SELMA TEREZINHA MILAGRE

**Análise do Número de Grupos em Bases de
Dados Incompletas Utilizando Agrupamen-
tos Nebulosos e Reamostragem *Bootstrap***

Tese apresentada à Escola de Engenharia de São Carlos,
da Universidade de São Paulo, como parte dos requisi-
tos para obtenção do Título de Doutor em Engenharia
Elétrica.

Área de Concentração: Sistemas Dinâmicos
Orientador: Prof. Dr. Carlos Dias Maciel

São Carlos
2008

FOLHA DE APROVAÇÃO

Selma Terezinha Milagre

Análise do Número de Grupos em Bases de Dados Incompletas Utilizando
Agrupamentos Nebulosos e Reamostragem *Bootstrap*

Tese apresentada à Escola de Engenharia de
São Carlos, da Universidade de São Paulo,
como parte dos requisitos para obtenção do
Título de Doutor em Engenharia Elétrica.
Área de Concentração: Sistemas Dinâmicos

Aprovado em:

Banca Examinadora

Prof. Dr. José Carlos Pereira

Instituição: Universidade de São Paulo

Prof. Dr. José Carlos Pereira

Instituição: Universidade de São Paulo

Prof. Dr. José Carlos Pereira

Instituição: Universidade de São Paulo

Prof. Dr. José Carlos Pereira

Instituição: Universidade de São Paulo

Prof. Dr. José Carlos Pereira

Instituição: Universidade de São Paulo

Dedico esta tese ao meu esposo Adriano, por seu apoio incansável, sua dedicação e seu grande amor;

Aos meus filhos Caroline e Adriano Júnior, pelo amor, paciência e compreensão pelos tão longos períodos de minha ausência.

Agradecimentos

Agradeço a DEUS por ter permitido que eu realizasse mais esse sonho;

Ao meu esposo por todas as horas em que esteve ao meu lado me incentivando,
estudando comigo e me ajudando a decifrar os caminhos a seguir;

Aos meus pais, Geraldo e Marina, por todo esforço e dedicação;

Ao meu orientador, Dr. Carlos Dias Maciel, que sempre é mais que um orientador e hoje
é um grande amigo;

Ao professor Dr. José Carlos Pereira, pelas palavras de incentivo que vieram na hora
certa e a quem eu aprendi a admirar e respeitar;

À Ivonete, Maísa e Diana por sempre terem me recebido com tanto carinho;

Ao amigo Paulo Scalassara que tantas vezes me auxiliou;

Aos amigos do laboratório, Eugênia, Soledad, Edwin, Julian e Mariana, que não
deixaram que a pouca convivência impedisse que crescesse uma grande amizade;

A todos os funcionários da USP São Carlos com os quais tive o prazer de conviver;

A todos que torceram por mim.

*Se vi mais longe, foi porque
estava nos ombros de gigantes*
Isaac Newton (1642-1727)

Resumo

MILAGRE, S. T. **Análise do Número de Grupos em Bases de Dados Incompletas Utilizando Agrupamentos Nebulosos e Reamostragem Bootstrap.** 157f. Tese (doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2008.

A técnica de agrupamento de dados é amplamente utilizada em análise exploratória, a qual é freqüentemente necessária em diversas áreas de pesquisa tais como medicina, biologia e estatística, para avaliar potenciais hipóteses a serem utilizadas em estudos subseqüentes. Em bases de dados reais, a ocorrência de dados incompletos, nos quais os valores de um ou mais atributos do dado são desconhecidos, é bastante comum. Este trabalho apresenta um método capaz de identificar o número de grupos presentes em bases de dados incompletas, utilizando a combinação das técnicas de agrupamentos nebulosos e reamostragem *bootstrap*. A qualidade da classificação é baseada em medidas de comparação tradicionais como F_1 , Classificação Cruzada, Hubert e outras. Os estudos foram feitos em oito bases de dados. As quatro primeiras são bases de dados artificiais, a quinta e a sexta são a *wine* e *iris*. A sétima e oitava bases são formadas por uma coleção brasileira de 119 estirpes de *Bradyrhizobium*. Para avaliar toda informação sem introduzir estimativas, fez-se a modificação do algoritmo *Fuzzy C-Means* (FCM) utilizando-se um vetor de índices de atributos, os quais indicam onde o valor de um atributo é observado ou não, modificando-se então, os cálculos do centro e distância ao centro. As simulações foram feitas de 2 até 8 grupos utilizando-se 100 sub-amostras. Os percentuais de valores faltando utilizados foram 2%, 5%, 10%, 20% e 30%. Os resultados deste trabalho demonstraram que nosso método é capaz de identificar partições relevantes, até em presença de altos índices de dados incompletos, sem a necessidade de se fazer nenhuma suposição sobre a base de dados. As medidas Hubert e Índice Randômico Ajustado encontraram os melhores resultados experimentais.

Palavras-chave: Dados Incompletos, *Fuzzy c-Means*, Agrupamento de Dados, *Bootstrap*.

Abstract

MILAGRE, S. T. **Análise do Número de Grupos em Bases de Dados Incompletas Utilizando Agrupamentos Nebulosos e Reamostragem Bootstrap.** 157f. Tese (doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2008.

Clustering in Exploratory data analysis is often necessary in several areas of the survey such as medicine, biology and statistics, to evaluate potential hypotheses for subsequent studies. In real datasets the occurrence of incompleteness, where the values of some of the attributes are unknown, is very common. This work presents a method capable to identifying the number of clusters present in incomplete datasets, using a combination of the fuzzy clustering and resampling (bootstrapping). The quality of classification is based on the traditional measures, like F_1 , Cross-Classification, Hubert and others. The studies were made on eight datasets. The first four are artificial datasets, the fifth and sixth are the wine and iris datasets. The seventh and eighth databases are composed of the Brazilian collection of 119 *Bradyrhizobium* strains. To evaluate all information without introducing estimates, a modification of the Fuzzy C-Means (FCM) algorithm was developed using an index vector of attributes, which indicates whether an attribute value is observed or not, and changing the center and distance calculations. The simulations were made from 2 to 8 clusters using 100 sub-samples. The percentage of the missing values used were 2%, 5%, 10%, 20% and 30%. Even lacking data and with no special requirements of the database, the results of this work demonstrate that the proposed method is capable to identifying relevant partitions. The best experimental results were found using Hubert and Corrected Randomness measures.

Keywords: Missing Values, Fuzzy c-Means, Clustering, Bootstrap.

Sumário

LISTA DE FIGURAS	xiii
LISTA DE TABELAS	xvii
LISTA DE ABREVIATURAS E SIGLAS	xxi
LISTA DE SÍMBOLOS	xxiii
1 INTRODUÇÃO	25
2 DADOS INCOMPLETOS	31
2.1 Tipos de Dados Imperfeitos	32
2.2 Tipos de Dados Incompletos	33
2.2.1 Dados MCAR	34
2.2.2 Dados MAR	35
2.2.3 Dados NMAR	36
2.3 Métodos para Tratar Dados Incompletos	37
2.3.1 Métodos que Ignoram Observações Incompletas	37
2.3.2 Métodos de Inserção Única	40
2.4 Métodos de Estimativas de Parâmetros	44
2.4.1 Método Maximização da Expectativa	44
2.4.2 Inserção Múltipla	45
2.5 Conclusão	46
3 AGRUPAMENTO DE DADOS	47
3.1 Critérios de Validação	49
3.2 Definições e Notações	50
3.3 Lógica Nebulosa	51
3.3.1 Conjuntos Nebulosos	53
3.3.2 Operações em Conjuntos Nebulosos	54
3.3.3 Normas Triangulares	56
3.3.4 Agrupamentos Nebulosos	57

3.3.5	Algoritmo <i>Fuzzy C-Means</i> (FCM)	60
3.4	Análise de Dados Incompletos Utilizando FCM	65
3.5	Medindo a Validade do Agrupamento	70
3.5.1	Comparando Matrizes de Partição	72
3.5.2	Comparando Matrizes de Coincidência	78
3.6	Reamostragem <i>Bootstrap</i>	82
3.7	Conclusão	85
4	MATERIAIS E MÉTODOS	87
4.1	Bases de Dados Utilizadas	87
4.1.1	Redução de Dimensionalidade	88
4.1.2	Bases de Dados de Validação	90
4.1.3	Base de Dados de Aplicação: <i>Bradyrhizobium</i>	97
4.2	O Método	101
4.2.1	Comparando Matrizes de Partição	108
4.2.2	Comparando Matrizes de Coincidência	116
5	RESULTADOS E DISCUSSÃO	123
5.1	Tempos de Processamento	124
5.2	Resultados nas Bases de Dados de Validação	125
5.3	Resultados nas Bases de Dados de Aplicação	132
5.4	Resumo dos Resultados Obtidos	140
6	CONCLUSÃO	145
	REFERÊNCIAS BIBLIOGRÁFICAS	149

Lista de Figuras

FIGURA 3.1 – <i>t-norma</i> Zadeh (Modificada de (JAFELICE; BARROS; BASSA-NEZI, 2005)).	58
FIGURA 3.2 – <i>t-norma</i> produto (Modificada de (JAFELICE; BARROS; BASSA-NEZI, 2005)).	58
FIGURA 3.3 – Agrupamentos nebulosos (Modificada de (JAIN; MURTY; FLYNN, 1999)).	59
FIGURA 4.1 – Dados artificiais bidimensionais (Artificial 1) contendo 400 dados e 4 grupos, sendo que cada grupo contém 100 dados.	91
FIGURA 4.2 – Dados artificiais bidimensionais (Artificial 2) contendo 350 dados e 7 grupos, sendo que cada grupo contém 50 dados.	92
FIGURA 4.3 – Análise de autovalores e variância da matriz de vetores resultantes de MDS (LLERENA, 2008) referentes à base de dados Artificial 3.	93
FIGURA 4.4 – Dados artificiais (Artificial 3) contendo 2000 dados, 4 grupos e 4 atributos, sendo que cada grupo contém 492, 482, 523 e 503 dados.	94
FIGURA 4.5 – Dados artificiais (Artificial 4) contendo 1000 dados, 5 grupos e 3 atributos, sendo que cada grupo contém 190, 207, 209 e 206 e 188 dados.	95
FIGURA 4.6 – Dados <i>wine</i> (BALKE; MERZ, 1998) contendo 178 dados e 3 grupos.	96
FIGURA 4.7 – Dados <i>iris</i> (BALKE; MERZ, 1998) contendo 150 dados e 3 grupos, dos quais somente um é linearmente separável.	97

FIGURA 4.8 – Análise de autovalores e variância da matriz de vetores resultantes de MDS (LLERENA, 2008) referentes à base de dados <i>iris</i> (BALKE; MERZ, 1998).	97
FIGURA 4.9 – Análise de autovalores e variância da matriz de vetores resultantes de MDS (LLERENA, 2008) referentes à base de dados Brady 1. . .	100
FIGURA 4.10 – Base de dados Brady 1 (LLERENA, 2008) contendo 119 dados e considerando os três primeiros atributos.	101
FIGURA 4.11 – Análise de autovalores e variância da matriz de vetores resultantes de MDS (LLERENA, 2008) referentes à base de dados Brady 2. . .	102
FIGURA 4.12 – Dados Brady 2 (LLERENA, 2008) contendo 119 dados e considerando os três primeiros atributos.	102
FIGURA 5.1 – Relação dos tempos médios de CPU, em minutos, gastos para as bases de dados Artificial 1, Artificial 2, iris, wine, Brady 1 e Brady 2 (descritas na Tabela 4.1), com número de grupos variando de 2 a 8 ($c = 2, 3, \dots, 8$).	126
FIGURA 5.2 – Relação dos tempos médios de CPU, em minutos, gastos para as bases de dados Artificial 3 e Artificial 4 (descritas na Tabela 4.1), com número de grupos variando de 2 a 8 ($c = 2, 3, \dots, 8$).	126
FIGURA 5.3 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Artificial 1, contendo 400 dados, 2 atributos e 4 grupos.	127
FIGURA 5.4 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Artificial 2, contendo 350 dados, 2 atributos e 7 grupos.	128
FIGURA 5.5 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Artificial 3, contendo 2000 dados, 4 atributos e 4 grupos.	130

- FIGURA 5.6 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados *iris*(BALKE; MERZ, 1998) contendo 150 dados, 4 atributos e 3 grupos (dos quais somente um é linearmente separável). 131
- FIGURA 5.7 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Artificial 4, contendo 1000 dados, 3 atributos e 5 grupos. 132
- FIGURA 5.8 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados *wine*(BALKE; MERZ, 1998) contendo 178 dados, 13 atributos (neste trabalho utilizou-se somente os atributos 7, 10 e 13) e 3 grupos. 133
- FIGURA 5.9 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Brady 1 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006), com 119 dados e 49 atributos (neste trabalho utilizou-se somente os atributos de 1 a 20). 134
- FIGURA 5.10 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Brady 1 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006), com 119 dados e 49 atributos (neste trabalho utilizou-se somente os atributos de 1 a 20). 135
- FIGURA 5.11 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Brady 2 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006), com 119 dados e 38 atributos (neste trabalho utilizou-se somente os atributos de 1 a 15). 136
- FIGURA 5.12 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Brady 2 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006), com 119 dados e 38 atributos (neste trabalho utilizou-se somente os atributos de 1 a 15). 137
- FIGURA 5.13 – Gel de eletroforese da região 16S rRNA das estirpes SEMIA (026 a 050) digeridas com enzima de restrição *Cfo* I (MILAGRE, 2003). . . 139

FIGURA 5.14 –Relação dos percentuais de resultados corretos obtidos para cada
medida de comparação em relação às simulações feitas, com número
de grupos variando de 2 a 8 ($c = 2, 3, \dots, 8$) e percentuais de valores
faltando iguais a 2%, 5%, 10%, 20% e 30%. 141

Lista de Tabelas

TABELA 2.1 – Tabela contendo exemplo de uma base de dados multivariada contendo p atributos sobre n dados, alguns dos quais são incompletos, onde a inexistência do valor para o atributo é representado por NaN (Adaptado de (SCHAFER, 1999)).	34
TABELA 3.1 – Tabela referente ao conjunto total de dados $\mathbf{X}$, contendo p atributos sobre n dados.	51
TABELA 3.2 – Relação das principais t -normas.	57
TABELA 3.3 – Tabela de contingência para comparar duas linhas de duas matrizes de partição <i>crisp</i> (Adaptado de (BORGELT, 2005)).	72
TABELA 3.4 – Tabela de contingência para o cálculo de <i>precisão</i> e <i>recuperação</i> (Adaptado de (BORGELT, 2005)).	75
TABELA 4.1 – Sumário das características das oito bases de dados utilizadas neste trabalho. Onde o símbolo '?' indica <i>em estudo</i>	89
TABELA 4.2 – Parâmetros das quatro distribuições gaussianas utilizadas na construção da base de dados Artificial 3 (Adaptado de (TALAVERA, 2007)).	93
TABELA 4.3 – Parâmetros das cinco distribuições gaussianas utilizadas na construção da base de dados Artificial 4 (Adaptado de (TALAVERA, 2007)).	94

TABELA 4.4 – Amostras dos 350 dados que constituem a base de dados Artificial 2, já normalizada e contendo dados incompletos (30% de valores faltando). Onde o NaN indica que o atributo não possui valor observado.	104
TABELA 4.5 – Amostras dos 178 dados da <i>wine</i> (BALKE; MERZ, 1998), já normalizada e contendo dados incompletos (20% de valores faltando). Onde o NaN indica que o atributo não possui valor observado. . . .	104
TABELA 5.1 – Relação dos tempos de CPU gastos na execução do algoritmo completo (Algoritmo 4.1) para cada base de dados (descritas na Tabela 4.1), com número de grupos variando de 2 a 8 ($c = 2, 3, \dots, 8$). . . .	125
TABELA 5.2 – Resultados das simulações para a base de dados Artificial 1, contendo 400 dados, 2 atributos e 4 grupos. O percentual de valores faltando é de 2%. Todas as medidas foram calculadas usando <i>t-norma</i> produto e <i>nil-potent</i> mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.	127
TABELA 5.3 – Resultados das simulações para a base de dados Artificial 2, contendo 350 dados, 2 atributos e 7 grupos. O percentual de valores faltando é de 20%. Todas as medidas foram calculadas usando <i>t-norma</i> produto e <i>nil-potent</i> mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.	128
TABELA 5.4 – Resultados das simulações para a base de dados Artificial 3, contendo 2000 dados, 4 atributos e 4 grupos. O percentual de valores faltando é de 30%. Todas as medidas foram calculadas usando <i>t-norma</i> produto e <i>nil-potent</i> mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.	129

