21.784	99.748	77.037	95.294	100.111	46.835
[14] 08-02/0.3-0.3	50.256	99.762	74.556	0.374	50.951	99.761	74.992	69.535	0.39	19.949	99.394	65.292	49.647	22.905	99.29	79.192	94.61	99.218	40.764
[15] 08-02/0.3-0.7	77.871	99.811	85.378	1.603	85.065	99.784	88.041	88.077	0.774	49.87	101.199	63.633	26.863	61.29	99.445	75.058	95.019	99.569	8.997
[16] 08-02/0.7-0.0	38.289	98.846	48.837	0.21	38.249	99.847	48.838	59.571	0.226	17.999	99.744	60.54	53.266	19.703	99.793	77.399	95.01	100.024	48.273
[17] 08-02/0.7-0.3	49.527	99.719	75.904	0.436	50.354	99.726	76.033	72.24	0.528	18.263	99.714	63.695	50.153	20.898	99.626	78.946	94.684	99.155	41.282
[18] 08-02/0.7-0.7	99.969	81.134	97.907	0.806	80.856	99.817	88.012	87.602	0.608	22.602	99.822	63.242	49.682	24.347	99.822	79.192	94.61	99.218	40.764
[19] 02-02/0.0-0.0	71.623	1.426	0.744	0.081	71.601	85.618	76.005	80.688	0.077	15.373	81.868	61.214	24.847	57.595	80.648	70.486	82.92	98.41	32.292
[20] 02-02/0.0-0.3	79.866	98.344	59.379	0.475	80.043	98.343	60.548	90.367	0.273	NaN	Inf	Inf	Inf	63.97	96.827	66.686	79.329	99.723	17.941
[21] 02-02/0.0-0.7	89.741	99.943	99.899	1.685	87.046	99.428	80.424	95.279	2.151	60.047	98.948	54.483	14.975	78.027	98.542	72.923	84.054	99.178	5.867
[22] 02-02/0.3-0.0	69.623	0.162	0.017	0.001	67.073	71.195	67.303	71.195	0.001	NaN	NaN	NaN	NaN	NaN	NaN	67.303	71.195	67.303	71.195
[23] 02-02/0.3-0.3	73.316	95.462	64.116	0.669	73.655	95.507	64.865	88.991	0.657	67.169	109.682	54.85	32.631	58.313	91.927	70.882	85.7	98.717	27.492
[24] 02-02/0.3-0.7	89.087	99.77	79.006	2.941	90.56	99.088	81.197	93.637	0.763	98.34	99.735	54.85	14.975	80.163	98.993	74.419	86.738	99.62	5.557
[25] 02-02/0.7-0.0	61.641	85.893	43.021	0.13	61.651	85.891	43.048	76.451	0.146	98.34	99.735	54.85	14.975	80.163	98.993	74.419	86.738	99.62	5.557
[26] 02-02/0.7-0.3	85.067	96.946	64.116	0.669	85.413	96.946	64.116	88.991	0.657	67.169	109.682	54.85	32.631	58.313	91.927	70.882	85.7	98.717	27.492
[27] 02-02/0.7-0.7	87.058	99.454	78.482	1.661	87.659	99.683	83.033	94.029	0.676	98.34	99.735	54.85	14.975	80.163	98.993	74.419	86.738	99.62	5.557

Table 94 – Results of experiment C2 - Average and standard deviation of numerical attributes weights

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)		Group 4 (Y)			
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTBL	KPTBR
[1] 02-08/0.0-0.0	1 ±0	1^{PY} ±0	1 ±0	1 ±0	0.5 ±0	0.289 ±0.247	0.036 ±0.153	0.5 ±0	0.401 ±0.409	1 ±0	1 ±0	1 ±0	1 ±0
[2] 02-08/0.0-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.317 ±0.275	0.018 ±0.09	0.5 ±0	0.386 ±0.409	1 ±0	1 ±0	1 ±0	1 ±0
[3] 02-08/0.0-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.381 ±0.314	0.009 ±0.003	0.5 ±0	0.4 ±0.402	1 ±0	1 ±0	1 ±0	1 ±0
[4] 02-08/0.3-0.0	1 ±0	1^{PY} ±0	1 ±0	1 ±0	0.5 ±0	0.299 ±0.259	0.224 ±0.385	0.5 ±0	0.431 ±0.427	1 ±0	1 ±0	1 ±0	1 ±0
[5] 02-08/0.3-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.32 ±0.281	0.116 ±0.291	0.5 ±0	0.409 ±0.413	1 ±0	1 ±0	1 ±0	1 ±0
[6] 02-08/0.3-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.403 ±0.327	0.027 ±0.125	0.5 ±0	0.411 ±0.422	1 ±0	1 ±0	1 ±0	1 ±0
[7] 02-08/0.7-0.0	1 ±0	1^{PY} ±0	1 ±0	1 ±0	0.5 ±0	0.344 ±0.289	0.224 ±0.385	0.5 ±0	0.43 ±0.425	1 ±0	1 ±0	1 ±0	1 ±0
[8] 02-08/0.7-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.343 ±0.297	0.107 ±0.281	0.5 ±0	0.39 ±0.409	1 ±0	1 ±0	1 ±0	1 ±0
[9] 02-08/0.7-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.436 ±0.333	0.027 ±0.125	0.5 ±0	0.405 ±0.422	1 ±0	1 ±0	1 ±0	1 ±0
[10] 08-02/0.0-0.0	1^{PY} ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.717 ±0.232	0.28 ±0.322	0.5 ±0	0.258 ±0.307	1 ±0	1 ±0	1 ±0	1 ±0
[11] 08-02/0.0-0.3	1^{PY} ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.806 ±0.141	0.277 ±0.31	0.5 ±0	0.341 ±0.316	1 ±0	1 ±0	1 ±0	1 ±0
[12] 08-02/0.0-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.807 ±0.154	0.329 ±0.305	0.5 ±0	0.387 ±0.355	1 ±0	1 ±0	1 ±0	1 ±0
[13] 08-02/0.3-0.0	1^{PY} ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.691 ±0.249	0.33 ±0.324	0.5 ±0	0.195 ±0.294	1 ±0	1 ±0	1 ±0	1 ±0
[14] 08-02/0.3-0.3	1^{PY} ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.781 ±0.173	0.305 ±0.294	0.5 ±0	0.265 ±0.313	1 ±0	1 ±0	1 ±0	1 ±0
[15] 08-02/0.3-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.73 ±0.229	0.295 ±0.277	0.5 ±0	0.368 ±0.349	1 ±0	1 ±0	1 ±0	1 ±0
[16] 08-02/0.7-0.0	1^{PY} ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.669 ±0.254	0.328 ±0.329	0.5 ±0	0.191 ±0.306	1 ±0	1 ±0	1 ±0	1 ±0
[17] 08-02/0.7-0.3	1^{PY} ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.777 ±0.174	0.284 ±0.312	0.5 ±0	0.3 ±0.318	1 ±0	1 ±0	1 ±0	1 ±0
[18] 08-02/0.7-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.734 ±0.232	0.282 ±0.287	0.5 ±0	0.387 ±0.357	1 ±0	1 ±0	1 ±0	1 ±0
[19] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.773 ±0.197	0.236 ±0.292	0.5 ±0	0.233 ±0.366	1 ±0	1 ±0	1 ±0	1 ±0
[20] 02-02/0.0-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.816 ±0.141	0.27 ±0.273	0.5 ±0	0.343 ±0.387	1 ±0	1 ±0	1 ±0	1 ±0
[21] 02-02/0.0-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.834 ±0.081	0.256 ±0.277	0.5 ±0	0.464 ±0.423	1 ±0	1 ±0	1 ±0	1 ±0
[22] 02-02/0.3-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.724 ±0.246	0.228 ±0.299	0.5 ±0	0.339 ±0.418	1 ±0	1 ±0	1 ±0	1 ±0
[23] 02-02/0.3-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.804 ±0.159	0.306 ±0.286	0.5 ±0	0.479 ±0.403	1 ±0	1 ±0	1 ±0	1 ±0
[24] 02-02/0.3-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.8 ±0.16	0.333 ±0.277	0.5 ±0	0.412 ±0.403	1 ±0	1 ±0	1 ±0	1 ±0
[25] 02-02/0.7-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.759 ±0.229	0.227 ±0.294	0.5 ±0	0.299 ±0.403	1 ±0	1 ±0	1 ±0	1 ±0
[26] 02-02/0.7-0.3	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.809 ±0.156	0.368 ±0.286	0.5 ±0	0.462 ±0.408	1 ±0	1 ±0	1 ±0	1 ±0
[27] 02-02/0.7-0.7	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.804 ±0.154	0.274 ±0.262	0.5 ±0	0.459 ±0.407	1 ±0	1 ±0	1 ±0	1 ±0