- TABELA 5.5 – Resultados das simulações feitas para a base de dados *iris*(BALKE; MERZ, 1998) contendo 150 dados, 4 atributos e 3 grupos (dos quais somente um é linearmente separável). O percentual de valores faltando é de 5%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida. 131
- TABELA 5.6 – Resultados das simulações para a base de dados Artificial 4, contendo 1000 dados, 3 atributos e 5 grupos. O percentual de valores faltando é de 30%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida. 132
- TABELA 5.7 – Resultados das simulações para a base de dados *wine* (BALKE; MERZ, 1998) contendo 178 dados, 13 atributos (neste trabalho utilizou-se somente os atributos 7, 10 e 13) e 3 grupos. O percentual de valores faltando é de 5%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida. 133
- TABELA 5.8 – Resultados das simulações feitas para a base de dados Brady 1 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006) contendo 119 dados e 49 atributos (neste trabalho utilizou-se somente os atributos de 1 a 20). O percentual de valores faltando é de 2%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida. 134
- TABELA 5.9 – Resultados das simulações feitas para a base de dados Brady 1 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006) contendo 119 dados e 49 atributos (neste trabalho utilizou-se somente os atributos de 1 a 20). O percentual de valores faltando é de 5%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida. 135

- TABELA 5.10 –Resultados das simulações feitas para a base de dados Brady 2 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006) contendo 119 dados e 38 atributos (neste trabalho utilizou-se somente os atributos de 1 a 15). O percentual de valores faltando é de 2%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.136
- TABELA 5.11 –Resultados das simulações feitas para a base de dados Brady 2 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006) contendo 119 dados e 38 atributos (neste trabalho utilizou-se somente os atributos de 1 a 15). O percentual de valores faltando é de 20%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.137
- TABELA 5.12 –Resumo do número de agrupamentos obtidos para as bases de dados Brady 1 e Brady 2, referentes às medidas Hubert e Índice Randômico Ajustado. 138
- TABELA 5.13 –Resumo dos métodos de análise da base de dados *Bradyrhizobium*. . 140
- TABELA 5.14 –Sumário dos resultados das simulações feitas neste trabalho para os percentuais de valores faltando de 2%, 5% e 10%. Os *pontos* em negrito representam onde a medida atingiu o seu valor máximo coincidindo com o número correto de grupos existente na base de dados. Pode-se observar que somente as medidas Hubert e Randômico Ajustado (Rand. Aj.) apresentaram o número correto de grupos em todas as simulações. 142
- TABELA 5.15 –Sumário dos resultados das simulações feitas neste trabalho para os percentuais de valores faltando de 20% e 30%. Os *pontos* em negrito representam onde a medida atingiu o seu valor máximo coincidindo com o número correto de grupos existente na base de dados. Pode-se observar que somente as medidas Hubert e Randômico Ajustado (Rand. Aj.) apresentaram o número correto de grupos em todas as simulações. 143

Lista de Abreviaturas e Siglas

MCAR	<i>Missing Completely at Random</i>
MAR	<i>Missing at Random</i>
NMAR	<i>Not Missing at Random</i>
LD	<i>Listwise Deletion</i>
CD	<i>Case Deletion</i>
AC	<i>Available Case Analysis</i>
MI	<i>Multiple Imputation</i>
EM	<i>Expectation Maximization</i>
ML	<i>Maximum Likelihood</i>
FCM	<i>Fuzzy C-Means</i>
WDS	<i>Whole Data Strategy</i>
WDSFCM	<i>Whole Data Strategy for Fuzzy C-Means</i>
OCS	<i>Optimal Completion Strategy</i>
OCSFCM	<i>Optimal Completion Strategy for Fuzzy C-Means</i>
NPS	<i>Nearest Prototype Strategy</i>
PDS	<i>Partial Distance Strategy</i>
PDSFCM	<i>Partial Distance Strategy for Fuzzy C-Means</i>
NaN	<i>Not-a-Number</i>

Diff.	Erro Médio Quadrático
Acc.	Classificação Cruzada
Rand.	Índice Randômico
Rand. Aj.	Índice Randômico Ajustado
F&M	Índice Fowlkes and Mallows
MDS	<i>Multidimensional Scaling</i>
Atrib.	Atributo
Nro	Número

Lista de Símbolos

$\mathbf{X}$	base de dados
$\mathbf{x}$	dado
n	número de dados na base de dados
p	número de atributos que compõem o dado
c	número de grupos nebulosos
w	índice do atributo
k	índice do atributo
i	índice do grupo
r	índice do grupo
j	índice do dado
l	índice do dado
f	índice do percentual de dados faltando
μ	grau de pertinência do dado dentro do grupo
μ_{ij}	grau de pertinência do dado j no grupo i
d_{ij}	distância do dado j ao centro do grupo i
$\top$	t -norma
$\top_{minnp}$	t -norma <i>nil-potent</i> mínimo
$\top_{prod}$	t -norma produto
$\mathbb{R}^p$	espaço p -dimensional

$\mathbf{U}$	matriz de graus de pertinência nebulosos
$\mathbf{V}$	matriz de centros dos grupos nebulosos
J_m	função objetivo do algoritmo <i>fuzzy c-means</i>
m	fator de fuzzificação
t	número de iterações
$\mathbf{v}$	matriz que contém os centros nebulosos
ε	critério de parada para o algoritmo <i>fuzzy c-means</i>
$\mathbf{I}$	vetor de índices incompletos
$\Pi(c)$	conjunto de todas as permutações dos c grupos
$\zeta(i)$	linha permutada i
Ψ	matriz de coincidência
ψ_{jl}	valor da coincidência entre os dados j e l

Capítulo 1

Introdução

A tarefa de classificar objetos em categorias é uma das atividades mais primitivas e comuns do Homem. Como resultado do grande volume de informações a serem avaliadas, esta atividade tem sido intensificada nas mais diversas áreas do conhecimento (BACKER, 1995).

A análise exploratória de dados é freqüentemente necessária em diversas áreas, tais como medicina e biologia, a fim de avaliar hipóteses potenciais para estudos futuros. Por exemplo, um modo de reduzir o número de tratamentos para avaliação subsequente é agrupar os dados de forma que reflitam a estrutura interna dos dados e identifique grupos *naturais* presentes, pois normalmente, existem diversos tratamentos para cada caso.

Quando se trabalha com bases de dados reais a ocorrência de dados incompletos é um problema bastante comum, pois esses dados freqüentemente contêm erros e inconsistências. A literatura sobre análise estatística desse tipo de dado ganhou força a partir dos anos 70, impulsionada pelos avanços das tecnologias computacionais, que transformaram cálculos numéricos laboriosos em mecanismos simples (LITTLE; RUBIN, 1987). O último quarto do século XX foi marcado por enormes desenvolvimentos em métodos estatísticos

voltados para a análise de dados incompletos (SCHAFFER, 1999).

Dados incompletos em uma base de dados significa que os valores de alguns dos atributos são desconhecidos. Bases de dados podem estar incompletas por várias razões, tais como: erros de operadores humanos, perdas de dados, ou ainda devido ao mal funcionamento de sensores em equipamentos. Essas diferentes formas de ocorrências de *dados incompletos* influenciam o pesquisador na escolha da técnica de modelagem que será implementada. Normalmente, os dados incompletos são resultantes de perdas aleatórias, assim, o valor *desconhecido* não tem o mesmo *status* que o valor de um atributo existente, ou seja, onde um atributo tem (ou não tem) um valor determinado não provê informação sobre o grupo ao qual pertence (LIU *et al.*, 1997). Assim, a ocorrência de um dado incompleto deve ser vista de forma cuidadosa e o tratamento dado a eles deve ser ajustado, dependendo do tipo de valores característicos e do tipo de sua ocorrência, aleatória ou não (TIMM; DÖRING; KRUSE, 2004).

A análise de agrupamentos é uma ferramenta importante para investigar e interpretar dados. Espera-se que os algoritmos de classificação produzam partições que reflitam a estrutura interna dos dados e identifique grupos *naturais* e hierarquias presentes (LEVINE; DOMANY, 2001). Entretanto, a atribuição de objetos para os grupos e a descrição desses grupos são desconhecidas. Ao agrupar-se objetos similares dentro de grupos, tenta-se representar o atual tipo de categorias desses objetos (DÖRING; LESOT; KRUSE, 2006), de forma que exista homogeneidade dentro dos grupos e heterogeneidade entre os mesmos (MILAGRE *et al.*, 2007b; MILAGRE *et al.*, 2007a).

Assim, diferentes algoritmos de classificação têm sido propostos (JAIN; DUBES, 1988) para descobrir automaticamente o particionamento *natural* da base de dados. Entretanto, esses algoritmos geram um particionamento de dados, independentemente se este parti-

cionamento corresponde ao número correto de grupos. Desta forma, necessita-se validar a saída do algoritmo de classificação respondendo à seguinte questão: os grupos apresentados representam uma estrutura significativa ou é somente um artefato do algoritmo? (LAW; JAIN, 2003).

Geralmente, validar soluções de classificação significa avaliar resultados dos algoritmos de forma quantitativa e objetiva. Essas avaliações podem ser baseadas em dois tipos de critérios. O primeiro é o *critério externo*, onde a solução do classificador é comparada com informações que não estão contidas na base de dados. Por outro lado, o *critério interno* usa exclusivamente a própria base de dados (ROTH *et al.*, 2002).

Um método para avaliar agrupamentos pode ser a utilização da reamostragem, cuja idéia central é comparar o agrupamento obtido em uma base de dados completa com os encontrados em sub-amostras deste conjunto inicial de dados. Espera-se que, quando o número de grupos estiver *correto*, obtenha-se a mesma estrutura presente nos dados originais. Com um número de grupos *incorreto*, entretanto, o agrupamento resultante deve mostrar considerável variação entre as diferentes sub-amostras (ROTH *et al.*, 2002; BORGELT, 2006).

Quando a base de dados contém dados incompletos executa-se um processo de análise em dois passos. No primeiro passo deve-se analisar as razões para a presença dos mesmos. Então, no segundo passo analisa-se as correlações na base de dados. Dependendo dos resultados dessas duas avaliações escolhe-se o método de análise.

Porém, a maioria dos algoritmos para agrupamento de dados exigem que os mesmos estejam completos. Alguns métodos tratam da exclusão dos dados que contêm valores de atributos faltando, outros trabalham reparando ou inserindo valores antes de iniciar a classificação, ou tentam reparar os dados durante o processo de classificação. Contudo,

tanto a exclusão de dados incompletos quanto a inclusão de atributos podem provocar alterações nas bases de dados que levam o pesquisador a conclusões errôneas. Pode-se citar como exemplo o caso de exclusão de dados, onde a amostra pode tornar-se pequena para qualquer tipo de análise, além do mais os dados excluídos podem ser um conjunto de observações bastante diferente dos dados restantes, assim o resultado não poderá ser estendido para o conjunto total de dados. Já no caso de inclusão são assumidas muitas suposições que podem distorcer a distribuição dos dados, levando a resultados tendenciosos e inexatos (MILAGRE *et al.*, 2007b) e em alguns casos, dependendo do método, pode-se imputar valores fora da escala disponível. Ressalta-se, que no caso de inclusão, o valor inserido acompanha todo o processo de agrupamento, assim, um dado incompleto que esteja mal estimado deteriorará toda a qualidade do processo. Um outro problema resultante dos dados incompletos é que nem sempre é possível investigar as causas de sua ocorrência para permitir uma melhor escolha do método a ser utilizado.

Assim, existe um interesse em métodos de análise que tolerem bases de dados incompletas (TIMM; KLAWONN, 1999). Nesse contexto, o objetivo deste trabalho é propor um método para identificar grupos presentes em bases de dados com dados incompletos sem a necessidade de exclusão de dados ou inclusão de valores, presentes nas metodologias tradicionais, desta forma o dado incompleto participa de todo o processo de agrupamento como tal. Além disso, o método deve utilizar um critério de avaliação interno, o qual não requer nenhum conhecimento *a priori* dos dados, tais como, tamanho, geometria e número de atributos da base de dados.

A metodologia é baseada no algoritmo de classificação nebuloso (BEZDEK *et al.*, 2005; HÖPPNER *et al.*, 1999) modificado (TIMM; DÖRING; KRUSE, 2004; HATHAWAY; BEZDEK, 2001; TIMM; KLAWONN, 1998). A análise foi feita introduzindo-se perturbações

no sistema, usando a técnica de reamostragem com reposição (*bootstrapping*) (MILAGRE; MACIEL, 2008a)¹. Foram utilizados dois conjuntos de bases de dados. O primeiro são as bases de dados de validação da qual fazem parte quatro bases de dados artificiais e as bases de dados iris e wine (BALKE; MERZ, 1998). O segundo conjunto corresponde às bases de dados de aplicação, composta por dados genotípicos do *Bradyrhizobium sp.* (MILAGRE, 2003; GERMANO *et al.*, 2006), (MILAGRE *et al.*, 2008)², (MILAGRE; MACIEL, 2008b)³. Os percentuais de valores faltando aplicados às bases de dados foram: 2%, 5%, 10%, 20% e 30%. Neste trabalho, em cada simulação gerou-se 100 sub-amostras, cada uma contendo 90% do total de dados originais, aleatoriamente escolhidos. Para combinar graus de pertinência e matrizes de coincidência utilizou-se as *t-normas* produto e *nil-potent* mínimo. As medidas de similaridade utilizadas foram Diferença Média Quadrática e Classificação Cruzada (BORGELT, 2006); F1 (RIJSBERGEN, 1979); Fowlkes & Mallows e Índice Randômico (LAW; JAIN, 2003); Jaccard (BOUTIN; HASCOËT, 2004); Hubert (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a) e Índice Randômico Ajustado (HUBERT; ARABIE, 1985).

Para elucidar o exposto acima, o trabalho foi organizado em seis capítulos. No Capítulo 2 descreve-se conceitos sobre dados incompletos. No Capítulo 3 aborda-se agrupamento de dados, lógica nebulosa, medidas de comparação e reamostragem. Em seguida, no Capítulo 4, descreve-se toda a metodologia utilizada para a análise do número de grupos. O Capítulo 5 trata dos resultados obtidos e a discussão sobre os mesmos. Finalmente, no Capítulo 6, são apresentadas as conclusões e são apontados possíveis tópicos

¹MILAGRE, S.T.; MACIEL, C.D. (2008). Clustering analysis in datasets within missing values using fuzzy c-means and bootstrapping. A ser submetido.

²MILAGRE, S.T. et al. (2008). Multidimensional cluster stability analysis from a brazilian *Bradyrhizobium sp.* RFLP/PCR data set. No prelo.

³MILAGRE, S.T. et al. (2008). Performance analysis of restriction enzymes based in results of cluster stability. No prelo.

para trabalhos futuros.

Capítulo 2

Dados Incompletos

”Porque dados incompletos criam dificuldades em pesquisas científicas? Porque muitos procedimentos de análises de dados não foram projetados para eles.” (SCHAFER; GRAHAM, 2002).

Dados incompletos são uma realidade em qualquer linha de pesquisa. Podem advir da falta de resposta a um item de pesquisa, podem ser ocasionados por perda acidental de dados, ou até mesmo ser resultante da morte precoce de um paciente. Além disso, uma significativa fração de dados pode estar errada, restando como única alternativa descartar esses dados errados, gerando-se, conseqüentemente dados incompletos (BATISTA; MONARD, 2003).

A presença de dados incompletos pode levar a conclusões incorretas sobre a base de dados. Segundo Bennett (2001) quando a taxa de dados incompletos é alta (maior que 10%) os resultados das subseqüentes análises estatísticas podem ser tendenciosas e para percentuais mais altos a análise pode ser inviável. Outro ponto a ser considerado é o fato de que alguns métodos de tratamento muitas vezes são simplistas, como por exemplo a exclusão dos dados que estão incompletos, podendo introduzir imprecisão e inconsistências nas análises.

Este capítulo apresenta os fundamentos teóricos sobre dados incompletos que são o objeto de estudo deste trabalho. O capítulo é dividido em quatro seções. A primeira seção traz uma classificação sobre dados imperfeitos. A segunda seção trata de definições de dados incompletos. As duas últimas descrevem métodos utilizados para trabalhar com as bases de dados incompletas.

2.1 Tipos de Dados Imperfeitos

”Incerteza permeia nossa compreensão do mundo real. O propósito de sistemas de informação é modelar o mundo real. Então, sistemas de informação devem ser capazes de trabalhar com incertezas.” (MOTRO; SMETS, 1993).

Segundo Parsons (1996) a informação imperfeita é ubíqua, pois quase toda informação que temos sobre o mundo real não é correta, nem completa e também não é precisa. Assim, insistir em estudar somente informações *certas*, como é o caso da maioria dos trabalhos em bases de dados, é concentrar-se somente sobre uma pequena parte do problema como um todo.