Table 95 – Results of experiment C2 - Average and standard deviation of categorical attributes weights

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)		Group 4 (Y)			
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-08/0.0-0.0	1 ±0	1^{PY} ±0	1 ±0	0.955 ±0.177	0.5 ±0	0.711 ±0.247	0.964 ±0.153	0.5 ±0	0.599 ±0.409	1 ±0	1 ±0	0.011 ±0.005	0.483 ±0.384
[2] 02-08/0.0-0.3	1 ±0	1 ±0	1 ±0	0.991 ±0.003	0.5 ±0	0.683 ±0.275	0.982 ±0.09	0.5 ±0	0.614 ±0.409	1 ±0	1 ±0	0.011 ±0.005	0.412 ±0.387
[3] 02-08/0.0-0.7	1 ±0	1 ±0	1 ±0	0.991 ±0.003	0.5 ±0	0.619 ±0.314	0.991 ±0.003	0.5 ±0	0.6 ±0.402	1 ±0	1 ±0	0.011 ±0.005	0.388 ±0.367
[4] 02-08/0.3-0.0	1 ±0	1^{PY} ±0	1 ±0	0.812 ±0.36	0.5 ±0	0.701 ±0.259	0.776 ±0.385	0.5 ±0	0.569 ±0.427	1 ±0	1 ±0	0.01 ±0.005	0.512 ±0.402
[5] 02-08/0.3-0.3	1 ±0	1 ±0	1 ±0	0.945 ±0.198	0.5 ±0	0.68 ±0.281	0.884 ±0.291	0.5 ±0	0.591 ±0.413	1 ±0	1 ±0	0.01 ±0.005	0.468 ±0.396
[6] 02-08/0.3-0.7	1 ±0	1 ±0	1 ±0	0.991 ±0.003	0.5 ±0	0.597 ±0.327	0.973 ±0.125	0.5 ±0	0.589 ±0.422	1 ±0	1 ±0	0.01 ±0.005	0.384 ±0.382
[7] 02-08/0.7-0.0	1 ±0	1^{PY} ±0	1 ±0	0.838 ±0.339	0.5 ±0	0.656 ±0.289	0.776 ±0.385	0.5 ±0	0.57 ±0.425	1 ±0	1 ±0	0.01 ±0.005	0.461 ±0.396
[8] 02-08/0.7-0.3	1 ±0	1 ±0	1 ±0	0.955 ±0.177	0.5 ±0	0.657 ±0.297	0.893 ±0.281	0.5 ±0	0.61 ±0.409	1 ±0	1 ±0	0.01 ±0.005	0.486 ±0.415
[9] 02-08/0.7-0.7	1 ±0	1 ±0	1 ±0	0.991 ±0.003	0.5 ±0	0.564 ±0.333	0.973 ±0.125	0.5 ±0	0.595 ±0.422	1 ±0	1 ±0	0.01 ±0.005	0.363 ±0.366
[10] 08-02/0.0-0.0	1^{PY} ±0	1 ±0	1 ±0	0.727 ±0.316	0.5 ±0	0.283 ±0.232	0.72 ±0.322	0.5 ±0	0.742 ±0.307	1 ±0	1 ±0	0.005 ±0.002	0.459 ±0.376
[11] 08-02/0.0-0.3	1^{PY} ±0	1 ±0	1 ±0	0.68 ±0.299	0.5 ±0	0.194 ±0.141	0.723 ±0.31	0.5 ±0	0.659 ±0.316	1 ±0	1 ±0	0.005 ±0.002	0.418 ±0.355
[12] 08-02/0.0-0.7	1 ±0	1 ±0	1 ±0	0.731 ±0.276	0.5 ±0	0.193 ±0.154	0.671 ±0.305	0.5 ±0	0.613 ±0.355	1 ±0	1 ±0	0.005 ±0.002	0.451 ±0.391
[13] 08-02/0.3-0.0	1^{PY} ±0	1 ±0	1 ±0	0.736 ±0.292	0.5 ±0	0.309 ±0.249	0.67 ±0.324	0.5 ±0	0.805 ±0.294	1 ±0	1 ±0	0.005 ±0.002	0.41 ±0.359
[14] 08-02/0.3-0.3	1^{PY} ±0	1 ±0	1 ±0	0.723 ±0.293	0.5 ±0	0.219 ±0.173	0.695 ±0.294	0.5 ±0	0.735 ±0.313	1 ±0	1 ±0	0.005 ±0.002	0.488 ±0.388
[15] 08-02/0.3-0.7	1 ±0	1 ±0	1 ±0	0.703 ±0.272	0.5 ±0	0.27 ±0.229	0.705 ±0.277	0.5 ±0	0.632 ±0.349	1 ±0	1 ±0	0.005 ±0.002	0.395 ±0.361
[16] 08-02/0.7-0.0	1^{PY} ±0	1 ±0	1 ±0	0.749 ±0.292	0.5 ±0	0.331 ±0.254	0.672 ±0.329	0.5 ±0	0.809 ±0.306	1 ±0	1 ±0	0.005 ±0.002	0.442 ±0.371
[17] 08-02/0.7-0.3	1^{PY} ±0	1 ±0	1 ±0	0.727 ±0.28	0.5 ±0	0.223 ±0.174	0.716 ±0.312	0.5 ±0	0.7 ±0.318	1 ±0	1 ±0	0.005 ±0.002	0.396 ±0.366
[18] 08-02/0.7-0.7	1 ±0	1 ±0	1 ±0	0.72 ±0.269	0.5 ±0	0.266 ±0.232	0.718 ±0.287	0.5 ±0	0.613 ±0.357	1 ±0	1 ±0	0.005 ±0.002	0.422 ±0.377
[19] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	0.84 ±0.242	0.5 ±0	0.227 ±0.197	0.764 ±0.292	0.5 ±0	0.767 ±0.366	1 ±0	1 ±0	0.011 ±0.005	0.437 ±0.358
[20] 02-02/0.0-0.3	1 ±0	1 ±0	1 ±0	0.776 ±0.244	0.5 ±0	0.184 ±0.141	0.73 ±0.273	0.5 ±0	0.657 ±0.387	1 ±0	1 ±0	0.011 ±0.005	0.485 ±0.399
[21] 02-02/0.0-0.7	1 ±0	1 ±0	1 ±0	0.767 ±0.27	0.5 ±0	0.166 ±0.081	0.744 ±0.277	0.5 ±0	0.536 ±0.423	1 ±0	1 ±0	0.011 ±0.005	0.471 ±0.391
[22] 02-02/0.3-0.0	1 ±0	1 ±0	1 ±0	0.837 ±0.253	0.5 ±0	0.276 ±0.246	0.772 ±0.299	0.5 ±0	0.661 ±0.418	1 ±0	1 ±0	0.01 ±0.005	0.517 ±0.379
[23] 02-02/0.3-0.3	1 ±0	1 ±0	1 ±0	0.718 ±0.278	0.5 ±0	0.196 ±0.159	0.694 ±0.286	0.5 ±0	0.521 ±0.403	1 ±0	1 ±0	0.01 ±0.005	0.467 ±0.395
[24] 02-02/0.3-0.7	1 ±0	1 ±0	1 ±0	0.777 ±0.238	0.5 ±0	0.2 ±0.16	0.667 ±0.277	0.5 ±0	0.588 ±0.403	1 ±0	1 ±0	0.01 ±0.005	0.442 ±0.384
[25] 02-02/0.7-0.0	1 ±0	1 ±0	1 ±0	0.826 ±0.257	0.5 ±0	0.241 ±0.229	0.773 ±0.294	0.5 ±0	0.701 ±0.403	1 ±0	1 ±0	0.01 ±0.005	0.515 ±0.366
[26] 02-02/0.7-0.3	1 ±0	1 ±0	1 ±0	0.733 ±0.281	0.5 ±0	0.191 ±0.156	0.632 ±0.286	0.5 ±0	0.538 ±0.408	1 ±0	1 ±0	0.01 ±0.005	0.52 ±0.39
[27] 02-02/0.7-0.7	1 ±0	1 ±0	1 ±0	0.765 ±0.257	0.5 ±0	0.196 ±0.154	0.726 ±0.262	0.5 ±0	0.541 ±0.407	1 ±0	1 ±0	0.01 ±0.005	0.437 ±0.392

Table 96 – Results of experiment C2 - Average and standard deviation of execution time

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBs	PMGBR	KPTBS	KPTNM	KPTBL	KPTBR
[1] 02-08/0.0-0.0	0.041 ±0.007	0.11^{Py} ±0.021	0.011 ±0.002	0.773 ±0.059	0.015 ±0.003	8.821 ±0.208	1.129 ±0.081	0.01 ±0.011	0.862 ±0.063	0.056 ±0.018	0.065 ±0.023	0.069 ±0.032	4.619 ±0.35
[2] 02-08/0.0-0.3	0.047 ±0.014	0.132 ±0.032	0.011 ±0.003	0.762 ±0.05	0.015 ±0.003	8.811 ±0.202	1.104 ±0.064	0.011 ±0.012	0.893 ±0.067	0.058 ±0.021	0.066 ±0.026	0.068 ±0.031	4.919 ±0.351
[3] 02-08/0.0-0.7	0.054 ±0.019	0.167 ±0.056	0.011 ±0.003	0.769 ±0.058	0.016 ±0.012	8.751 ±0.135	1.111 ±0.059	0.01 ±0.004	0.91 ±0.082	0.061 ±0.022	0.062 ±0.025	0.071 ±0.033	4.882 ±0.332
[4] 02-08/0.3-0.0	0.042 ±0.014	0.109^{Py} ±0.024	0.011 ±0.002	0.772 ±0.054	0.015 ±0.003	8.703 ±0.135	1.102 ±0.051	0.01 ±0.003	0.85 ±0.059	0.054 ±0.02	0.061 ±0.023	0.063 ±0.025	4.654 ±0.423
[5] 02-08/0.3-0.3	0.049 ±0.017	0.162 ±0.048	0.012 ±0.003	0.835 ±0.096	0.015 ±0.003	9.06 ±0.33	1.145 ±0.082	0.01 ±0.003	0.921 ±0.085	0.063 ±0.024	0.065 ±0.024	0.068 ±0.025	5.142 ±0.501
[6] 02-08/0.3-0.7	0.057 ±0.017	0.24 ±0.101	0.012 ±0.003	0.785 ±0.067	0.015 ±0.011	9.192 ±0.426	1.128 ±0.07	0.011 ±0.004	0.915 ±0.086	0.057 ±0.023	0.061 ±0.02	0.072 ±0.029	4.916 ±0.407
[7] 02-08/0.7-0.0	0.045 ±0.022	0.121^{Py} ±0.025	0.012 ±0.003	0.803 ±0.053	0.017 ±0.004	9.118 ±0.257	1.153 ±0.068	0.01 ±0.003	0.874 ±0.068	0.06 ±0.02	0.064 ±0.025	0.068 ±0.027	4.727 ±0.422
[8] 02-08/0.7-0.3	0.048 ±0.024	0.158 ±0.052	0.013 ±0.013	0.826 ±0.081	0.015 ±0.002	8.991 ±0.247	1.148 ±0.067	0.01 ±0.004	0.942 ±0.093	0.065 ±0.024	0.066 ±0.027	0.073 ±0.027	4.943 ±0.38
[9] 02-08/0.7-0.7	0.053 ±0.025	0.184 ±0.068	0.011 ±0.003	0.769 ±0.06	0.014 ±0.002	8.853 ±0.201	1.12 ±0.079	0.01 ±0.012	0.912 ±0.077	0.057 ±0.02	0.061 ±0.025	0.072 ±0.03	4.818 ±0.36
[10] 08-02/0.0-0.0	0.043^{Py} ±0.013	0.119 ±0.061	0.004 ±0.001	0.301 ±0.037	0.005 ±0.002	3.417 ±0.098	0.436 ±0.044	0.017 ±0.015	1.272 ±0.066	0.064 ±0.024	0.058 ±0.031	0.075 ±0.036	4.752 ±0.307
[11] 08-02/0.0-0.3	0.045^{Py} ±0.013	0.108 ±0.035	0.004 ±0.001	0.299 ±0.036	0.005 ±0.001	3.495 ±0.192	0.433 ±0.037	0.017 ±0.007	1.267 ±0.068	0.064 ±0.026	0.065 ±0.031	0.084 ±0.041	4.752 ±0.315
[12] 08-02/0.0-0.7	0.05 ±0.015	0.124 ±0.045	0.004 ±0.002	0.302 ±0.039	0.006 ±0.002	3.402 ±0.104	0.444 ±0.047	0.019 ±0.016	1.278 ±0.076	0.063 ±0.025	0.068 ±0.034	0.078 ±0.034	4.745 ±0.265
[13] 08-02/0.3-0.0	0.041^{Py} ±0.012	0.114 ±0.039	0.005 ±0.002	0.306 ±0.044	0.006 ±0.003	3.393 ±0.081	0.439 ±0.038	0.018 ±0.007	1.27 ±0.07	0.065 ±0.026	0.06 ±0.032	0.077 ±0.039	4.737 ±0.321