De acordo com Bonnisone e Tong (1985) dados imperfeitos podem ser categorizados em incompletos, imprecisos e incertos:

- Incompletos

Surtem da ausência do valor para um atributo e podem ser de dois tipos: existencial ou universal. O tipo existencial ocorre quando uma determinada ocorrência do valor de um atributo do dado é desconhecida (PARSONS, 1996), por outro lado, é universal quando todas as ocorrências dos valores de um determinado atributo são desconhecidas.

- Imprecisos

É quando o valor do atributo existe mas não pode ser medido com precisão. Imprecisão pode ocorrer de diversas formas, como por exemplo, através de um intervalo de valores: "o fato ocorreu entre 3 e 4 da tarde". Pode surgir de conceitos nebulosos como uma negação: "a avaliação não será amanhã", uma informação excludente: "a bactéria morreu dia 5, 6 ou 7", etc.

- Incertos

Dados são considerados incertos quando o valor medido para o atributo é provavelmente próximo do atributo real.

É inerentemente subjetivo. Assim sendo, é feita uma estimativa da verdade sobre algum fato através de uma avaliação individual, utilizando-se, por exemplo, estimativas de probabilidade, ou fazendo-se alguma declaração sobre uma crença de que a proposição é verdadeira (PARSONS, 1996).

Por exemplo, se estão sendo feitas medidas em um determinado equipamento e o valor dessas medidas encontram-se próximas uma da outra, de forma que o equipamento não consegue diferenciá-las perfeitamente, uma alternativa é lançar mão de alguma técnica, que pode ser por exemplo a visual, para atribuir valores diferenciados para essas medidas.

2.2 Tipos de Dados Incompletos

De acordo com Little e Rubin (1987), os mecanismos que levam a dados incompletos podem ser classificados em três tipos: MCAR (Dado incompleto de forma totalmente

aleatória), MAR (Dado incompleto de forma aleatória) e NMAR (Dado incompleto de forma não aleatória).

A Tabela 2.1, traz uma representação esquemática de uma base de dados com p atributos e n dados, que possui dados incompletos, onde a inexistência do valor para o atributo é denotado por NaN.

Tabela 2.1 – Tabela contendo exemplo de uma base de dados multivariada contendo p atributos sobre n dados, alguns dos quais são incompletos, onde a inexistência do valor para o atributo é representado por NaN (Adaptado de (SCHAFFER, 1999)).

	<i>Atrib. 1</i>	<i>Atrib. 2</i>	<i>Atrib. 3</i>	<i>...</i>	<i>Atrib. p</i>
<i>Dado 1:</i>					
<i>Dado 2:</i>			<i>NaN</i>		
<i>Dado 3:</i>	<i>NaN</i>				
$\vdots$				<i>NaN</i>	
$\vdots$	<i>NaN</i>				
$\vdots$		<i>NaN</i>			
$\vdots$				<i>NaN</i>	
<i>Dado n:</i>		<i>NaN</i>			<i>NaN</i>

2.2.1 Dados MCAR

Dados MCAR (*Missing Completely at Random*) são os que possuem maior grau de aleatoriedade. Ocorre quando a probabilidade de um dado possuir um atributo sem valor não depende de qualquer outro valor, seja da base de dados como um todo ou de outro atributo que também não possua valor. Podem surgir, por exemplo, devido ao mal funcionamento de um equipamento, por erro humano na coleta do dado, por perda de dados coletados, e outros.

Para qualquer conjunto de dados pode-se definir um indicador de variáveis $\mathbf{R}$ que identifica o que é conhecido e o que é incompleto. Assim, $\mathbf{R} = \{R_{jk}\}$ é uma matriz n

x p de variáveis indicadoras, onde $j = 1, \dots, n$ representa o índice do dado e $k = 1, \dots, p$ representa o índice do atributo do dado. Desta forma, $R_{jk} = 1$ indica que o atributo está presente e $R_{jk} = 0$ indica que o atributo está ausente (SCHAFFER; GRAHAM, 2002).

Denotando-se o conjunto completo de dados por $\mathbf{X}_{com} = \{x_{jk}\}$. A partição de $\mathbf{X}_{com} = (\mathbf{X}_{obs}, \mathbf{X}_{inc})$, onde $\mathbf{X}_{obs}$ representa as partes observadas e $\mathbf{X}_{inc}$ representa as partes incompletas do conjunto total de dados (LITTLE; RUBIN, 1987; SCHAFFER; GRAHAM, 2002).

Valores MCAR ocorrem quando a distribuição de $\mathbf{R}$ não depende de $\mathbf{X}_{com}$,

$$P(\mathbf{R}|\mathbf{X}_{com}) = P(\mathbf{R}) \quad (2.1)$$

Assim, a probabilidade que o valor para um atributo x_{jk} esteja faltando não está relacionada ao valor de x_{jk} ou ao valor de qualquer outro atributo.

Como exemplos de dados MCAR pode-se citar a perda de uma amostra em laboratório por deixá-la cair ao chão, queda de energia desligando temporariamente um aparelho de medida, morte de um paciente prematuramente. Ou, segundo Howell (2006), outra forma de pensar dados MCAR é perceber que qualquer valor de atributo é exatamente tão provável de estar faltando quanto qualquer outro.

2.2.2 Dados MAR

Em dados MAR (*Missing at Random*) a distribuição de $\mathbf{R}$ não depende dos atributos incompletos, mas pode estar relacionada a outros atributos observados. Em outras palavras, a probabilidade de um dado estar incompleto pode ser definido através da relação

com valores observados.

$$P(\mathbf{R}|\mathbf{X}_{com}) = P(\mathbf{R}|\mathbf{X}_{obs}) \quad (2.2)$$

Como exemplo de dados MAR pode-se citar uma pesquisa onde participantes com dados incompletos diferem daqueles que possuem dados completos, mas o padrão de existência de dados incompletos é delineado a partir de outros atributos do conjunto de dados (BENNETT, 2001). Desta forma, mulheres são menos prováveis de revelar seu peso do que os homens, ou seja, a probabilidade da falta do dado depende do gênero e não depende do peso em si.

2.2.3 Dados NMAR

Quando os dados são incompletos de forma não aleatória diz-se que eles são NMAR (*Not Missing at Random*). Em contraste com MAR onde as probabilidades dos dados incompletos são determinadas completamente pelos dados observados (Equação 2.2), em dados NMAR o padrão de distribuição de $\mathbf{R}$ não é delineável a partir de outros atributos do conjunto de dados. Desta forma, a probabilidade que o valor de um atributo esteja faltando é diretamente relacionada aos dados que não foram coletados ou observados (BENNETT, 2001).

Por exemplo, se um estudo é feito sobre saúde mental e pessoas com diagnóstico de depressão participam da pesquisa, é menos provável que elas reportem seu *status* mental do que outras pessoas. Assim, os dados são não aleatórios (HOWELL, 2006). Outros exemplos: é menos provável que uma pessoa obesa revele seu peso, pessoas com maiores salários são menos prováveis de revelá-los.

Dados NMAR não podem ser ignorados em processos de modelagem, por essa razão são às vezes referenciados como *não-ignoráveis* (BENNETT, 2001).

2.3 Métodos para Tratar Dados Incompletos

Existem diferentes métodos que podem ser usados para manipular dados incompletos que se baseiam no tipo e grau de ocorrência desses dados na base de dados (BENNETT, 2001). Conhecer as circunstâncias e razões pelas quais os dados estão incompletos é sempre útil aos pesquisadores (BENNETT, 2001), pois o conhecimento, ou a ausência de conhecimento do mecanismo que levou a determinados valores estarem faltando, é um elemento chave na escolha da análise apropriada e na interpretação dos resultados (LITTLE; RUBIN, 1987).

Os métodos para tratar dados incompletos podem ser divididos em duas formas principais, a primeira compreende os métodos que ignoram os dados incompletos, por meio da simples remoção, ou seja, descarta-se esses dados da base de dados. A outra forma é a inserção, onde os valores que faltam são *preenchidos* por valores estimados. Nesta seção serão abordados alguns desses métodos, em (LITTLE; RUBIN, 2002; SCHAFER, 1999; SCHAFER; GRAHAM, 2002) encontram-se maiores detalhes.

2.3.1 Métodos que Ignoram Observações Incompletas

De acordo com Little e Rubin (1987) são métodos apropriados para bases de dados MCAR, onde a existência de dados incompletos não está relacionada aos demais dados.

1. Análise de Caso Completo

Este método é também conhecido como remoção *listwise* (LD - *Listwise Deletion*) ou remoção de casos (CD - *Case Deletion*). É um método tradicional e também o mais popular, sendo pré-assumido em muitos pacotes estatísticos (BENNETT, 2001). O objetivo da análise são os dados que possuem valores para todos os atributos, desta forma, todos os dados que possuem atributos incompletos são descartados.

Duas vantagens desse método são a simplicidade e possibilidade de comparações posteriores. A simplicidade diz respeito às análises estatísticas padrões, que podem ser aplicadas sem modificações. A possibilidade de comparação refere-se às estatísticas univariadas, como por exemplo a média, que podem ser comparadas, pois todas são calculadas sobre a mesma base de dados. A desvantagem reside no fato da potencial perda de informações quando descarta-se parte da base de dados (LITTLE; RUBIN, 1987).

Uma preocupação fundamental sobre esse método baseia-se em quando a análise pode tornar-se tendenciosa.

No caso de dados MCAR, a base de dados representa efetivamente uma sub-amostra escolhida aleatoriamente do conjunto original e as inferências feitas são válidas. Usualmente, entretanto, a base de dados com valores descartados difere de forma importante da amostra original, como quando tem-se perda de indivíduos em uma pesquisa, aqueles que permanecem são freqüentemente diferentes dos que foram perdidos. Nesses casos, a análise de caso completo pode fornecer resultados fortemente tendenciosos.

Se os dados incompletos não são MCAR, a base de dados composta apenas pelos dados completos pode ser uma sub-amostra tendenciosa. Assim, os resultados

produzidos podem ser tendenciosos.

Segundo Schafer e Graham (2002) deve-se considerar que a principal virtude deste método é a sua simplicidade. Deve-se levar em conta, porém, que se a quantidade de dados incompletos é alta, uma grande porção da base de dados será descartada. Se o problema de dados incompletos pode ser resolvido descartando-se apenas uma pequena parte da base de dados, então o método pode ser bastante eficiente, mas mesmo neste caso não se pode deixar de explorar a base de dados para que se esteja certo de que as exclusões não afetarão fortemente os resultados.

2. Análise de Caso Disponível

A maneira mais comum de excluir dados incompletos para calcular-se uma matriz de correlação é utilizando-se a análise de caso completo, assim obtém-se a verdadeira matriz de correlação, onde todas as correlações são obtidas do mesmo conjunto de dados. Entretanto, se os atributos sem valor são aleatoriamente distribuídos através dos dados, pode-se chegar ao final do processo de exclusão desses dados sem nenhum dado válido.

A Análise de Caso Disponível (*Available-Case (AC) analysis*) é um método onde inclui-se todos os dados nos quais o atributo de interesse, a ser utilizado para um cálculo estatístico em particular, está presente, assim esta análise utiliza todos os atributos disponíveis, com a desvantagem do tamanho da amostra base alterar de atributo para atributo de acordo com o padrão de dado incompleto (LITTLE; RUBIN, 1987).

Por exemplo, se uma matriz de correlação é requerida para três atributos A , B e C e o atributo C não possui alguns valores, todos os dados serão incluídos no cálculo

da correlação entre A e B , mas somente alguns dados serão incluídos na correlação entre A e C e entre B e C .

De acordo com Schafer e Graham (2002) o princípio da análise AC, que é fazer uso de toda informação disponível, é eminentemente sensível, mas deletar casos é uma forma mais pobre ainda. Como os parâmetros são estimados a partir de diferentes unidades amostrais, existe também a dificuldade de comparar erros padrões ou outras medidas de incertezas.

Na análise de Little e Rubin (1987) como o método de caso disponível aparentemente faz uso de todos os dados, espera-se então que seja superior à análise de caso completo. Isto é verdade quando os dados são MCAR e as correlações são pequenas. Entretanto, quando as correlações são altas, as simulações indicam superioridade para o método de caso completo. Porém, geralmente nenhum dos dois métodos é satisfatório.

2.3.2 Métodos de Inserção Única

O método de inserção, ou preenchimento, é um método geral e flexível para lidar com dados incompletos. Entretanto, não está livre de perigos (LITTLE; RUBIN, 1987). Nas palavras de Dempster e Rubin (DEMPSTER; RUBIN, 1983):

"A idéia de inserção é ao mesmo tempo sedutora e perigosa. É sedutora porque pode embalar o usuário em um estado de satisfação de acreditar que depois de tudo, os dados estão completos, e é perigosa porque reúne situações onde o problema é suficientemente pequeno e pode, assim, legitimamente ser trabalhado desta forma e, situações onde as estimativas padrão aplicadas aos dados reais e aos dados que passaram por inserção têm substanciais tendências."

Nestes métodos de inserção ao invés de descartar os dados incompletos, tenta-se substituí-los por valores que sejam plausíveis. É potencialmente mais eficiente que o método de exclusão de casos, pois mantendo-se a base de dados completa evita-se a perda de potência do teste devido à diminuição do tamanho da amostra. Produz uma base de dados aparentemente completa, o que facilita a aplicação de métodos padrões de análise e utilização de softwares. Um problema com este método pode ser a distorção da distribuição dos dados e suas relações, pois algumas medidas estatísticas são tendenciosas usando uma base de dados *preenchida* (ZHOU, 2006).

Apresenta-se aqui os métodos mais comuns de inserção única.

1. Substituição pela Média

A substituição pela média (*Mean Imputation*) é uma forma particularmente simples de inserção. Substitui os atributos incompletos pela média dos dados observados para esse atributo.

O uso da substituição pela média pode ser justificada no fato de que a média é uma suposição razoável de valor para uma observação aleatoriamente selecionada de uma distribuição normal. Entretanto, quando dados incompletos não são estritamente aleatórios (MCAR) a substituição pela média pode não ser uma suposição razoável. Outro ponto a ser considerado é que este método é especialmente problemático quando existe um alto percentual de dados incompletos (ACOCK, 2005).

Uma vantagem deste método é que ele preserva a parte dos dados que seriam eliminados pelos métodos de deleção. Por outro lado, leva a estimativas tendenciosas que diferem dos parâmetros da população em geral (ZHOU, 2006). Assim, nem sempre é recomendado porque apesar de preservar a média, outros aspectos como variância

e covariância podem ser sub-estimados.

Na análise de Howell (2006) é um procedimento antigo que deve certamente ser relegado ao passado. Não adiciona nenhuma informação nova, pois a média total, com ou sem a substituição de atributos incompletos pela média dos atributos observados será a mesma. Adicionado a isso, esse método leva a subestimar-se o erro padrão devido ao fato de aumentar o tamanho da amostra e aumentar o denominador, reduzindo então o erro padrão.

2. Métodos de Regressão

Se em uma base de dados, os dados completos se comportam de acordo com uma distribuição normal com média μ e matriz de covariância Σ então, os dados incompletos têm regressões lineares sobre os dados observados, com coeficientes de regressão que são funções bem conhecidas de μ e Σ . Assim, este método estima primeiramente μ e Σ a partir dos dados completos e utiliza essas estimativas para calcular as regressões lineares dos dados incompletos sobre os dados completos, caso a caso, ou seja, neste método desenvolve-se uma equação de regressão a partir dos dados observados para determinado atributo.

Este método fornece boas estimativas de médias, particularmente se as suposições de normalidade são plausíveis mas subestima a variância, ainda que seja menor que os obtidos pelo método de substituição pela média (LITTLE; RUBIN, 2002). Outra desvantagem é que se os padrões da base de dados e dos dados incompletos não forem simples, os pesos e desvio-padrão podem se tornar extremamente complexos (BENNETT, 2001).

3. Inserção *Hot-deck*

Nesta inserção a idéia é preservar a forma da distribuição das variáveis. Cada atributo incompleto é substituído por um valor obtido a partir da comparação dos valores observados para esse atributo. Ele substitui o atributo incompleto por um valor observado (*o doador*). Uma forma de melhor escolher o *doador* é escolhê-lo não em toda a base de dados, mas entre aqueles atributos que tenham características similares (ZHOU, 2006).

Assim, este método é tipicamente implementado em dois estágios. No primeiro estágio particiona-se a base de dados em grupos disjuntos e homogêneos. No segundo estágio os dados completos pertencentes a um determinado grupo são utilizados para estimar os valores dos atributos incompletos pertencentes a esse mesmo grupo. Para cálculo da estimativa pode-se utilizar, por exemplo, a média dos atributos completos pertencentes ao grupo em questão (BATISTA; MONARD, 2003).

É menos tendencioso que a inserção pela média e análise de caso completo. Assume que os dados são MAR. Uma desvantagem é que às vezes são necessários algoritmos complexos de comparação (BENNETT, 2001). Como a variabilidade para o item de dado não é distorcida, este método soluciona parcialmente o problema de entender-se a incerteza. Porém, poderá distorcer correlações e outras medidas de associação se métodos refinados não forem utilizados.

4. Inserção *Cold-deck*

É semelhante à inserção *hot-deck*. A diferença é que além de utilizar as informações presentes na base de dados, a estratégia para avaliar a similaridade subjetiva baseia-se em informações externas ou conhecimentos *a priori* que se tem sobre os dados,

tais como estudos pilotos, pesquisas similares encontradas na literatura, opinião de especialistas. Assim, esse método é dependente da qualidade da avaliação externa (BENNETT, 2001).

2.4 Métodos de Estimativas de Parâmetros

Incluem os métodos que trabalham com inserção múltipla e aqueles que se baseiam em soluções de máxima verossimilhança. A idéia desta técnica é combater o problema de sub-estimar a variabilidade da base de dados. Serão vistos aqui os métodos Maximização da Expectativa e Inserção Múltipla.

2.4.1 Método Maximização da Expectativa

Conhecido como algoritmo EM (*Expectation-Maximization*), usa o fato que o dado incompleto contém informações relevantes a serem utilizadas na estimativa do parâmetro de interesse, como por exemplo efeito de um tratamento, taxa de casos positivos (BENNETT, 2001).

É um eficiente processo iterativo para calcular estimativa de *Máxima Verossimilhança* (*ML - Maximum Likelihood*) para dados incompletos. Em uma estimativa ML, deseja-se estimar o(s) modelo(s) de parâmetros no qual os dados observados são mais prováveis (LITTLE; RUBIN, 1987).

É dividido em dois passos em cada iteração:

- Passo **E** (Expectativa):

Os valores que faltam são estimados a partir dos valores observados e a estimativa

dos parâmetros de interesse é encontrada;

- Passo **M** (Maximização):

Substitui-se os valores esperados (como média e variância) para os valores que faltam (obtidos no passo-**E**) e a função de máxima verossimilhança é maximizada sob a suposição de que os valores que faltam agora são conhecidos, obtendo-se assim novas estimativas de parâmetros. Esses novos parâmetros estimados são substituídos novamente no passo-**E** e um novo passo-**M** é feito.

Repete-se o processo até que a convergência seja alcançada, ou seja, até que a diferença entre os parâmetros estimados entre a iteração anterior e a atual seja desprezível.

2.4.2 Inserção Múltipla

Ao invés de preencher cada atributo incompleto com um único valor, o procedimento de Inserção Múltipla (MI - *Multiple Imputation*), substitui cada valor de valor faltando por um conjunto de M valores plausíveis (tipicamente M está entre 5 e 10 (BENNETT, 2001; ACOCK, 2005)) que representem a incerteza sobre o correto valor a ser inserido, para criar M bases de dados completas (LITTLE; RUBIN, 1987). Para gerar os valores de inserção, deve ser aplicado sobre a base de dados completa (com dados completos e dados incompletos) um modelo de probabilidade.

Cada base M de dados é analisada da mesma forma por um método de análise de dados completos. Os resultados, que podem variar entre si (porque os valores inseridos podem ser diferentes), são combinados para obter a estimativa completa.

A validade do método MI depende de como as inserções são criadas e como este

procedimento se relaciona com o modelo usado para subsequente análise dos dados. Criar MI's requer algoritmos especiais (SCHAFFER; GRAHAM, 2002).

Atualmente tornou-se um método popular, porque o alto custo computacional já não representa limitações diante das modernas tecnologias computacionais que se tem hoje.

2.5 Conclusão

Atualmente existem vários métodos que permitem análise de bases de dados incompletas. Esses métodos podem ser simples, como a exclusão dos dados incompletos da base de dados, ou métodos mais complexos, como por exemplo a inserção múltipla. Em todos eles é necessário que se conheça a realidade da base de dados e o porque da existência de dados incompletos para que se escolha o método a ser aplicado, pois a aplicação dos mesmos pode levar a resultados imprecisos e inconsistências nas análises.

Capítulo 3

Agrupamento de Dados

Os procedimentos exploratórios são, freqüentemente, úteis no entendimento da natureza complexa das relações multivariadas. Buscar nos dados uma estrutura de *agrupamentos naturais* é uma importante técnica exploratória. Agrupamentos podem fornecer um meio informal para acesso à dimensionalidade, identificando tendências e sugerindo hipóteses relativas aos parentescos (JOHNSON, 1992).

A análise de agrupamentos é uma das mais antigas técnicas em que não são feitas suposições com relação ao número de grupos ou à estrutura dentro do grupo. A classificação é feita com base nas similaridades ou distâncias (dissimilaridades) e nenhuma informação sobre os grupos e variáveis é assumida (JOHNSON, 1992).

O objetivo no problema de agrupamento é particionar a base de dados em grupos homogêneos e bem separados, de forma que itens que pertençam a um mesmo grupo sejam tão similares quanto possível e itens pertencentes a grupos diferentes sejam o mais dissimilares possível (XIE; BENI, 1991; HÖPPNER *et al.*, 1999). Classificação também pode ser vista como uma forma de compressão de dados, onde um grande número de amostras são convertidas dentro de um pequeno número de protótipos representativos ou

grupos.

A análise de agrupamentos é uma ferramenta importante em processamento de imagens, biologia, reconhecimento de dados, mineração de dados, sensoriamento remoto, bioinformática, etc. Entretanto, em muitos desses problemas, existe, *a priori*, poucas informações disponíveis sobre os dados e na tomada de decisão deve-se fazer tão poucas suposições quanto possível. É baseado nessas restrições que a metodologia de agrupamento é particularmente apropriada para a exploração de inter-relações entre os dados para dar uma contribuição (ainda que preliminar) de sua estrutura (JAIN; MURTY; FLYNN, 1999). Baseada em diferentes teorias, muitas soluções têm sido propostas para esta técnica, e muitas pesquisas focando sobre tipos especiais de algoritmos de agrupamentos têm sido apresentadas (YU, 2005).

Segundo Jain, Murty e Flynn (1999) análise de agrupamento é a classificação não supervisionada de padrões (observações, dados, ou vetor de características) dentro de grupos (*clusters*). Em uma classificação supervisionada tem-se inicialmente uma coleção de dados rotulados (pré-classificados), sendo que o problema é rotular um novo dado encontrado e ainda não rotulado. Tipicamente, os dados rotulados (conjunto de treinamento) são usados para aprender a descrição das classes e voltam a ser utilizados para rotular o novo dado. Na classificação não supervisionada o problema é agrupar uma dada coleção de dados não rotulados dentro de grupos expressivos. Neste sentido, rótulos são também associados com grupos, mas nessa categoria, rótulos são *dirigidos a dados*, ou seja, eles são obtidos somente dos dados.

Existem três tipos de métodos de agrupamentos que são utilizados para encontrar o particionamento ótimo: métodos hierárquicos, métodos baseados em teoria dos grafos e métodos baseados em função objetivo. Os métodos baseados em função objetivo são

muito populares (SARKAR; LEONG, 2001), dentre eles pode-se destacar a classificação nebulosa (*fuzzy*) que é utilizada neste trabalho, assim são apresentados neste capítulo os conceitos relativos à lógica nebulosa e também à análise de dados incompletos utilizando esta técnica.

3.1 Critérios de Validação

Como é avaliada a saída de um algoritmo de agrupamento? O que caracteriza um resultado de agrupamento *bom* e outro *ruim*? Todos os algoritmos de agrupamento, quando apresentados a uma base de dados, irão produzir classes, independente se os dados contêm classes ou não (JAIN; MURTY; FLYNN, 1999).

A avaliação dos procedimentos de agrupamento têm então, diferentes facetas. Uma é a avaliação do domínio dos dados mais do que do próprio algoritmo de agrupamento, dados que não possuem grupos não devem ser processados por um algoritmo de agrupamento. Esse estudo faz parte da área de pesquisa chamada *tendência de classe* (*cluster tendency*), uma área de pesquisa relativamente inativa, na qual examina-se os dados de entrada para verificar se existe qualquer mérito para a análise de grupos, antes de executar o algoritmo de agrupamento. Por outro lado, na análise de validade de grupos (*cluster validity analysis*) avalia-se a saída do procedimento de agrupamento. Os critérios de validade são objetivos e executados para determinar o quão significativa é a saída. Uma estrutura de agrupamento é válida se não tiver ocorrido razoavelmente por acaso ou como um artefato do algoritmo (JAIN; MURTY; FLYNN, 1999).

A validação dos resultados dos algoritmos de forma quantitativa e objetiva pode ser baseada em dois tipos de critérios (ROTH *et al.*, 2002):

- Critério externo

A solução do classificador é comparada com informações que não estão contidas na base de dados;

- Critério interno

Usa exclusivamente o próprio conjunto de dados.

3.2 Definições e Notações

Antes de prosseguir, faz-se necessária a definição de alguns termos e notações que serão utilizados no decorrer deste trabalho. Em (JAIN; MURTY; FLYNN, 1999) define-se os seguintes termos:

1. **Dado:** ou vetor de características, observações ou amostras $\mathbf{x}$ é um item de dado único utilizado pelo algoritmo de agrupamento. Tipicamente consiste de um vetor que possui p medidas (ou atributos): $\mathbf{x} = (x_1, \dots, x_p)$;
2. **Atributo:** ou característica são os componentes individuais do dado $\mathbf{x}$;
3. **Dimensionalidade:** p é a dimensão do dado ou o espaço do dado;
4. **Base de Dados:** é o conjunto total de dados, denotado por $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$. O j -ésimo dado em $\mathbf{X}$ é denotado por $\mathbf{x}_j = (x_{j1}, \dots, x_{jp})$. Na Tabela 3.1, exemplifica-se o conjunto total de dados $\mathbf{X}$, com p atributos e n dados;
5. **Grupo:** é um conjunto de dados que são similares entre si e diferentes dos outros dados presentes na base de dados. Pode ser visto como uma fonte de dados cuja distribuição no espaço de características é governada por uma função densidade

probabilidade específica para o grupo. As técnicas de agrupamento tentam agrupar dados de forma que os grupos obtidos reflitam os processos de geração de diferentes dados representados no conjunto total de dados.

Tabela 3.1 – Tabela referente ao conjunto total de dados $\mathbf{X}$, contendo p atributos sobre n dados.

	<i>Atrib. 1</i>	<i>Atrib. 2</i>	...	<i>Atrib. w</i>	...	<i>Atrib. p</i>
<i>Dado 1:</i>	x_{11}	x_{12}	...	x_{1w}	...	x_{1p}
<i>Dado 2:</i>	x_{21}	x_{22}	...	x_{2w}	...	x_{2p}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	...
<i>Dado j:</i>	x_{j1}	x_{j2}	...	x_{jw}	...	x_{jp}
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	...
<i>Dado n:</i>	x_{n1}	x_{n2}	...	x_{nw}	...	x_{np}

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1w} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2w} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j1} & x_{j2} & \dots & x_{jw} & \dots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nw} & \dots & x_{np} \end{pmatrix} \quad (3.1)$$

3.3 Lógica Nebulosa

No mundo real é comum a utilização de termos como: alto, baixo, gordo, magro, bonito, feio, muito quente, muito frio, etc. para descrever situações de nossa vida cotidiana. Em se tratando de pensamento humano, a resolução deste tipo de problema torna-se trivial, pois o cérebro utiliza conhecimento, experiência, intuição, realiza comparações.

Assim o ser humano é capaz de lidar com informações incertas, imprecisas, incompletas da forma que lhe é natural, mas quando se deseja transformar esse tipo de informação em valores numéricos o problema não é tão simples assim.

Na lógica tradicional, booleana, trabalha-se com os valores *0* e *1* (ou *falso* e *verdadeiro*), não existem valores intermediários. Desta forma, a lógica clássica (binária) não consegue tratar declarações que sejam *parcialmente falsas* ou *parcialmente verdadeiras*, logo esses termos *imprecisos* não podem ser modelados pela matemática tradicional de conjuntos.

Na década de 60, o matemático, professor da Universidade de Berkeley (EUA), Lofti Asker Zadeh, percebeu que os recursos tecnológicos disponíveis não eram suficientes para automatizar problemas de incertezas, ambigüidades, imprecisão utilizando lógica computacional (binária). Em 1965 introduziu a lógica nebulosa (*fuzzy logic*) (ZADEH, 1965), onde o termo *nebuloso* significa não totalmente delineado. Assim, o objetivo da introdução dessa lógica era definir uma estrutura formal para representação e gerenciamento do conhecimento vago e incerto. Desta forma, na lógica nebulosa os valores se encontram "entre 0 e 1". Neste contexto, a lógica booleana é, então, um caso particular da lógica nebulosa.

Os sistemas que têm como base a lógica nebulosa e as redes neurais são os principais integrantes do que tem sido definido como computação flexível (*soft computing*). O princípio da computação flexível é explorar a tolerância à imprecisão, à incerteza e à veracidade parcial para alcançar tratabilidade, robustez e baixo custo de solução (ZADEH, 1994).

3.3.1 Conjuntos Nebulosos

De acordo com (ZADEH, 1965), um conjunto clássico é caracterizado por sua função característica, definida da seguinte forma: Seja $\mathbf{X}$ um espaço de pontos (dados) com um elemento genérico de $\mathbf{X}$ denotado por $\mathbf{x}$, assim $\mathbf{X} = \{\mathbf{x}\}$ e seja $\mathbf{F}$ um subconjunto de $\mathbf{X}$. A função característica de $\mathbf{F}$ é dada por:

$$C_{\mathbf{F}}(\mathbf{x}) = \begin{cases} 1, & \text{se } \mathbf{x} \in \mathbf{F} \\ 0, & \text{se } \mathbf{x} \notin \mathbf{F} \end{cases}$$

Assim, $C_{\mathbf{F}}$ é uma função cujo domínio é $\mathbf{X}$ e a imagem está contida no conjunto $\{0, 1\}$.

Partindo da definição clássica de conjunto e *ampliando* a imagem da função característica, Zadeh formalizou o conjunto nebuloso (*fuzzy set*).

Na teoria dos conjuntos nebulosos existe um grau de pertinência de cada elemento a um determinado conjunto. Onde a função de pertinência (também chamada de função característica) define o grau de pertinência de um elemento em relação a um determinado conjunto, ou seja, *o quanto* um elemento pertence a um determinado conjunto. Na lógica booleana existem somente dois graus de pertinência que são 0% e 100%. A lógica nebulosa é multi-valorada, permite suaves transições para os graus de pertinência *entre* 0% e 100%. Desta forma, a função de pertinência associa cada dado $\mathbf{x}$ pertencente a $\mathbf{X}$ a um número real $\mu(\mathbf{x})$ no intervalo $[0, 1]$.

Definição 1: Um conjunto nebuloso $\mathbf{A}$ na base de dados $\mathbf{X}$ é caracterizado por uma função de pertinência $\mu_{\mathbf{A}}(\mathbf{x})$ a qual associa cada ponto em $\mathbf{X}$ a um número real no intervalo $[0,1]$, com o valor de $\mu_{\mathbf{A}}(\mathbf{x})$ representando o *grau de pertinência* de $\mathbf{x}$ em $\mathbf{A}$, ou

seja, $\mathbf{A}$ é um conjunto de pares ordenados do elemento genérico $\mathbf{x}$ dado por:

$$\mathbf{A} = \{(\mathbf{x}, \mu_{\mathbf{A}}(\mathbf{x})) \mid \mathbf{x} \in \mathbf{X}\}$$

O conjunto nebuloso $\mathbf{A}$ é caracterizado por: $\mu_{\mathbf{A}}(\mathbf{x}) : \mathbf{X} \rightarrow [0, 1]$. Onde $\mu_{\mathbf{A}}(\mathbf{x}) = 1$ indica pertinência total ao conjunto nebuloso $\mathbf{A}$ e $\mu_{\mathbf{A}}(\mathbf{x}) = 0$ indica não pertinência ao conjunto nebuloso $\mathbf{A}$.

Um valor de $\mu(\mathbf{x})$ próximo de zero indica um baixo grau de pertinência e $\mu(\mathbf{x})$ próximo de 1 indica alto grau de pertinência.

3.3.2 Operações em Conjuntos Nebulosos

Da mesma forma que os conjuntos clássicos, os conjuntos nebulosos também necessitam de operações tais como: *união*, *intersecção*, *negação*, *subtração*, entre outras.