[14] 08-02/0.3-0.3	0.045^{Py} ±0.011	0.113 ±0.033	0.004 ±0.002	0.305 ±0.04	0.006 ±0.002	3.416 ±0.098	0.449 ±0.042	0.017 ±0.006	1.274 ±0.068	0.067 ±0.027	0.063 ±0.027	0.071 ±0.034	4.751 ±0.317
[15] 08-02/0.3-0.7	0.045 ±0.011	0.114 ±0.04	0.004 ±0.002	0.303 ±0.039	0.005 ±0.002	3.412 ±0.106	0.451 ±0.045	0.017 ±0.007	1.269 ±0.069	0.061 ±0.021	0.063 ±0.034	0.069 ±0.033	4.727 ±0.293
[16] 08-02/0.7-0.0	0.041^{Py} ±0.011	0.115 ±0.039	0.004 ±0.001	0.297 ±0.039	0.006 ±0.002	3.408 ±0.106	0.443 ±0.049	0.017 ±0.006	1.274 ±0.072	0.064 ±0.027	0.058 ±0.028	0.08 ±0.04	4.734 ±0.317
[17] 08-02/0.7-0.3	0.045^{Py} ±0.013	0.111 ±0.035	0.004 ±0.002	0.3 ±0.039	0.005 ±0.001	3.389 ±0.097	0.447 ±0.046	0.018 ±0.015	1.27 ±0.068	0.066 ±0.029	0.063 ±0.03	0.07 ±0.034	4.741 ±0.282
[18] 08-02/0.7-0.7	0.046 ±0.011	0.115 ±0.04	0.004 ±0.001	0.304 ±0.041	0.006 ±0.002	3.379 ±0.102	0.441 ±0.041	0.017 ±0.014	1.263 ±0.071	0.058 ±0.017	0.06 ±0.03	0.068 ±0.034	4.637 ±0.26
[19] 02-02/0.0-0.0	0.05 ±0.021	0.145 ±0.057	0.004 ±0.001	0.271 ±0.031	0.005 ±0.001	2.716 ±0.057	0.393 ±0.031	0.009 ±0.003	0.819 ±0.063	0.036 ±0.016	0.039 ±0.018	0.054 ±0.024	3.095 ±0.239
[20] 02-02/0.0-0.3	0.051 ±0.016	0.147 ±0.07	0.004 ±0.002	0.272 ±0.036	0.005 ±0.002	2.739 ±0.074	0.393 ±0.031	0.009 ±0.004	0.823 ±0.07	0.036 ±0.013	0.04 ±0.02	0.058 ±0.027	3.078 ±0.293
[21] 02-02/0.0-0.7	0.053 ±0.019	0.15 ±0.062	0.004 ±0.001	0.273 ±0.033	0.005 ±0.001	2.772 ±0.074	0.39 ±0.032	0.009 ±0.004	0.815 ±0.073	0.037 ±0.014	0.042 ±0.02	0.059 ±0.032	3.105 ±0.277
[22] 02-02/0.3-0.0	0.047 ±0.016	0.163 ±0.075	0.004 ±0.001	0.275 ±0.03	0.005 ±0.002	2.777 ±0.063	0.392 ±0.031	0.009 ±0.003	0.816 ±0.059	0.037 ±0.015	0.036 ±0.015	0.054 ±0.026	3.071 ±0.254
[23] 02-02/0.3-0.3	0.051 ±0.018	0.164 ±0.065	0.004 ±0.002	0.276 ±0.032	0.005 ±0.001	2.749 ±0.075	0.389 ±0.028	0.009 ±0.004	0.819 ±0.067	0.039 ±0.016	0.044 ±0.021	0.056 ±0.026	3.065 ±0.237
[24] 02-02/0.3-0.7	0.052 ±0.018	0.154 ±0.067	0.004 ±0.001	0.27 ±0.029	0.005 ±0.001	2.733 ±0.07	0.393 ±0.033	0.009 ±0.004	0.824 ±0.073	0.036 ±0.013	0.041 ±0.017	0.057 ±0.027	3.074 ±0.263
[25] 02-02/0.7-0.0	0.048 ±0.017	0.162 ±0.069	0.004 ±0.001	0.269 ±0.026	0.005 ±0.002	2.784 ±0.086	0.392 ±0.03	0.009 ±0.004	0.819 ±0.062	0.036 ±0.014	0.041 ±0.02	0.054 ±0.025	3.084 ±0.272
[26] 02-02/0.7-0.3	0.052 ±0.018	0.161 ±0.074	0.004 ±0.002	0.277 ±0.031	0.005 ±0.001	2.777 ±0.098	0.395 ±0.031	0.009 ±0.004	0.819 ±0.071	0.036 ±0.015	0.04 ±0.018	0.055 ±0.025	3.07 ±0.254
[27] 02-02/0.7-0.7	0.052 ±0.015	0.156 ±0.07	0.004 ±0.002	0.274 ±0.033	0.005 ±0.002	2.777 ±0.103	0.399 ±0.037	0.009 ±0.004	0.816 ±0.07	0.036 ±0.012	0.041 ±0.019	0.055 ±0.026	3.089 ±0.259