Dados os conjuntos nebulosos $\mathbf{A}$, $\mathbf{B}$ e $\mathbf{C}$ definidos em $\mathbf{X}$, Zadeh (1965) descreve as seguintes definições para os mesmos:

Definição 2: Um conjunto nebuloso é vazio *se e somente se* sua função de pertinência for idêntica a zero em $\mathbf{X}$.

Definição 3: Dois conjuntos nebulosos $\mathbf{A}$ e $\mathbf{B}$ são iguais (escreve-se $\mathbf{A} = \mathbf{B}$), *se e somente se* $\mu_{\mathbf{A}}(\mathbf{x}) = \mu_{\mathbf{B}}(\mathbf{x})$ para todo $\mathbf{x}$ em $\mathbf{X}$.

Definição 4: O complemento do conjunto nebuloso $\mathbf{A}$ (denotado por $\mathbf{A}'$) é definido por:

$$\mu_{\mathbf{A}'}(\mathbf{x}) = 1 - \mu_{\mathbf{A}}(\mathbf{x})$$

Definição 5: $\mathbf{A}$ está contido em $\mathbf{B}$ (ou equivalentemente $\mathbf{A}$ é subconjunto de $\mathbf{B}$, ou

$\mathbf{A}$ é menor ou igual a $\mathbf{B}$) se e somente se $\mu_{\mathbf{A}}(\mathbf{x}) \leq \mu_{\mathbf{B}}(\mathbf{x})$:

$$\mathbf{A} \subset \mathbf{B} \Leftrightarrow \mu_{\mathbf{A}}(\mathbf{x}) \leq \mu_{\mathbf{B}}(\mathbf{x})$$

Definição 6: A união de dois conjuntos nebulosos $\mathbf{A}$ e $\mathbf{B}$ com as respectivas funções de pertinência $\mu_{\mathbf{A}}(\mathbf{x})$ e $\mu_{\mathbf{B}}(\mathbf{x})$ é um conjunto nebuloso $\mathbf{C}$ (escreve-se como $\mathbf{C} = \mathbf{A} \cup \mathbf{B}$), cuja função de pertinência é relacionada às de $\mathbf{A}$ e $\mathbf{B}$ por:

$$\mu_{\mathbf{C}}(\mathbf{x}) = \text{Max}[\mu_{\mathbf{A}}(\mathbf{x}), \mu_{\mathbf{B}}(\mathbf{x})], \quad \mathbf{x} \in \mathbf{X}$$

Note-se que a união ($\cup$) tem propriedade associativa, ou seja, $\mathbf{A} \cup (\mathbf{B} \cup \mathbf{C}) = (\mathbf{A} \cup \mathbf{B}) \cup \mathbf{C}$.

Definição 7: A intersecção de dois conjuntos nebulosos $\mathbf{A}$ e $\mathbf{B}$ com as respectivas funções de pertinência $\mu_{\mathbf{A}}(\mathbf{x})$ e $\mu_{\mathbf{B}}(\mathbf{x})$ é um conjunto nebuloso $\mathbf{C}$ (escreve-se como $\mathbf{C} = \mathbf{A} \cap \mathbf{B}$), cuja função de pertinência é relacionada às de $\mathbf{A}$ e $\mathbf{B}$ por:

$$\mu_{\mathbf{C}}(\mathbf{x}) = \text{Min}[\mu_{\mathbf{A}}(\mathbf{x}), \mu_{\mathbf{B}}(\mathbf{x})], \quad \mathbf{x} \in \mathbf{X}$$

Como nos conjuntos comuns, $\mathbf{A}$ e $\mathbf{B}$ são disjuntos se $\mathbf{A} \cap \mathbf{B}$ é vazio. Note-se que a intersecção ($\cap$) tem propriedade associativa, ou seja, $\mathbf{A} \cap (\mathbf{B} \cap \mathbf{C}) = (\mathbf{A} \cap \mathbf{B}) \cap \mathbf{C}$.

Definição 8: O produto algébrico de $\mathbf{A}$ e $\mathbf{B}$ é denotado por $\mathbf{AB}$ e definido em termos das funções de pertinência $\mu_{\mathbf{A}}(\mathbf{x})$ e $\mu_{\mathbf{B}}(\mathbf{x})$:

$$\mu_{\mathbf{AB}}(\mathbf{x}) = \mu_{\mathbf{A}}(\mathbf{x})\mu_{\mathbf{B}}(\mathbf{x})$$

Claramente,

$$\mathbf{AB} \subset \mathbf{A} \cap \mathbf{B}$$

Definição 9: A soma algébrica de $\mathbf{A}$ e $\mathbf{B}$ é denotada por $\mathbf{A} + \mathbf{B}$ e definida por:

$$\mu_{\mathbf{A}+\mathbf{B}}(\mathbf{x}) = \mu_{\mathbf{A}}(\mathbf{x}) + \mu_{\mathbf{B}}(\mathbf{x})$$

Gerando a soma que é menor ou igual a um. Assim, diferentemente do produto algébrico, a soma algébrica só é significativa quando a condição $\mu_{\mathbf{A}}(\mathbf{x}) + \mu_{\mathbf{B}}(\mathbf{x}) \leq 1$ é satisfeita para todo $\mathbf{x}$.

Definição 10: A diferença absoluta de $\mathbf{A}$ e $\mathbf{B}$ é denotada por $|\mathbf{A} - \mathbf{B}|$ e definida por:

$$\mu_{|\mathbf{A}-\mathbf{B}|}(\mathbf{x}) = |\mu_{\mathbf{A}}(\mathbf{x}) - \mu_{\mathbf{B}}(\mathbf{x})|$$

3.3.3 Normas Triangulares

As normas triangulares, introduzidas por (MENGER, 1942), são generalizações dos operadores união e intersecção.

Definição 1: Uma norma triangular (*t-norma*) é uma operação binária $\top : [0, 1] \times [0, 1] \rightarrow [0, 1]$, satisfazendo as seguintes condições (para $\forall a, b, c \in [0, 1]$):

1. Comutatividade: $\top(a, b) = \top(b, a)$;
2. Associatividade: $\top(a, \top(b, c)) = \top(\top(a, b), c)$;
3. Monotonicidade: $\top(a, b) \leq \top(c, d)$ se $(a \leq c)$ e $(b \leq d)$;
4. Elemento neutro: $\top(a, 1) = a$;

5. $\top(0, 0) = 0$.

A *Comutatividade* indica que o operador $\top$ independe da ordem dos conjuntos nebulosos que estão sendo combinados. A *Associatividade* generaliza o operador $\top$ para qualquer número de conjuntos nebulosos. A *Monotonicidade* implica que um decréscimo nos valores de pertinência de dois conjuntos nebulosos não pode produzir um aumento no valor de pertinência após a aplicação do operador $\top$. Os critérios 4 e 5 indicam que as *t-normas* podem ser generalizadas para conjuntos clássicos.

Uma *t-norma* aplicada aos conjuntos nebulosos é utilizada para implementar a intersecção (*and*) entre os mesmos. Uma co-norma triangular $\perp$ (ou *s-norma*) implementa a união (SANDRI; CORREA, 1999).

Algumas *t-normas* bem conhecidas são definidas na Tabela 3.2. Neste trabalho utilizou-se as *t-normas nil-potent* mínimo ($\top_{minnp}(a, b)$) e produto ($\top_{prod}(a, b)$).

Tabela 3.2 – Relação das principais *t-normas*.

<i>t-norma</i>	<i>nome</i>
$\min(a, b)$	Zadeh
$a \cdot b$	produto
$\min(a, b)$, se $a+b \geq 1$, 0 se não	<i>nil-potent</i> mínimo
$\max(0, a+b-1)$	Lukasiewicz

As Figuras 3.1 e 3.2 trazem o exemplo das *t-normas* Zadeh e produto.

3.3.4 Agrupamentos Nebulosos

Uma técnica de agrupamento não nebulosa (*hard* ou *crisp*) identifica o grupo de cada dado $\mathbf{x}_j$ atribuindo a cada um deles um rótulo único l_j , com $l_j \in \{1, \dots, c\}$, onde c é o número de grupos.

Figura 3.1 – *t-norma* Zadeh (Modificada de (JAFELICE; BARROS; BASSANEZI, 2005)).

Figura 3.2 – *t-norma* produto (Modificada de (JAFELICE; BARROS; BASSANEZI, 2005)).

Em um agrupamento nebuloso cada grupo é um conjunto nebuloso de todos os dados (JAIN; MURTY; FLYNN, 1999), assim, cada dado de entrada $\mathbf{x}_j$ pode pertencer a mais de um grupo, e associados com cada um dos dados estão os graus de pertinência μ_{ij} , indicando o grau com o qual cada dado j pertence ao grupo i .

A Figura 3.3 traz um exemplo gráfico para agrupamentos nebulosos e não nebulos-

sos. Os retângulos, H_1 e H_2 , representam dois agrupamentos *hard*. As elipses F_1 e F_2 representam a saída de um algoritmo nebuloso.

Figura 3.3 – Agrupamentos nebulosos (Modificada de (JAIN; MURTY; FLYNN, 1999)).

Em agrupamentos nebulosos, cada agrupamento é definido por pares ordenados (j, μ_j) , onde j representa o j -ésimo dado e μ_j é seu grau de pertinência. Altos valores de pertinência (próximo de 1) indicam maior confiança em atribuir o dado para o grupo. O dado 4 em F_1 , por exemplo, poderia ser definido pelo par ordenado $(4, 0.8)$ e conseqüentemente em F_2 seria $(4, 0.2)$, indicando que o dado 4 pertence ao grupo F_1 com um grau maior de pertinência.

A análise de agrupamentos nebulosos (*fuzzy cluster analysis*) é uma técnica de aprendizado não supervisionado. Nesta técnica apenas os atributos de entrada estão disponíveis, diferindo do aprendizado supervisionado o qual possui como exemplo atributos de entrada e saída. A aprendizagem não supervisionada é também conhecida como aprendizagem *sem professor*, pois não é necessária uma resposta correta associada a cada dado de entrada no conjunto de dados avaliado.

Segundo Everitt, Landau e Leese (2001), a principal vantagem dos agrupamentos nebulosos em relação aos demais métodos particionais está no fato de representar com muito mais detalhes a informação sobre a estrutura dos dados. Por outro lado, isso poderia ser considerado uma desvantagem, pois a quantidade de informação cresce rapidamente com o número de dados e o número de grupos, tornando a compreensão difícil. Outra desvantagem é a ausência de dados representativos dos grupos formados e o fato de que geralmente os algoritmos são mais complicados e consomem um maior tempo de computação. No entanto, em problemas reais, dificilmente consegue-se incluir um dado em apenas um grupo, de maneira rígida, o que torna os princípios nebulosos muito interessantes.

3.3.5 Algoritmo *Fuzzy C-Means* (FCM)

O algoritmo *fuzzy c-means* (BEZDEK; PAL, 1992) é um típico algoritmo de agrupamento, e tem sido amplamente utilizado em uma variedade de áreas de engenharia e científicas, tais como medicina, processamento de imagens, bioinformática, reconhecimento de padrões e mineração de dados (ZHANG; CHEN, 2003).

Suas principais características são facilidade de implementação e baixa complexidade computacional, a qual é linear para todos os parâmetros (NUOVO *et al.*, 2006; HöPPNER *et al.*, 1999). É também um algoritmo rápido e que não impõe restrições ao conjunto de dados, nem mesmo com relação ao tamanho. Na prática poucas iterações já fornecem boas aproximações para a solução final (HöPPNER *et al.*, 1999).

Esse algoritmo é o equivalente *nebuloso* do algoritmo de agrupamento *hard k-means* (XIE; BENI, 1991). É empregado para classificar um universo de amostras em categorias nebulosas, que são *hiper-esferas* do espaço p -dimensional $\mathbb{R}^p$, aproximadamente de mesmo

volume e caracterizadas pelos seus centros.

A letra c no nome do algoritmo refere-se ao número de grupos. Mas essa forma não é utilizada, e o c permanece intocado. Referente a esse fato é importante notar que o algoritmo é voltado para um número fixo de grupos, ou seja, ele não determina esse número (HöPPNER *et al.*, 1999).

O *fuzzy c-means* é uma técnica de análise de grupos baseada em *função objetivo*, na qual a base está em avaliar uma função que atribui a cada partição (grupo) um valor de qualidade, ou erro, que deve ser otimizado. A solução ideal é o particionamento em grupos que obtenha a melhor avaliação. Logo, o problema de particionamento pode ser equacionado como um problema de otimização (HöPPNER *et al.*, 1999).

O processo de cálculo do algoritmo *fuzzy c-means* é iterativo. O método tem início com uma suposição sobre os centros de cada agrupamento (protótipo) ou atribuindo-se para cada dado um grau de pertinência inicial em cada agrupamento. Sendo essa apenas uma suposição inicial, é na maioria das vezes incorreta.

Seja $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ um conjunto de dados, que em geral é um vetor de p características (atributos), $\mathbf{x}_j = (x_{j1}, \dots, x_{jp}) \in \mathbb{R}^p$ para todo $j \in \{1, 2, \dots, n\}$ e $\mathbb{R}^p$ um espaço p -dimensional.

O algoritmo FCM tem como propósito minimizar o índice de desempenho da pseudo-partição nebulosa J_m (ou função objetivo) que mede a distância entre os centros dos grupos e os elementos dentro dos grupos, desta forma, quanto menor seu valor, mais otimizada estará a pseudo-partição $\mathbf{U}$:

$$J_m(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^c \sum_{j=1}^n (\mu_{ij})^m (d_{ij})^2 \quad (3.2)$$

onde $\mathbf{V}$ é um conjunto de vetores que representa os c centros dos grupos (centróides), ou seja, $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_c\}$, $\mathbf{U}$ é a matriz de graus de pertinência nebulosos, μ_{ij} é a pertinência do dado j no grupo i e d_{ij} é a distância do dado j ao centro do grupo i .

Uma partição nebulosa ou uma pseudo-partição nebulosa em um conjunto finito de dados $\mathbf{X}$ ($\mathbf{U} = \{\mu_{1j}, \mu_{2j}, \dots, \mu_{cj}\}$), são normalmente denominadas c -partições nebulosas onde c indica o número de grupos nebulosos na partição, sendo que c deve ser um número inteiro positivo, $c \in [2, n]$.

$$\mathbf{U} = \begin{pmatrix} \mu_{11} & \mu_{12} & \dots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \dots & \mu_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{c1} & \mu_{c2} & \dots & \mu_{cn} \end{pmatrix}$$

O valor para $c = 1$ não está inserido no intervalo porque neste caso todos os n dados pertenceriam a um mesmo grupo, o único existente. O intervalo também exclui $c = n$, pois neste caso o número de centros seria igual ao número de dados, então, cada centro coincidiria exatamente com um dado e a pertinência seria igual a 1 neste grupo e zero nos demais, tornando-se um particionamento rígido (*crisp*). Não existe uma norma rígida para estabelecer o valor máximo de grupos (c_{max}) que pode ser aplicado ao método de agrupamento. De acordo com (PAL; BEZDEK, 1995) o que existe é uma regra simples que muitos pesquisadores utilizam onde $c_{max} \leq \sqrt{n}$.

Desta forma, a pseudo-partição nebulosa de $\mathbf{X}$ é a família de c subgrupos nebulosos de $\mathbf{X}$, que satisfaz as equações (3.3) e (3.4):

$$\sum_{i=1}^c \mu_i(\mathbf{x}_j) = 1, \forall j \in \{1, 2, \dots, n\} \quad (3.3)$$

sendo n o número de dados no conjunto total de dados $\mathbf{X}$. Assim, a soma de todos os graus de pertinência de um determinado dado deve ser igual a 1 nos c número de agrupamentos para todo $i \in \{1, 2, \dots, c\}$.

$$0 < \sum_{j=1}^n \mu_i(\mathbf{x}_j) < n, \forall i \in \{1, 2, \dots, c\} \quad (3.4)$$

A equação (3.4) garante que não existe nenhum grupo vazio ou que contenha todos os dados.

Para o cálculo da distância d_{ij} , normalmente utiliza-se como métrica a distância Euclideana, mostrada nas Equações 3.5 e 3.6:

$$d_{ij} = d(\mathbf{x}_j - \mathbf{v}_i) \quad (3.5)$$

$$d_{ij}^2 = \|\mathbf{x}_j - \mathbf{v}_i\|^2 = \sum_{k=1}^p (x_{jk} - v_{ik})^2 \quad (3.6)$$

O problema do agrupamento nebuloso é então, encontrar uma pseudo-partição e os centros dos agrupamentos associados que represente a estrutura dos dados da melhor forma possível. Este objetivo é alcançado quando as associações dos dados são fortes dentro do agrupamento e fracas entre os agrupamentos. Assim, define-se um índice de desempenho baseado nos centros dos agrupamentos. Dado uma pseudo-partição $\mathbf{U} = \{\mu_{1j}, \mu_{2j}, \dots, \mu_{cj}\}$

e o centro de agrupamento c , pode-se calcular os centros associados a cada partição:

$$\mathbf{v}_i = \frac{\sum_{j=1}^n (\mu_{ij})^m (\mathbf{x}_j)}{\sum_{j=1}^n (\mu_{ij})^m} \quad (3.7)$$

Onde μ_{ij} é o grau de inclusão do dado j no grupo i , para $\forall i \in \{1, 2, \dots, c\}$ e $m \in (1, \infty)$ é um número real que representa o fator de nebulosidade e define a faixa de nebulosidade que existe entre um grupo e outro.

Quando $m \rightarrow 1$ o *fuzzy c-means* converge para o seu histórico predecessor *hard k-means*, ou seja, a matriz $\mathbf{U}$ tende a ser rígida e, $m = 1$ levaria a uma matriz totalmente rígida. Quando $m \rightarrow \infty$, as pertinências tendem a $1/c$, ou seja os pontos têm o mesmo grau de inclusão em todos os grupos, desta forma a partição fica mais nebulosa.

A precisão pela qual o algoritmo classifica os dados depende fortemente do valor de m , que normalmente varia entre 1.5 e 2, mas às vezes agrupamentos melhores podem ser alcançados com valores maiores ou menores. A escolha ótima depende da base de dados que está sendo analisada, variando, portanto, de acordo com a aplicação (NUOVO *et al.*, 2006). A escolha de $m = 2$ tem sido amplamente aceita como uma boa escolha do índice de nebulosidade (HATHAWAY; BEZDEK, 2001).

Sumarizando-se, tem-se os seguintes passos para o algoritmo *fuzzy c-means*:

1. Dada uma base de dados contendo n dados (ou amostras), define-se o índice de nebulosidade $m \in (1, \infty)$, o número de agrupamentos c ($1 < c < n$), o número máximo de iterações t ($t = 0, 1, 2, \dots$), e um critério de parada (um pequeno valor positivo) $\varepsilon > 0$
2. No passo $t = 0$ inicializa-se a matriz de partições nebulosas $\mathbf{U}^{(0)} = [\mu_{ij}]$, aleatoria-

mente, respeitando-se a restrição dada pela Equação 3.3

3. Nos passos subseqüentes ($t > 0$) calcula-se os centros dos grupos $\mathbf{V}^{(t)}$, ou seja o vetor de protótipos $\mathbf{v}_i$ (dado pela Equação 3.7), para $i = 1, 2, \dots, c$, utilizando-se $\mathbf{U}^{(t-1)}$
4. Com os novos centros obtidos, atualiza-se os graus de pertinência μ_{ij} (Equação 3.8), obtendo-se $\mathbf{U}^{(t)}$

$$\mu_{ij} = \frac{1}{\sum_{r=1}^c \left(\frac{d^2(\mathbf{x}_j, \mathbf{v}_i)}{d^2(\mathbf{x}_j, \mathbf{v}_r)} \right)^{\left(\frac{1}{m-1}\right)}} \quad (3.8)$$

5. Repete-se os passos (3) e (4) até que um dos critérios de parada sejam satisfeitos, ou seja, $\|\mathbf{U}^{(t)} - \mathbf{U}^{(t-1)}\| < \varepsilon$, ou o número de iterações t seja atingido.

Outra versão do algoritmo consiste em inicializar-se $\mathbf{V}^{(0)}$ ao invés de $\mathbf{U}^{(0)}$. Neste caso, o critério de parada referente ao cálculo de ε seria: $\|\mathbf{V}^{(t)} - \mathbf{V}^{(t-1)}\| < \varepsilon$ (HATHAWAY; BEZDEK, 2001).

3.4 Análise de Dados Incompletos Utilizando FCM

O algoritmo *fuzzy c-means* tem sido amplamente utilizado, adaptado e generalizado (HATHAWAY; BEZDEK, 2001). No caso de valores que por um motivo ou outro não tenham sido coletados, ou tenham sido perdidos, durante a obtenção dos dados (dados incompletos), é possível integrar-se, dentro das iterações FCM, uma eficiente estimativa para os mesmos (NUOVO *et al.*, 2006).

Por outro lado, o *fuzzy c-means* não é diretamente aplicável para o caso de dados

incompletos. Em (HATHAWAY; BEZDEK, 2001) descreve-se quatro métodos para trabalhar com dados incompletos utilizando FCM, sendo que três consistem de novas adaptações para o FCM:

1. Estratégia dos Dados Completos - WDS (*Whole Data Strategy*)

Esta técnica utiliza um dos métodos descritos na Seção 2.3, *análise de caso completo*, onde exclui-se da base de dados todos os dados incompletos. Assim o algoritmo FCM, sem modificações, é aplicado na base de dados *completa* (estratégia referenciada como WDSFCM).

Esta estratégia é utilizada quando o percentual de dados incompletos é pequena. A sugestão é que deve-se utilizar WDS quando o percentual $n_p/n_s \geq 0.75$, onde n_p representa a quantidade total de atributos (ou características) que estão presentes na base de dados e $n_s = (\text{quantidade de dados na base de dados} \times \text{número de atributos})$.

2. Estratégia para Complementação Ótima - OCS (*Optimal Completion Strategy*)

Neste método os valores faltando são vistos como variáveis adicionais sobre as quais será feita a otimização com o objetivo de buscar o menor valor para a função objetivo J_m (Equação 3.2). Assim, a estratégia é atribuir valores aos atributos incompletos de forma que resulte no menor valor possível para J_m . Esta versão OCS do algoritmo FCM é referenciada como OCSFCM. O processo é iterativo.

3. Estratégia do Protótipo mais Próximo - NPS (*Nearest Prototype Strategy*)

É uma simples modificação do OCSFCM. Calculam-se as distâncias parciais e os valores dos atributos que faltam são calculados utilizando-se o protótipo (centróide)

mais próximo.

4. Estratégia das Distâncias Parciais - PDS (*Partial Distance Strategy*)

Quando o percentual de dados incompletos é alto, Dixon (1979) recomenda esta estratégia. Nesta técnica calcula-se a distância parcial utilizando-se os atributos que estão presentes.

Na versão PDS do algoritmo FCM (referenciada como PDSFCM) realizam-se duas modificações no algoritmo FCM, que são: altera-se o cálculo da distância (Equação 3.6) e dos centróides (Equação 3.7). Para isso, introduz-se um vetor de índices para dados incompletos (Equação 3.9), o qual possui o valor 0 no caso do valor do atributo estar faltando e 1 no caso do valor do atributo estar presente.

Esta técnica também é referenciada em (HATHAWAY; BEZDEK, 2001; TIMM; KLAWONN, 1998; TIMM; DÖRING; KRUSE, 2004). Neste trabalho utiliza-se este método para trabalhar com dados incompletos.

Para exemplificar a estratégia PDS, seja a base de dados ou conjunto total de dados, denotado por:

$$\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$$

Sejam também $\mathbf{X}_C$ o conjunto dos dados completos, $\mathbf{X}_P$ o conjunto de atributos que estão presentes e $\mathbf{X}_M$ o conjunto de atributos que não estão presentes, onde w representa o índice do atributo, j o índice do dado e i o índice do centróide:

$$\mathbf{X}_C = \{\mathbf{x}_j \in \mathbf{X} | \mathbf{x}_j \text{ é um dado completo}\}$$

$$\mathbf{X_P} = \{x_{jw} \text{ para } 1 \leq w \leq p, 1 \leq j \leq n |$$

o valor para x_{jw} está presente em $\mathbf{X}\}$

$$\mathbf{X_M} = \{x_{jw} = \text{NaN para } 1 \leq w \leq p, 1 \leq j \leq n |$$

o valor para x_{jw} não está presente em $\mathbf{X}\}$

Utilizando o exemplo dado em (HATHAWAY; BEZDEK, 2001), seja uma base de dados $\mathbf{X}$ contendo quatro dados ($n = 4$) com três atributos cada ($p = 3$):

$$\mathbf{X} = \begin{pmatrix} 2 & NaN & NaN \\ 3 & 2 & 4 \\ 2 & 1 & NaN \\ 1 & 1 & 2 \end{pmatrix}$$

Então,

$$\mathbf{X_C} = \begin{pmatrix} 3 & 2 & 4 \\ 1 & 1 & 2 \end{pmatrix}$$

$$\mathbf{X_M} = \{x_{12}, x_{13}, x_{33}\}$$

$$\mathbf{X_P} = \{x_{11} = 2, x_{21} = 3, x_{22} = 2, x_{23} = 4, x_{31} = 2, x_{32} = 1, x_{41} = 1, x_{42} = 1, x_{43} = 2\}$$

Assim, define-se os valores do vetor de índices incompletos $\mathbf{I}$:

$$I_{jw} = \begin{cases} 0, & \text{se } x_{jw} \in X_M \\ 1, & \text{se } x_{jw} \in X_P \end{cases} \quad (3.9)$$

Sendo,

$$\mathbf{I}_j = \sum_{w=1}^p I_{jw} \quad (3.10)$$

De acordo com o exemplo,

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

Utilizando o vetor de índices incompletos $\mathbf{I}_j$, obtém-se a versão modificada do algoritmo FCM, alterando-se a equação de cálculo dos centróides (Equação 3.7) para a Equação 3.11:

$$\mathbf{v}_i = \frac{\sum_{j=1}^n (\mu_{ij})^m (\mathbf{x}_j)(\mathbf{I}_j)}{\sum_{j=1}^n (\mu_{ij})^m \mathbf{I}_j} \quad (3.11)$$

ou,

$$v_{iw} = \frac{\sum_{j=1}^n (\mu_{ij})^m x_{jw} I_{jw}}{\sum_{j=1}^n (\mu_{ij})^m I_{jw}} \quad (3.12)$$

que representa o w -ésimo atributo do centro i ($i \in \{1, 2, \dots, c\}$) e, alterando-se também a equação da distância (Equação 3.6), que se torna:

$$d_{ij}^2 = \|\mathbf{x}_j - \mathbf{v}_i\|^2 = \frac{p}{\sum_{w=1}^p I_{jw}} \left(\sum_{w=1}^p I_{jw} (x_{jw} - v_{iw})^2 \right) \quad (3.13)$$

3.5 Medindo a Validade do Agrupamento

O objetivo de *agrupar* é descobrir, automaticamente, o intrínseco agrupamento de um conjunto de dados (LAW; JAIN, 2003), assim, a principal preocupação em um processo de agrupamento é revelar a organização dos padrões dentro de grupos *sensíveis*, os quais nos permitem descobrir similaridades e diferenças, assim como derivar inferências úteis sobre eles (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a).

Diferentes algoritmos de agrupamento já foram propostos para diferentes aplicações e tamanhos de bases de dados (JAIN; MURTY; FLYNN, 1999; LAW; JAIN, 2003; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a). A aplicação de um algoritmo a uma base de dados tem como objetivo, assumindo-se que a base de dados oferece uma tendência ao agrupamento, descobrir suas partições naturais. Entretanto, o processo de agrupamento é considerado um processo não supervisionado, tendo em vista que não são pré-definidas classes e nem exemplos que mostrem que tipo de relações desejáveis devem ser válidas entre os dados. Desta forma, os vários algoritmos de agrupamento baseiam-se em algumas suposições para definir o particionamento de uma base de dados, o que pode levar a resultados diferentes dependendo das características do conjunto de dados, como por exemplo geometria e densidade de distribuição dos grupos, e dos valores dos parâmetros

de entrada (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a).

Assim, adicionado às diferenças entre os resultados dos algoritmos, deve-se considerar que os algoritmos de agrupamentos produzem um modelo de particionamento para uma base de dados, quer existam ou não, tornando-se necessária a validação de tais agrupamentos, seja verificando se o modelo de agrupamento obtido é o que mais se adequa ao conjunto de dados ou avaliando-se a qualidade do agrupamento.

O procedimento de avaliação que pode fornecer uma resposta quantitativa para estas questões é denominado *validade do agrupamento* (*cluster validity*), o qual é o objetivo de muitos esforços de pesquisadores (LAW; JAIN, 2003; PAL; BEZDEK, 1995; XIE; BENI, 1991; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002b).

Em termos gerais existem três métodos para investigar a validade de um agrupamento que são os critérios de avaliação externos, internos e relativos. Nos critérios de avaliação externos a avaliação dos resultados do algoritmo de agrupamento baseia-se na comparação entre o particionamento obtido pelo algoritmo com uma estrutura de grupos pré-especificada que é imposta ao conjunto de dados e reflete nossa intuição sobre a estrutura que a mesma deve possuir. Nos critérios internos, o objetivo é avaliar o agrupamento resultante, utilizando-se somente características intrínsecas da base de dados, ou seja, baseiam-se em medidas de avaliação internas. No método de critérios relativos, a idéia básica é avaliar a estrutura do agrupamento por meio da comparação do mesmo com outros esquemas de agrupamentos resultantes do mesmo algoritmo, mas com diferentes valores para os parâmetros (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002b; LAW; JAIN, 2003)

A reamostragem com reposição (*bootstrapping* (EFRON; TIBSHIRANI, 1993)) pode

ser usada para estimar índices relativos, comparando-se o agrupamento resultante para uma dada amostra da base de dados e um agrupamento de referência. Assim, utiliza-se os métodos baseados em reamostragem para verificar onde a base de dados possui uma estrutura de agrupamentos ou onde o resultado é um artefato do algoritmo e também para seleccionar-se o modelo do número de grupos (LAW; JAIN, 2003; BORGELT, 2005). Neste trabalho utilizou-se o método interno de comparação combinado com o critério relativo reamostragem *bootstrap*.

3.5.1 Comparando Matrizes de Partição

Medidas de avaliação relativas para grupos comparam duas matrizes de partição $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$, onde $\mathbf{U}$ é uma matriz ($c \times n$) e $\mathbf{U} = (\mu_{ij})$. Lembrando que: c é o número de grupos e n é o número de dados contidos na base de dados, e ainda, $(1 \leq i \leq c)$ e $(1 \leq j \leq n)$. As matrizes de partição a serem comparadas podem ser, por exemplo, uma matriz obtida pelo algoritmo de agrupamento e a outra uma matriz de referência (modelo). Faz-se a comparação diretamente, aos pares, ou seja, compara-se cada par de linhas, sendo uma de cada matriz de partição (índice i refere-se à $\mathbf{U}^{(1)}$ e o índice r refere-se à $\mathbf{U}^{(2)}$). Formalmente, utiliza-se uma tabela de contingência 2×2 (Tabela 3.3) para cada par de linhas, uma de cada matriz de partição.

Tabela 3.3 – Tabela de contingência para comparar duas linhas de duas matrizes de partição *crisp* (Adaptado de (BORGELT, 2005)).

	$\mu_{rj}^{(2)} = 1$	$\mu_{rj}^{(2)} = 0$	$\sum$
$\mu_{ij}^{(1)} = 1$	$n_{11}^{(i,r)}$	$n_{10}^{(i,r)}$	$n_{1.}^{(i,r)}$
$\mu_{ij}^{(1)} = 0$	$n_{01}^{(i,r)}$	$n_{00}^{(i,r)}$	$n_{0.}^{(i,r)}$
$\sum$	$n_{.1}^{(i,r)}$	$n_{.0}^{(i,r)}$	n

Assim, para cada par de índices de grupo, $(i, r) \in \{1, \dots, c\}^2$ e cada par $(a, b) \in \{0, 1\}^2$,

calcula-se:

$$n_{ab}^{(i,r)}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \sum_{j=1}^n \left((1-a) + (2a-1)\mu_{ij}^{(1)} \right) \cdot \left((1-b) + (2b-1)\mu_{rj}^{(2)} \right) \quad (3.14)$$

Substituindo-se os valores de a e b , tem-se:

$$n_{11}^{(i,r)}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \sum_{j=1}^n \mu_{ij}^{(1)} \cdot \mu_{rj}^{(2)} \quad (3.15)$$

$$n_{01}^{(i,r)}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \sum_{j=1}^n (1 - \mu_{ij}^{(1)}) \cdot \mu_{rj}^{(2)} \quad (3.16)$$

$$n_{10}^{(i,r)}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \sum_{j=1}^n \mu_{ij}^{(1)} \cdot (1 - \mu_{rj}^{(2)}) \quad (3.17)$$

$$n_{00}^{(i,r)}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \sum_{j=1}^n (1 - \mu_{ij}^{(1)}) \cdot (1 - \mu_{rj}^{(2)}) \quad (3.18)$$

Sendo que o termo n_{11} representa os verdadeiros positivos, n_{01} os falso positivos, n_{10} os falso negativos, e n_{00} os verdadeiros negativos. Por exemplo, no agrupamento rígido (ou *crisp*) n_{11} representa o número de dados que estão no i -ésimo grupo da partição $\mathbf{U}^{(1)}$ e no r -ésimo grupo da partição $\mathbf{U}^{(2)}$. Onde a operação booleana 'e' é formalmente expressa pelo produto nas Equações 3.14 a 3.18.