Table 97 – Results of experiment D1 - Average and standard deviation of ARI values

Dataset		All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
[ID]	Scenario	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBs	PMGBR	KPTBS	KPTNM	KPTBL	KPTBR
[1]	02-02/0.0-0.0	1	1	0.999	1	0.999	0.999	1	0.999	1	0.999	1	1	1
		±0.002	±0	±0.003	±0	±0.003	±0.003	±0	±0.003	±0	±0.003	±0.002	±0	±0
[2]	02-02/0.0-0.0	1	1	1	1	1	1	1	1	1	1	1	1	1
		±0	±0	±0.001	±0	±0.001	±0.001	±0	±0.001	±0	±0.001	±0.001	±0	±0
[3]	02-02/0.0-0.0	1	1	1	1	1	1	1	1	1	1	1	1	1
		±0	±0	±0.001	±0	±0.001	±0.001	±0	±0.001	±0	±0.001	±0.001	±0	±0
[4]	02-02/0.0-0.0	1	1	1	1	1	1	1	1	1	1	1	1	1
		±0.001	±0	±0.001	±0	±0.001	±0.001	±0	±0.001	±0	±0.001	±0.001	±0	±0
[5]	02-02/0.0-0.0	1	1	1	1	1	1	1	1	1	1	1	1	1
		±0	±0	±0	±0	±0	±0	±0	±0	±0	±0	±0	±0	±0
[6]	02-02/0.0-0.0	1	1	1	1	1	1	1	1	1	1	1	1	1
		±0	±0	±0.001	±0	±0.001	±0.001	±0	±0.001	±0	±0.001	±0.001	±0	±0

Table 98 – Results of experiment D1 - ARPD values considering the ARI values of each repetition

Dataset		Group 1 (K)				Group 2 (W)				Group 3 (P)				Group 4 (Y)						
[ID]	Scenario	KAMLA	FMMMD	KMDBS	KMDBR	KAMLA	FMMMD	WKMBs	WKMMs	WKMBR	KAMLA	FMMMD	PMGBS	PMGBR	KAMLA	FMMMD	KPTBS	KPTNM	KPTBL	KPTBR
[1]	02-02/0.0-0.0	0.016	0	0.064	0	0.016	0	0.064	0.064	0	0.016	0	0.064	0	0.016	0	0.064	0.016	0	0
[2]	02-02/0.0-0.0	0	0	0.008	0	0	0	0.008	0.008	0	0	0	0.008	0	0	0	0.008	0.008	0	0
[3]	02-02/0.0-0.0	0	0	0.01	0	0	0	0.01	0.01	0	0	0	0.01	0	0	0	0.01	0.005	0	0
[4]	02-02/0.0-0.0	0.02	0	0.008	0	0.02	0	0.008	0.008	0	0.02	0	0.008	0	0.02	0	0.008	0.008	0	0
[5]	02-02/0.0-0.0	0.008	0	0.004	0	0.008	0	0.004	0.004	0	0.008	0	0.006	0.002	0.008	0	0.006	0.004	0.002	0.002
[6]	02-02/0.0-0.0	0.003	0	0.013	0	0.003	0	0.013	0.013	0	0.003	0	0.013	0	0.003	0	0.013	0.013	0	0

Table 99 – Results of experiment D1 - Average and standard deviation of numerical attributes weights

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)			Group 4 (Y)		
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.282 ±0.118	0.874 ±0.015	0.5 ±0	0.874 ±0.015	1 ±0	1 ±0	1 ±0	1 ±0
[2] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.332 ±0.12	0.883 ±0.012	0.5 ±0	0.883 ±0.012	1 ±0	1 ±0	1 ±0	1 ±0
[3] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.345 ±0.098	0.887 ±0.01	0.5 ±0	0.887 ±0.01	1 ±0	1 ±0	1 ±0	1 ±0
[4] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.337 ±0.085	0.879 ±0.067	0.5 ±0	0.891 ±0.009	1 ±0	1 ±0	1 ±0	1 ±0
[5] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.335 ±0.09	0.88 ±0.078	0.5 ±0	0.896 ±0.008	1 ±0	1 ±0	1 ±0	1 ±0
[6] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.34 ±0.088	0.882 ±0.08	0.5 ±0	0.899 ±0.008	1 ±0	1 ±0	1 ±0	1 ±0

Table 100 – Results of experiment D1 - Average and standard deviation of categorical attributes weights

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)			Group 4 (Y)		
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	0.126 ±0.015	0.5 ±0	0.718 ±0.118	0.126 ±0.015	0.5 ±0	0.126 ±0.015	1 ±0	1 ±0	0.13 ±0.013	0.126 ±0.015
[2] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	0.117 ±0.012	0.5 ±0	0.668 ±0.12	0.117 ±0.012	0.5 ±0	0.117 ±0.012	1 ±0	1 ±0	0.121 ±0.011	0.117 ±0.012
[3] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	0.113 ±0.01	0.5 ±0	0.655 ±0.098	0.113 ±0.01	0.5 ±0	0.113 ±0.01	1 ±0	1 ±0	0.117 ±0.01	0.113 ±0.01
[4] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	0.121 ±0.067	0.5 ±0	0.663 ±0.085	0.121 ±0.067	0.5 ±0	0.109 ±0.009	1 ±0	1 ±0	0.113 ±0.01	0.109 ±0.009
[5] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	0.132 ±0.119	0.5 ±0	0.665 ±0.09	0.12 ±0.078	0.5 ±0	0.104 ±0.008	1 ±0	1 ±0	0.106 ±0.009	0.104 ±0.008
[6] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	0.126 ±0.096	0.5 ±0	0.66 ±0.088	0.118 ±0.08	0.5 ±0	0.101 ±0.008	1 ±0	1 ±0	0.102 ±0.007	0.101 ±0.008

Table 101 – Results of experiment D1 - Average and standard deviation of execution time