No caso de agrupamentos nebulosos ($\mu \in [0, 1]$) e considerando-se que o produto expressa a operação booleana 'e' (*and*), pode-se utilizar o conceito de *t-norma* (Seção 3.3.2) para calcular-se os termos definidos nas Equações 3.15 a 3.18.

Por exemplo, reescrevendo-se a Equação 3.14, temos:

$$n_{ab}^{(i,r)}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \sum_{j=1}^n \left(\underbrace{(1-a) + (2a-1)\mu_{ij}^{(1)}}_x \right) \cdot \left(\underbrace{(1-b) + (2b-1)\mu_{rj}^{(2)}}_y \right)$$

Desta forma, considerando-se o produto entre x e y tem-se a t -norma $\top(x, y)$. Aplicando-se a t -norma Zadeh (Tabela 3.2) na Equação 3.16, tem-se:

$$x = (1 - \mu_{ij}^{(1)}) \quad \text{e} \quad y = \mu_{rj}^{(2)}$$

Assim:

$$\top_{min}(x, y) = \min((1 - \mu_{ij}^{(1)}), \mu_{rj}^{(2)}) \quad (3.19)$$

(Note-se que mostra-se aqui apenas um dos termos referentes ao somatório).

Calculados os números $n_{11}^{(i,r)}$, $n_{01}^{(i,r)}$, $n_{10}^{(i,r)}$ e $n_{00}^{(i,r)}$, pode-se utilizá-los no cálculo de medidas de comparação binária, tais como F_1 (RIJSBERGEN, 1979) e classificação cruzada (ou *accuracy*). Quanto maior o valor dessas duas medidas, mais similares são as matrizes de partição $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$.

Antes de definir-se a medida F_1 , faz-se necessário o detalhamento de duas medidas, as quais são necessárias no cálculo de F_1 , que são *precisão* e *recuperação*, definidas em (RIJSBERGEN, 1979). Originalmente, essas medidas são calculadas para cada grupo e descrevem o quão bem um grupo é identificado pelo método classificador. A comparação é feita entre o grupo real e o predito pelo classificador (problema dois grupos) (BORGELT,

2005).

Para calcular *precisão* e *recuperação*, faz-se uma tabela de contingência 2 x 2 (Tabela 3.4) para cada grupo, baseadas nos grupos verdadeiros v associados com os dados x_j , com $1 \leq j \leq n$, e os grupos o que são preditos pelo classificador. Os elementos para a tabela de contingência é qualquer grupo r ($\forall r : 1 \leq r \leq c$)

Tabela 3.4 – Tabela de contingência para o cálculo de *precisão* e *recuperação* (Adaptado de (BORGELT, 2005)).

	$o = r$	$o \neq r$	$\sum$
$v = r$	$n_{11}^{(r)}$	$n_{10}^{(r)}$	$n_{1.}^{(r)}$
$v \neq r$	$n_{01}^{(r)}$	$n_{00}^{(r)}$	$n_{0.}^{(r)}$
$\sum$	$n_{.1}^{(r)}$	$n_{.0}^{(r)}$	n

Assim, dado um grupo r , a *precisão* representa a taxa de verdadeiros positivos para todos os dados classificados como grupo r , ou seja, a fração de dados que o classificador agrupa corretamente:

$$\pi_r = \frac{n_{11}^{(r)}}{n_{11}^{(r)} + n_{01}^{(r)}} \quad (3.20)$$

A *recuperação* representa a taxa de verdadeiros positivos para todos os dados que atualmente pertencem ao grupo r , isto é, é a fração de dados do grupo r que são identificados pelo classificador:

$$\rho_r = \frac{n_{11}^{(r)}}{n_{11}^{(r)} + n_{10}^{(r)}} \quad (3.21)$$

Quando deseja-se comparar vários grupos (em toda a base de dados), é preciso calcular a média das medidas individuais. Pode-se utilizar, por exemplo a *média-macro* (*macro-*

averaging) (BORGELT, 2005):

$$\pi_{macro} = \frac{1}{c} \sum_{r=1}^c \frac{n_{11}^{(r)}}{n_{01}^{(r)} + n_{11}^{(r)}} \quad (3.22)$$

$$\rho_{macro} = \frac{1}{c} \sum_{r=1}^c \frac{n_{11}^{(r)}}{n_{10}^{(r)} + n_{11}^{(r)}} \quad (3.23)$$

- A medida de comparação binária F_1 :

A medida F_1 é a média harmônica entre *precisão* e *recuperação* (RIJSBERGEN, 1979):

$$F_1 = \frac{\pi \rho}{\left(\frac{\pi + \rho}{2}\right)} = \frac{2\pi \rho}{\pi + \rho} \quad (3.24)$$

Quando deseja-se encontrar a linha da matriz de partição $\mathbf{U}^{(1)}$ que mais se iguale a uma linha da matriz de partição $\mathbf{U}^{(2)}$, a melhor forma é utilizar-se permutação de todas as linhas da matriz ($\mathbf{U}$ é uma matriz $c \times n$), ou seja, a permutação de todos os grupos, assim, calcula-se *precisão* e *recuperação* por:

$$\pi = \max_{\zeta \in \Pi(c)} \frac{1}{c} \sum_{i=1}^c \frac{n_{11}^{(i, \zeta(i))}}{n_{01}^{(i, \zeta(i))} + n_{11}^{(i, \zeta(i))}} \quad (3.25)$$

$$\rho = \max_{\zeta \in \Pi(c)} \frac{1}{c} \sum_{i=1}^c \frac{n_{11}^{(i, \zeta(i))}}{n_{10}^{(i, \zeta(i))} + n_{11}^{(i, \zeta(i))}} \quad (3.26)$$

Assim, tem-se:

$$F_1(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \max_{\zeta \in \Pi(c)} \frac{1}{c} \sum_{i=1}^c \frac{2\pi_{i,\zeta(i)}\rho_{i,\zeta(i)}}{\pi_{i,\zeta(i)} + \rho_{i,\zeta(i)}} \quad (3.27)$$

Onde $\Pi(c)$ representa o conjunto de todas as permutações de linhas c da matriz de partição.

Por exemplo, para um agrupamento em cinco grupos ($c = 5$), teríamos a comparação do grupo $c = 1$ da matriz de partição $\mathbf{U}^{(1)}$ com os grupos $c = 1, 2, 3, 4, 5$ da matriz de partição $\mathbf{U}^{(2)}$. A seguir compara-se o grupo $c = 2$ da matriz de partição $\mathbf{U}^{(1)}$ com os grupos $c = 1, 2, 3, 4, 5$ da matriz de partição $\mathbf{U}^{(2)}$, e assim por diante até que se compare o último grupo de $\mathbf{U}^{(1)}$ (no caso $c = 5$) com todos os grupos de $\mathbf{U}^{(2)}$.

- **A medida de comparação binária classificação cruzada (Acc.):**

Indica o quanto um teste de classificação binária identifica corretamente ou exclui uma condição, ou seja, é a proporção de resultados verdadeiros (verdadeiros positivos e verdadeiros negativos) em relação às matrizes de partição $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$.

$$Acc.(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \max_{\zeta \in \Pi(c)} \frac{1}{cn} \sum_{i=1}^c (n_{00}^{(i,\zeta(i))} + n_{11}^{(i,\zeta(i))}) \quad (3.28)$$

- **A medida de comparação erro médio quadrático (Diff.):**

Representa o valor esperado do quadrado do erro (diferença) entre as pertinências de $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$:

$$Diff.(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \min_{\zeta \in \Pi(c)} \frac{1}{cn} \sum_{i=1}^c \sum_{j=1}^n \left(\mu_{ij}^{(1)} - \mu_{\zeta(i)j}^{(2)} \right)^2 \quad (3.29)$$

3.5.2 Comparando Matrizes de Coincidência

Matriz de coincidência é uma matriz $n \times n$, e indica para cada par de dados onde esses dados estão em um mesmo grupo ou não. Calcula-se uma matriz de coincidência Ψ a partir de uma matriz de partição $\mathbf{U}$, sendo $\Psi = (\psi_{jl})$, onde $1 \leq j, l \leq n$ (BORGELT, 2005):

$$\psi_{jl} = \sum_{i=1}^c \mu_{ij} \cdot \mu_{il} \quad (3.30)$$

Após calcular-se $\Psi^{(1)}$ e $\Psi^{(2)}$, pode-se gerar os números que comparam as coincidências na distribuição dos dados dentro dos grupos. Sendo cada par $(a, b) \in \{0, 1\}^2$, tem-se:

$$N_{ab}(\Psi^{(1)}, \Psi^{(2)}) = \sum_{j=2}^n \sum_{l=1}^{j-1} \left((1-a) + (2a-1)\psi_{jl}^{(1)} \right) \cdot \left((1-b) + (2b-1)\psi_{jl}^{(2)} \right) \quad (3.31)$$

Substituindo-se os valores de a e b , tem-se:

$$N_{11}(\Psi^{(1)}, \Psi^{(2)}) = \sum_{j=2}^n \sum_{l=1}^{j-1} \psi_{jl}^{(1)} \cdot \psi_{jl}^{(2)} \quad (3.32)$$

$$N_{10}(\Psi^{(1)}, \Psi^{(2)}) = \sum_{j=2}^n \sum_{l=1}^{j-1} \psi_{jl}^{(1)} \cdot (1 - \psi_{jl}^{(2)}) \quad (3.33)$$

$$N_{01}(\Psi^{(1)}, \Psi^{(2)}) = \sum_{j=2}^n \sum_{l=1}^{j-1} (1 - \psi_{jl}^{(1)}) \cdot \psi_{jl}^{(2)} \quad (3.34)$$

$$N_{00}(\Psi^{(1)}, \Psi^{(2)}) = \sum_{j=2}^n \sum_{l=1}^{j-1} (1 - \psi_{jl}^{(1)}) \cdot (1 - \psi_{jl}^{(2)}) \quad (3.35)$$

Sendo que N_{11} representa o número de pares de dados que estão em um mesmo grupo em ambas as partições, N_{10} e N_{01} representam o número de pares de dados que estão em um mesmo grupo em uma partição mas em grupos diferentes na outra partição e, N_{00} indica o número de pares de dados que estão grupos diferentes em ambas as partições.

Da mesma forma que citou-se na Seção 3.5.1, o produto representado na Equação 3.31 também pode ser substituído por uma *t-norma*, quando trabalha-se com conjuntos nebulosos.

Note-se que como essa substituição pela *t-norma* pode ser feita tanto na Equação 3.30 quanto na Equação 3.31, quando trabalha-se com matrizes de coincidência pode-se fazer várias combinações entre a primeira e segunda *t-normas*, ou seja, pode-se utilizar na Equação 3.30 a *t-norma* produto e na Equação 3.31 também a *t-norma* produto, ou utilizar-se *t-norma* mínimo e *t-norma* produto nas Equações 3.30 e 3.31, respectivamente, e assim por diante.

De posse dos números N_{11} , N_{01} , N_{10} e N_{00} pode-se calcular várias medidas de comparação, tais como:

- **Coefficiente de Jaccard**

Define a similaridade entre duas partições ($\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$). Esse índice calcula a probabilidade de que dados pertencentes ao mesmo grupo em uma das partições também

pertencam ao mesmo grupo na outra partição (BOUTIN; HASCOËT, 2004), ignorando a informação negativa N_{00} , que representa os pares que são atribuídos para grupos diferentes em ambas as partições (BORGELT, 2005).

$$Jaccard = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \quad (3.36)$$

- **Índice Randômico**

Esse índice calcula a proporção de concordância entre as partições. Assim, calcula a probabilidade de que os dados que pertençam a um grupo na partição $\mathbf{U}^{(1)}$ também pertençam ao mesmo grupo na partição $\mathbf{U}^{(2)}$ ou, que pertençam a grupos diferentes (BOUTIN; HASCOËT, 2004; JAIN; DUBES, 1988).

$$Rand. = \frac{N_{11} + N_{00}}{N_{11} + N_{01} + N_{10} + N_{00}} \quad (3.37)$$

- **Índice de Fowlkes & Mallows**

Foi proposto por Fowlkes e Mallows (1983):

$$Fowlkes-Mallows = \sqrt{\underbrace{\frac{N_{11}}{N_{11} + N_{10}}}_x \cdot \underbrace{\frac{N_{11}}{N_{11} + N_{01}}}_y} \quad (3.38)$$

Onde o termo x indica a probabilidade de que determinados dados pertençam a um mesmo grupo na partição $\mathbf{U}^{(1)}$ se eles fazem parte de um mesmo grupo na partição $\mathbf{U}^{(2)}$. O termo y é a probabilidade de que dados pertençam ao mesmo grupo em $\mathbf{U}^{(2)}$ se eles pertencem ao mesmo grupo em $\mathbf{U}^{(1)}$ (BOUTIN; HASCOËT, 2004).

Reescrevendo a Equação 3.38,

$$Fowlkes-Mallows = \frac{N_{11}}{\sqrt{(N_{11} + N_{10})(N_{11} + N_{01})}} \quad (3.39)$$

- **Índice Hubert**

Foi proposto por (HUBERT; ARABIE, 1985). É uma normalização do índice randômico para corrigir a estatística que ocorre *ao acaso*. Desta forma, o valor máximo é 1 quando é alcançada uma perfeita concordância e 0 (se o valor das partições são selecionadas selecionadas aleatoriamente (BOUTIN; HASCOËT, 2004; JAIN; DUBES, 1988)). É também chamado de estatística Phi.

Pode ser interpretado como a raiz quadrada da medida qui-quadrado (χ^2). Valores altos indicam alta probabilidade de aceitar a partição (BORGELT, 2006).

$$Hubert = \frac{N_{00}N_{11} - N_{10}N_{01}}{\sqrt{(N_{11} + N_{10})(N_{11} + N_{01})(N_{01} + N_{00})(N_{10} + N_{00})}} \quad (3.40)$$

- **Índice Randômico Ajustado**

Freqüentemente utiliza-se medidas que são correções de medidas existentes. Uma correção bastante comum é a normalização do índice correspondente para que ele apresente o valor 1 quando a partição casa perfeitamente com a partição real (ou modelo) e o valor 0 quando as partições são selecionadas ao acaso (JAIN; DUBES, 1988; GORDON, 1999).

$$Rand. Aj. = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{2(N_{11}N_{00}) + (N_{11}N_{00})(N_{01} + N_{10}) + N_{10}^2 + N_{01}^2} \quad (3.41)$$

Neste trabalho utilizou-se todas essas medidas na comparação entre matrizes de coincidência, sendo que para todas utilizou-se os valores máximos obtidos para a análise do número de grupos.

3.6 Reamostragem *Bootstrap*

Técnicas de reamostragem foram propostas há várias décadas, mas devido a serem métodos estatísticos computacionalmente intensivos, sua difusão só foi possível após o advento dos computadores modernos, com grande capacidade de processamento e custo relativamente baixo.

O procedimento *bootstrap* é uma técnica de reamostragem, bastante utilizada em diferentes situações estatísticas, que procura substituir a análise estatística teórica utilizando o poder da computação. O termo *bootstrap* surgiu da frase "*to pull oneself up one's bootstrap*" retirado do texto *Adventures of Baron Munchausem* de Rudolpf Erick Raspe, século XVII, que dizia: "The Baron had fallen to the of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps". Descreve-se nesta parte do texto a cena em que o Barão Munchausem ao se ver afundando em um lago e sentindo que tudo estava perdido, acha que conseguirá emergir puxando os cadarços de seus próprios sapatos (EFRON; TIBSHIRANI, 1993).

A base desta técnica é a obtenção de um novo conjunto de dados por reamostragem do conjunto de dados original (EFRON; TIBSHIRANI, 1993). Assim, explica-se a metáfora *bootstrap*, que refere-se ao fato de que a base de dados é usada em sua própria análise estatística, de modo que, todo resultado obtido depende diretamente da base de dados original.

Os métodos de reamostragem se diferenciam, principalmente, pela forma em que as amostras são extraídas do conjunto original de dados, que pode ser com reposição ou sem reposição. Na amostragem com reposição obtém-se um dado da base de dados para montar a amostra e então coloca-se o mesmo de volta para ser usado posteriormente. Assim, uma mesma amostra pode conter dados repetidos. Na amostragem sem reposição, uma vez que obtém-se um determinado dado para compor a amostra, este torna-se indisponível, ou seja não se pode mais utilizá-lo para compor esta amostra.