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)			Group 4 (Y)		
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-02/0.0-0.0	0.025 ±0.01	0.039 ±0.014	0.002 ±0.001	0.134 ±0.011	0.003 ±0.001	1.071 ±0.202	0.167 ±0.018	0.011 ±0.014	0.602 ±0.045	0.018 ±0.004	0.019 ±0.005	0.018 ±0.005	0.92 ±0.049
[2] 02-02/0.0-0.0	0.034 ±0.011	0.049 ±0.018	0.003 ±0.002	0.199 ±0.036	0.003 ±0.001	2.018 ±0.102	0.204 ±0.019	0.061 ±0.044	2.237 ±0.124	0.026 ±0.011	0.028 ±0.008	0.032 ±0.011	1.447 ±0.24
[3] 02-02/0.0-0.0	0.034 ±0.01	0.051 ±0.013	0.004 ±0.001	0.225 ±0.023	0.005 ±0.003	3.009 ±0.098	0.237 ±0.02	0.114 ±0.048	6.404 ±0.23	0.035 ±0.013	0.036 ±0.011	0.039 ±0.011	1.764 ±0.121
[4] 02-02/0.0-0.0	0.041 ±0.013	0.067 ±0.009	0.005 ±0.002	0.274 ±0.033	0.004 ±0.001	4.121 ±0.172	0.27 ±0.021	0.197 ±0.062	9.287 ±0.365	0.038 ±0.016	0.04 ±0.01	0.044 ±0.011	2.248 ±0.136
[5] 02-02/0.0-0.0	0.05 ±0.014	0.086 ±0.01	0.006 ±0.002	0.446 ±0.048	0.008 ±0.002	8.173 ±0.444	0.422 ±0.032	0.57 ±0.055	33.206 ±1.324	0.073 ±0.019	0.087 ±0.021	0.082 ±0.023	4.085 ±0.409
[6] 02-02/0.0-0.0	0.063 ±0.017	0.112 ±0.023	0.008 ±0.002	0.602 ±0.051	0.011 ±0.011	12.176 ±0.447	0.621 ±0.041	1.295 ±0.07	1.166 ±0.046	0.109 ±0.029	0.12 ±0.034	0.118 ±0.035	6.088 ±0.417

Table 102 – Results of experiment D2 - Average and standard deviation of ARI values

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)			Group 4 (Y)		
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0.001	1 ±0	1 ±0.001	1 ±0.001	1 ±0	1 ±0.001	1 ±0	1 ±0.001	1 ±0.001	1 ±0	1 ±0
[2] 20-20/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0
[3] 200-200/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0	1 ±0

Table 103 – Results of experiment D2 - ARPD values considering the ARI values of each repetition

Dataset [ID] Scenario	Group 1 (K)				Group 2 (W)				Group 3 (P)				Group 4 (Y)						
	KAMLA	FMMMD	KMDBS	KMDBR	KAMLA	FMMMD	WKMBs	WKMMs	WKMBR	KAMLA	FMMMD	PMGBS	PMGBR	KAMLA	FMMMD	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-02/0.0-0.0	0	0	0.008	0	0	0	0.008	0.008	0	0	0	0.008	0	0	0	0.008	0.008	0	0
[2] 20-20/0.0-0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[3] 200-200/0.0-0.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 104 – Results of experiment D2 - Average and standard deviation of numerical attributes weights

Dataset [ID] Scenario	All groups		Group 1 (K)		Group 2 (W)			Group 3 (P)			Group 4 (Y)		
	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-02/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.332 ±0.12	0.883 ±0.012	0.5 ±0	0.883 ±0.012	1 ±0	1 ±0	1 ±0	1 ±0
[2] 20-20/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.208 ±0.131	0.902 ±0.006	0.5 ±0	0.902 ±0.006	1 ±0	1 ±0	1 ±0	1 ±0
[3] 200-200/0.0-0.0	1 ±0	1 ±0	1 ±0	1 ±0	0.5 ±0	0.198 ±0.105	0.918 ±0.005	0.5 ±0	0.918 ±0.005	1 ±0	1 ±0	1 ±0	1 ±0

Table 105 – Results of experiment D2 - Average and standard deviation of categorical attributes weights

Dataset	All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
[ID] Scenario	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTBL	KPTBR
[1] 02-02/0.0-0.0	1	1	1	0.117	0.5	0.668	0.117	0.5	0.117	1	1	0.121	0.117
	± 0	± 0	± 0	± 0.012	± 0	± 0.12	± 0.012	± 0	± 0.012	± 0	± 0	± 0.011	± 0.012
[2] 20-20/0.0-0.0	1	1	1	0.098	0.5	0.792	0.098	0.5	0.098	1	1	0.099	0.098
	± 0	± 0	± 0	± 0.006	± 0	± 0.131	± 0.006	± 0	± 0.006	± 0	± 0	± 0.006	± 0.006
[3] 200-200/0.0-0.0	1	1	1	0.082	0.5	0.802	0.082	0.5	0.082	1	1	0.082	0.082
	± 0	± 0	± 0	± 0.005	± 0	± 0.105	± 0.005	± 0	± 0.005	± 0	± 0	± 0.005	± 0.005

Table 106 – Results of experiment D2 - Average and standard deviation of execution time

Dataset	All groups		Group 1 (K)		Group 2 (W)		Group 3 (P)		Group 4 (Y)				
[ID] Scenario	KAMLA	FMMMD	KMDBS	KMDBR	WKMBs	WKMMs	WKMBR	PMGBS	PMGBR	KPTBS	KPTNM	KPTL	KPTBR
[1] 02-02/0.0-0.0	0.029	0.044	0.003	0.179	0.003	2.129	0.211	0.062	2.397	0.032	0.035	0.029	1.456
	± 0.01	± 0.013	± 0.001	± 0.018	± 0.001	± 0.426	± 0.022	± 0.045	± 0.202	± 0.017	± 0.01	± 0.007	± 0.182
[2] 20-20/0.0-0.0	0.053	0.098	0.015	1.045	0.021	11.688	1.014	0.078	4.229	0.103	0.104	0.104	4.678
	± 0.015	± 0.013	± 0.01	± 0.115	± 0.004	± 0.671	± 0.058	± 0.036	± 0.243	± 0.025	± 0.025	± 0.026	± 0.153
[3] 200-200/0.0-0.0	0.361	0.689	0.222	14.039	0.238	1.874	14.87	0.487	26.155	0.853	0.86	0.851	43.892
	± 0.053	± 0.077	± 0.04	± 0.299	± 0.045	± 0.083	± 0.831	± 0.062	± 0.278	± 0.127	± 0.118	± 0.119	± 1.547

(a) Experiment A1

(b) Experiment A2

(c) Experiment B1

(d) Experiment B2

(e) Experiment C1

(f) Experiment C2

Figure 50 – Group 1

(a) Experiment A1

(b) Experiment A2

(c) Experiment B1

(d) Experiment B2

(e) Experiment C1

(f) Experiment C2

Figure 51 – Group 2

(a) Experiment A1

(b) Experiment A2

(c) Experiment B1

(d) Experiment B2

(e) Experiment C1

(f) Experiment C2

Figure 52 – Group 3

(a) Experiment A1

(b) Experiment A2

(c) Experiment B1

(d) Experiment B2

(e) Experiment C1

(f) Experiment C2

Figure 53 – Group 4

(a) Group 1

(b) Group 2

(c) Group 3

(d) Group 4

Figure 54 – Execution time in different numbers of instances

Figure 55 – Execution time in different numbers of attributes

E APPENDIX E

Feature weighting

E.1 Parameters

Tables from 107 to 109 show the values of the parameters in each experiment.

Table 107 – Baseline methods - Parameters' values defined by irace method

Experiment	GA				BRKGA				
	n_{pop}	p_c	p_m	p_{mm}	n_{pop}	p_e	p_m	p_{mm}	ρ
A	14	0.3	0.7	0.3	15	0.1	0.5	0.9	0.8
B	15	0.2	0.9	0.4	14	0.2	0.3	0.7	0.5
C	13	1	0.7	0.3	14	0.2	0.5	0.9	0.4

Table 108 – Proposed methods, Part 1 - Parameters' values defined by irace method

Experiment	ECS								ILS		
	n_{pop}	p_c	p_m	$n_{Clusters}$	$PMILS1$	$PMILS2$	$PMILS31$	$PMILS32$	n_{pop}	p_{IndILS}	$p_{AttrILS}$
A	14	0.9	0.2	3	0.5	0.4	0.9	0.4	11	0.4	0.9
B	14	0.6	0.9	2	0.6	0.6	0.6	0.8	11	0.9	0.3
C	4411	0.6	0.2	3	0.3	0.6	0.9	0.7	9	0.8	0.7

Table 109 – Proposed methods, Part 2 - Parameters' values defined by irace method

Experiment	ECSBR								
	n_{pop}	p_c	p_m	ρ	$n_{Clusters}$	$PMILS1$	$PMILS2$	$PMILS31$	$PMILS32$
A	13	0.2	0.4	0.7	3	0.6	0.4	1	0.5
B	14	0.2	0.5	0.5	2	0.9	0.5	0.9	0.6
C	13	0.3	0.5	0.3	5	0.4	0.9	1	0.5

E.2 Results

In the following subsections, we show the execution time values of each experiment. Note that in all tables the bold values are statistically better than the other ones which respect to the solution quality (ARI).

E.2.1 Experiment A

E.2.2 Experiment B

E.2.3 Experiment C

Table 110 – Results of experiment A - Average and standard deviation of the execution time.

Dataset scenario	Baseline			Proposed algorithms		
	KAMBSLN	KAMGAFW	KAMBRKG	KAMILSF	KAMECSF	KAMECBR
[1] 05-10/00-00	0.046 ±0.012	2.636 ±0.187	2.927 ±0.209	1.301 ±0.092	5.623 ±0.444	5.204 ± 0.628
[2] 05-10/04-02	0.043 ±0.011	2.721 ±0.125	2.938 ±0.155	1.375 ±0.117	5.817 ±0.55	5.414 ± 0.766
[3] 05-10/02-04	0.036 ±0.004	2.829 ±0.191	3.031 ±0.087	1.392 ±0.066	6.066 ±0.557	5.678 ± 0.508
[4] 05-10/01-08	0.04 ±0.005	3.71 ±0.229	3.87 ±0.148	1.828 ±0.088	7.691 ±0.499	6.715 ± 0.61
[5] 10-05/00-00	0.034 ±0.005	2.591 ±0.163	2.806 ±0.115	1.234 ±0.038	5.299 ±0.215	5.072 ± 0.244
[6] 10-05/08-01	0.038 ±0.008	2.715 ±0.109	2.951 ±0.099	1.547 ±0.228	5.816 ±0.352	5.454 ± 0.298
[7] 10-05/04-02	0.036 ±0.005	3.132 ±0.328	3.351 ±0.267	1.634 ±0.192	6.352 ±0.469	5.426 ± 0.261
[8] 10-05/02-04	0.038 ±0.003	3.73 ±0.244	3.862 ±0.222	1.816 ±0.144	7.291 ±0.424	6.424 ± 0.57
[9] 05-05/00-00	0.029 ±0.003	2.302 ±0.039	2.53 ±0.054	1.137 ±0.032	4.886 ±0.148	4.584 ± 0.116
[10] 05-05/04-01	0.035 ±0.004	2.514 ±0.076	3.236 ±0.624	1.371 ±0.223	5.738 ±0.474	5.547 ± 0.714
[11] 05-05/02-02	0.042 ±0.005	3.09 ±0.26	3.269 ±0.272	1.509 ±0.17	6.257 ±0.588	5.625 ± 0.638
[12] 05-05/01-04	0.044 ±0.01	4.055 ±0.632	4.022 ±0.517	1.702 ±0.098	6.962 ±0.44	6.234 ± 0.451
[13] 50-100/00-00	0.097 ±0.01	7.683 ±0.113	8.488 ±0.096	3.728 ±0.084	15.68 ±0.6	15.022 ± 0.319
[14] 50-100/40-20	0.098 ±0.008	7.637 ±0.097	8.458 ±0.066	3.72 ±0.066	15.785 ±0.497	15.029 ± 0.162
[15] 50-100/20-40	0.095 ±0.009	7.577 ±0.064	8.393 ±0.07	3.687 ±0.058	15.587 ±0.426	15.819 ± 0.789
[16] 50-100/10-80	0.096 ±0.008	7.732 ±0.058	8.573 ±0.067	3.779 ±0.071	16.015 ±0.397	15.243 ± 0.186
[17] 100-50/00-00	0.092 ±0.007	7.317 ±0.065	8.135 ±0.078	3.56 ±0.052	15.027 ±0.42	14.501 ± 0.191
[18] 100-50/80-10	0.092 ±0.006	7.396 ±0.063	8.212 ±0.05	3.604 ±0.064	15.226 ±0.401	14.677 ± 0.309
[19] 100-50/40-20	0.093 ±0.006	7.435 ±0.088	8.252 ±0.058	3.626 ±0.052	15.314 ±0.38	14.691 ± 0.169
[20] 100-50/20-40	0.092 ±0.006	7.818 ±0.091	8.739 ±0.207	3.812 ±0.062	16.211 ±0.382	15.477 ± 0.275
[21] 50-50/00-00	0.068 ±0.007	5.478 ±0.054	6.09 ±0.077	2.677 ±0.049	11.298 ±0.269	10.913 ± 0.297
[22] 50-50/40-10	0.071 ±0.008	5.557 ±0.058	6.161 ±0.076	2.703 ±0.033	11.532 ±0.336	11.153 ± 0.146
[23] 50-50/20-20	0.073 ±0.004	5.7 ±0.063	6.325 ±0.072	2.774 ±0.039	11.787 ±0.304	11.5 ± 0.828
[24] 50-50/10-40	0.074 ±0.008	6.08 ±0.088	6.729 ±0.097	2.958 ±0.069	12.528 ±0.352	12.047 ± 0.24

Table 111 – Results of experiment B - Average and standard deviation of the execution time.

Dataset scenario	Baseline			Proposed algorithms		
	KAMBSLN	KAMGAFW	KAMBRKG	KAMILSF	KAMECSF	KAMECBR
[1] 05-10/00-00	0.044	3.282	3.269	2.879	5.503	6.717
	± 0.01	± 0.222	± 0.474	± 0.347	± 0.375	± 0.734
[2] 05-10/04-02	0.045	3.285	3.185	2.583	5.839	6.579
	± 0.011	± 0.232	± 0.242	± 0.166	± 0.573	± 0.723
[3] 05-10/02-04	0.05	3.287	3.158	2.587	5.563	6.841
	± 0.013	± 0.149	± 0.187	± 0.132	± 0.298	± 0.692
[4] 05-10/01-08	0.062	4.383	4.155	3.444	7.374	8.672
	± 0.014	± 0.267	± 0.308	± 0.228	± 0.489	± 0.946
[5] 10-05/00-00	0.048	2.97	2.859	2.348	5.045	6.01
	± 0.013	± 0.12	± 0.12	± 0.094	± 0.181	± 0.315
[6] 10-05/08-01	0.049	3.058	2.94	2.426	5.213	6.134
	± 0.014	± 0.123	± 0.138	± 0.104	± 0.241	± 0.356
[7] 10-05/04-02	0.049	3.177	3.019	2.493	5.395	6.274
	± 0.013	± 0.122	± 0.156	± 0.094	± 0.28	± 0.39
[8] 10-05/02-04	0.06	3.566	3.309	2.885	6.174	7.476
	± 0.019	± 0.168	± 0.205	± 0.188	± 0.48	± 0.954
[9] 05-05/00-00	0.049	2.922	2.83	2.331	4.981	6.101
	± 0.015	± 0.153	± 0.221	± 0.138	± 0.25	± 0.732
[10] 05-05/04-01	0.05	3.001	2.898	2.385	5.131	6.311
	± 0.014	± 0.163	± 0.226	± 0.128	± 0.297	± 0.716
[11] 05-05/02-02	0.054	3.125	3.001	2.475	5.327	6.432
	± 0.016	± 0.161	± 0.178	± 0.132	± 0.309	± 0.647
[12] 05-05/01-04	0.064	3.505	3.311	2.733	5.954	6.912
	± 0.02	± 0.208	± 0.249	± 0.19	± 0.41	± 0.749
[13] 50-100/00-00	0.092	8.031	7.798	6.319	13.326	15.991
	± 0.024	± 0.217	± 0.414	± 0.187	± 0.546	± 1.255
[14] 50-100/40-20	0.093	7.995	7.794	6.269	13.405	16.101
	± 0.019	± 0.244	± 0.279	± 0.174	± 0.459	± 0.973
[15] 50-100/20-40	0.094	8.013	7.788	6.334	13.517	16.13
	± 0.014	± 0.168	± 0.25	± 0.179	± 0.444	± 0.81
[16] 50-100/10-80	0.102	8.238	8.081	6.501	13.89	16.702
	± 0.023	± 0.182	± 0.321	± 0.181	± 0.518	± 0.985
[17] 100-50/00-00	0.089	7.603	7.391	5.972	12.755	15.314
	± 0.017	± 0.252	± 0.276	± 0.186	± 0.472	± 0.881
[18] 100-50/80-10	0.091	7.671	7.805	6.336	13.406	17.027
	± 0.023	± 0.218	± 0.646	± 0.483	± 0.872	± 1.676
[19] 100-50/40-20	0.094	7.977	7.705	6.119	13.013	15.715
	± 0.017	± 0.477	± 0.666	± 0.251	± 0.437	± 0.9
[20] 100-50/20-40	0.101	8.209	7.956	6.495	13.833	16.615
	± 0.02	± 0.237	± 0.314	± 0.234	± 0.518	± 0.855
[21] 50-50/00-00	0.072	5.896	5.785	4.631	9.955	12.023
	± 0.023	± 0.218	± 0.355	± 0.131	± 0.429	± 1.061
[22] 50-50/40-10	0.076	5.987	5.837	4.723	10.049	12.098
	± 0.022	± 0.181	± 0.287	± 0.197	± 0.472	± 0.953
[23] 50-50/20-20	0.075	6.049	5.861	4.749	10.053	12.042
	± 0.016	± 0.164	± 0.278	± 0.167	± 0.41	± 0.763
[24] 50-50/10-40	0.08	6.308	6.126	4.961	10.622	14.112
	± 0.031	± 0.16	± 0.284	± 0.139	± 0.416	± 1.527

Table 112 – Results of experiment C - Average and standard deviation of the execution time.

Dataset scenario	Baseline			Proposed algorithms		
	KAMBSLN	KAMGAFW	KAMBRKG	KAMILSF	KAMECSF	KAMECBR
[1] 05-05/02-02	0.046	3.084	3.288	2.386	5.383	5.662
	± 0.007	± 0.24	± 0.149	± 0.347	± 0.482	$\pm \mathbf{0.338}$
[2] 05-05/02-02	0.055	3.636	4.086	2.886	7.143	7.536
	± 0.008	± 0.165	± 0.217	± 0.231	± 0.679	$\pm \mathbf{0.702}$
[3] 05-05/02-02	0.085	5.06	5.566	3.925	8.611	10.211
	± 0.016	± 0.377	± 0.413	± 0.278	± 0.558	$\pm \mathbf{0.985}$
[4] 05-05/02-02	0.092	6.55	6.989	4.605	10.798	11.149
	± 0.015	± 0.506	± 0.517	± 0.271	± 0.453	$\pm \mathbf{0.443}$
[5] 05-05/02-02	0.112	8.303	8.937	6.07	14.857	15.327
	± 0.015	± 0.426	± 0.418	± 0.21	± 0.956	$\pm \mathbf{0.893}$