A inferência estatística procura estabelecer as propriedades da população, partindo-se de uma amostra aleatória retirada da mesma. Assim, o método *bootstrap* obtém sua amostra via amostragem com reposição, para revelar algum padrão estrutural presente na base de dados. A idéia básica é de que os dados em si, vistos como distribuição de freqüências, representam a melhor imagem disponível da distribuição de freqüências da qual eles são amostrados (CROWLEY, 1992).

Seja uma amostra aleatória de tamanho n obtida de uma distribuição de probabilidade F :

$$F \rightarrow (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n) \quad (3.42)$$

A função de distribuição empírica $\hat{F}$ (o símbolo chapéu indica quantidades calculadas a partir dos dados observados), é definida como sendo a distribuição discreta que atribui a probabilidade $1/n$ para cada valor observado $\mathbf{x}_j$, onde $j = 1, 2, \dots, n$ (EFRON; TIBSHIRANI, 1993).

O método *bootstrap* depende da noção de *amostra bootstrap*. Uma amostra *bootstrap*

é definida com sendo uma amostra aleatória de tamanho n retirada de F :

$$\hat{F} \rightarrow (\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_n^*) \quad (3.43)$$

assim,

$$\mathbf{x}^* = (\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_n^*) \quad (3.44)$$

desta forma, os dados *bootstrap* $\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_n^*$ formam uma amostra aleatória de tamanho n retirada da base de dados de n objetos. A notação utilizando-se o asterisco significa que $\mathbf{x}^*$ não é um novo conjunto de dados, mas apenas uma versão reamostrada de $\mathbf{x}$, onde os dados podem aparecer zero vezes, uma vez, duas vezes, etc. Assim, pode-se ter, por exemplo, $\mathbf{x}^* = (\mathbf{x}_2, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_2, \dots, \mathbf{x}_4)$, onde $\mathbf{x}_1^* = \mathbf{x}_2$, $\mathbf{x}_2^* = \mathbf{x}_4$, $\mathbf{x}_3^* = \mathbf{x}_5$, $\mathbf{x}_4^* = \mathbf{x}_2$ e $\mathbf{x}_n^* = \mathbf{x}_4$ (EFRON; TIBSHIRANI, 1993).

O método *bootstrap* é baseado no princípio *plug-in*, um método simples de estimar parâmetros a partir de amostras. Seja um parâmetro θ obtido aplicando-se algum procedimento de avaliação numérica $t(\cdot)$ da função de distribuição F :

$$\theta = t(F) \quad (3.45)$$

A estimativa *plug-in* de um parâmetro $\theta = t(F)$ é definida como sendo:

$$\hat{\theta} = t(\hat{F}) \quad (3.46)$$

Assim, por este princípio, estima-se a função $\theta = t(F)$ da distribuição de probabilidade

F pela mesma função da distribuição empírica $\hat{F}$, $\hat{\theta} = t(\hat{F})$. A vantagem deste método está em produzir-se tendências e erros-padrões de forma automática, não importando a complexidade de θ (EFRON; TIBSHIRANI, 1993).

3.7 Conclusão

A análise de agrupamentos é uma importante ferramenta utilizada em diversas áreas sendo útil no entendimento da natureza complexa das relações multivariadas e que possui a característica de não se fazer suposições com relação ao número de grupos ou à estrutura dos mesmos. Mas é preciso que se tenha em mente que os particionamentos gerados por um algoritmo de agrupamento devem sempre ser validados, para que se verifique a consistência dos resultados.

A lógica nebulosa apresenta-se mais adequada, que a lógica booleana, para tratar problemas do mundo real, pois são capazes de modelar valores *intermediários*. Um algoritmo de agrupamento nebuloso bastante utilizado é o *fuzzy c-means* o qual pode ser modificado para trabalhar com bases de dados incompletas, bases essas que são uma realidade no cotidiano de pesquisadores.

Capítulo 4

Materiais e Métodos

Neste trabalho foram analisadas oito bases de dados utilizando-se a técnica de reamostragem com reposição (*bootstrapping*) associada com o método de agrupamento de dados nebuloso *fuzzy c-means*, modificado para trabalhar com dados incompletos. Utilizou-se o sistema operacional Linux, sendo que os programas foram desenvolvidos utilizando-se o software MATLAB®(MATHWORKS, 1998) e os experimentos foram realizados em um computador Pentium Dual Core E2160, 1.8 GHz, HD de 160 GB e 1 GB de RAM.

Este capítulo é composto por duas seções. Na Seção 4.1 descreve-se as bases de dados utilizadas, que são compostas por seis bases de dados de validação e duas bases de dados de aplicação que fazem parte de um banco de dados genotípico. Na Seção 4.2 descreve-se o método de agrupamento implementado e os experimentos realizados.

4.1 Bases de Dados Utilizadas

Nesta seção são descritas as bases de dados que foram objetos de teste para o método proposto. Objetivando-se a validação do método, primeiramente as análises foram

realizadas em bases de dados com números de grupos conhecidos, realizando-se a seguir a checagem entre o resultado obtido e a informação que se tem sobre o número de grupos em cada base de dados. Após esta validação inicial aplicou-se o método em duas bases de dados (Brady 1 e Brady 2) que compõem um banco de dados genotípico de bactérias simbióticas fixadoras de nitrogênio, coletivamente chamadas de rizóbios e cuja taxonomia tem sido objeto de estudo de muitos pesquisadores (MILAGRE, 2003; MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006; TALAVERA, 2007; CHRIST, 2007; LLERENA, 2008).

Na Tabela 4.1 faz-se um sumário das características das bases de dados utilizadas neste trabalho onde descreve-se o número de grupos, o número de atributos (ou características) em cada dado, o número de dados (número total de itens de dados) e o número de dados em cada grupo, sendo que em relação às bases de dados Brady 1 e Brady 2 os pontos de interrogação indicam que as mesmas encontram-se em estudo, portanto ainda não apresentam números definidos.

4.1.1 Redução de Dimensionalidade

De acordo com a Tabela 4.1 pode-se observar que existem bases de dados (Artificial 3, *iris*, *Brady 1* e *Brady 2*) que possuem mais de três atributos (dimensões), portanto não sendo possível representá-las graficamente de forma convencional. Como neste trabalho é importante que se possa visualizar a representação gráfica das bases de dados para uma melhor compreensão da estrutura das mesmas, utilizou-se os resultados obtidos por (LLERENA, 2008) para justificar a elaboração dos gráficos utilizando somente três dimensões. Esses resultados também podem ser utilizados como critério de escolha para determinação do número de atributos a serem utilizados no agrupamento, caso a base de dados possua um grande número de dimensões (*Brady 1* e *Brady 2*), pois baseado

nessa análise pode-se reduzir o número de atributos e essa redução levaria à diminuição do tempo de processamento sem afetar significativamente os resultados.

Tabela 4.1 – Sumário das características das oito bases de dados utilizadas neste trabalho. Onde o símbolo '?' indica *em estudo*.

<i>Base de Dados</i>	<i>Número de Grupos</i>	<i>Número de atributos</i>	<i>Número de dados</i>	<i>Número de dados por grupo</i>
Artificial 1	4	2	400	100
Artificial 2	7	2	350	50
Artificial 3	4	4	2000	492-482-523-503
Artificial 4	5	3	1000	190-207-209-206-188
wine	3	3	178	59-71-48
iris	3	4	150	50-50-50
Brady 1	?	20	119	?
Brady 2	?	15	119	?

Em (LLERENA, 2008) utilizou-se o Escalonamento Multidimensional (MDS - *Multidimensional Scaling*) que refere-se a um conjunto de técnicas que têm por finalidade representar objetos em um baixo espaço dimensional, tendo como entrada a informação de proximidade entre pares (matriz de distância) e como saída uma representação de vetores de coordenadas que são um conjunto de pontos em um espaço euclidiano cuja matriz de distância seja próxima da matriz de distância original.

A matriz de vetores gerada possui coordenadas ordenadas em ordem decrescente de acordo com a sua importância, desta forma, as primeiras coordenadas agregam a maior variância entre os dados e esta informação pode ser utilizada para selecionar-se o número adequado de coordenadas a serem consideradas no agrupamento e na elaboração dos gráficos somente em três dimensões (TALAVERA, 2007).

4.1.2 Bases de Dados de Validação

Base de Dados Artificial 1

Esta base de dados consiste de um conjunto bidimensional artificial contendo 400 dados e 2 atributos divididos em 4 grupos.

Os atributos foram gerados através da variação controlada de incrementos para as duas coordenadas (atributos), de forma que atingissem a distribuição de dados desejada. Por exemplo, os dados representados em vermelho na Figura 4.1 foram construídos utilizando-se inicialmente a função *rand* (MATHWORKS, 1998) para gerar aleatoriamente 100 dados com dois atributos cada (Equação 4.1). A seguir, para todos os dados (indicado pelo símbolo :) o valor do atributo 1 foi acrescido de 1.2 (Equação 4.2) e o valor do atributo 2 foi acrescido de 1.6 (Equação 4.3).

$$data1 = rand(100, 2) \quad (4.1)$$

$$data1(:, 1) = data1(:, 1) + 1.2 \quad (4.2)$$

$$data1(:, 2) = data1(:, 2) + 1.6 \quad (4.3)$$

Na Figura 4.1 pode-se visualizar a estrutura dessa base de dados a qual possui quatro grupos perfeitamente separados.

Figura 4.1 – Dados artificiais bidimensionais (Artificial 1) contendo 400 dados e 4 grupos, sendo que cada grupo contém 100 dados.

Base de Dados Artificial 2

Esta base de dados também é um conjunto bidimensional artificial. Contém 350 dados, 7 grupos e dois atributos. Os atributos também foram gerados, como na Artificial 1, através da variação controlada de incrementos para as duas coordenadas (atributos), de forma que atingissem a distribuição de dados desejada. Exemplificando-se, os dados em azul escuro na Figura 4.2 foram construídos utilizando-se as Equações 4.4, 4.5 e 4.6.

$$data1 = rand(50, 2) \quad (4.4)$$

$$data1(:, 1) = data1(:, 1) + 2.2 \quad (4.5)$$

$$data1(:, 2) = data1(:, 2) + 1.2 \quad (4.6)$$

A Figura 4.2 traz a representação desses dados e pode-se notar que os grupos encontram-se também claramente separados.

Figura 4.2 – Dados artificiais bidimensionais (Artificial 2) contendo 350 dados e 7 grupos, sendo que cada grupo contém 50 dados.

Base de Dados Artificial 3

A terceira base de dados é um conjunto artificial que contém 2000 dados, 4 grupos e 4 atributos gerados aleatoriamente a partir de uma mistura de quatro distribuições gaussianas (TALAVERA, 2007). A Tabela 4.2 traz os parâmetros para cada uma das distribuições.

Como esta base de dados possui 4 atributos não é possível representá-la graficamente de forma convencional. Foi aplicado então o método de redução de dimensionalidade de dados multivariados que mapeia os dados de um espaço de alta dimensionalidade para outro espaço de baixa dimensionalidade (maiores detalhes em (LLERENA, 2008)).

Tabela 4.2 – Parâmetros das quatro distribuições gaussianas utilizadas na construção da base de dados Artificial 3 (Adaptado de (TALAVERA, 2007)).

Parâmetros	Componentes			
	G1	G2	G3	G4
Proporção	0.25	0.24	0.26	0.25
Média X_1	0.0584	0.7276	0.4014	0.3124
Média X_2	0.8979	0.4873	0.5106	0.2555
Média X_3	0.9803	0.6801	0.0512	0.6078
Média X_4	0.6578	0.3650	0.9957	0.4415
Variância	0.0909	0.0937	0.1026	0.0845

A Figura 4.3 traz a fração da variância que representa cada autovalor referente aos vetores resultantes, ordenados em ordem crescente de importância, conforme descrito na Seção 4.1.1. Como pode-se notar os três primeiros atributos somam aproximadamente 97% da variância total.

Figura 4.3 – Análise de autovalores e variância da matriz de vetores resultantes de MDS (LLERENA, 2008) referentes à base de dados Artificial 3.

Na Figura 4.4 plota-se os dados Artificial 3, utilizando-se os três primeiros atributos. Pode-se notar que as classes são claramente separadas. Para a aplicação do método proposto foram utilizados os dados originais com quatro atributos.

Figura 4.4 – Dados artificiais (Artificial 3) contendo 2000 dados, 4 grupos e 4 atributos, sendo que cada grupo contém 492, 482, 523 e 503 dados.

Base de Dados Artificial 4

Esta base de dados é formada por 1000 dados artificiais distribuídos em 5 grupos e possuindo 3 atributos gerados aleatoriamente a partir de uma mistura de cinco distribuições gaussianas (TALAVERA, 2007). A Tabela 4.3 traz os parâmetros para cada uma das distribuições.

Tabela 4.3 – Parâmetros das cinco distribuições gaussianas utilizadas na construção da base de dados Artificial 4 (Adaptado de (TALAVERA, 2007)).

Parâmetros	Componentes				
	G1	G2	G3	G4	G5
Proporção	<i>0.19</i>	<i>0.21</i>	<i>0.21</i>	<i>0.21</i>	<i>0.19</i>
Média X_1	<i>0.4501</i>	<i>0.7082</i>	<i>0.7557</i>	<i>0.2355</i>	<i>0.4811</i>
Média X_2	<i>0.3983</i>	<i>0.1062</i>	<i>0.1562</i>	<i>0.7552</i>	<i>0.8636</i>
Média X_3	<i>0.4121</i>	<i>0.3780</i>	<i>0.6606</i>	<i>0.2536</i>	<i>0.8326</i>
Variância	<i>0.0963</i>	<i>0.0932</i>	<i>0.0905</i>	<i>0.0929</i>	<i>0.1052</i>

A Figura 4.5 traz a representação desses dados e conforme pode ser observado, os grupos não estão totalmente separados uns dos outros, existindo alguns dados que se

encontram sobrepostos.

Figura 4.5 – Dados artificiais (Artificial 4) contendo 1000 dados, 5 grupos e 3 atributos, sendo que cada grupo contém 190, 207, 209 e 206 e 188 dados.

Base de Dados *wine*

A quinta e sexta bases de dados são a *wine* e *iris* obtidas do *UCI Machine Learning Repository* (BALKE; MERZ, 1998), sendo que as mesmas são bases de dados reais.

A base de dados *wine* compreende a análise química de três classes de vinhos italianos, de uma mesma região, mas derivados de três diferentes cultivos. Possui 178 dados com 13 atributos cada. Neste trabalho utilizou-se apenas os atributos 7, 10 e 13 por terem os grupos mais representativos (TIMM; DÖRING; KRUSE, 2004; BORGELT, 2006).

Na Figura 4.6 pode-se visualizar essa base de dados e como pode-se observar os grupos não se encontram perfeitamente separados, existindo dados sobrepostos e alguns ligeiramente espalhados.

Figura 4.6 – Dados *wine* (BALKE; MERZ, 1998) contendo 178 dados e 3 grupos.

Base de Dados *iris*

A base de dados *iris* é talvez a mais conhecida na literatura de reconhecimento de padrões. Os dados referem-se a largura da sépala, comprimento da sépala, largura da pétala e comprimento da pétala da planta *Iris*. Contém 150 dados com 4 atributos cada, igualmente distribuídos em três grupos (50 dados por grupo) denominados *Iris virgínica*, *Iris versicolor* e *Iris setosa*, das quais somente uma (*Iris setosa*) é linearmente separável das demais (BALKE; MERZ, 1998).

Na Figura 4.7 plota-se os dados da *iris* onde pode-se observar que dois grupos encontram-se sobrepostos e esses apresentam também dados ligeiramente espalhados.

Para a construção do gráfico utilizou-se três atributos que de acordo com a Figura 4.8 somam aproximadamente 100% da variância.

Figura 4.7 – Dados *iris* (BALKE; MERZ, 1998) contendo 150 dados e 3 grupos, dos quais somente um é linearmente separável.

Figura 4.8 – Análise de autovalores e variância da matriz de vetores resultantes de MDS (LLERENA, 2008) referentes à base de dados *iris* (BALKE; MERZ, 1998).

4.1.3 Base de Dados de Aplicação: *Bradyrhizobium*

Os rizóbios são conhecidos como bactérias simbióticas que, de modo geral, revelam sua capacidade de fixar N_2 somente quando em simbiose com plantas leguminosas e algumas

não-leguminosas. No entanto, os rizóbios são bactérias de vida livre e sobrevivem *in situ* às custas de compostos nitrogenados do solo, continuando a carregar o equipamento genético que lhes permite fixar o N_2 , até que encontrem as plantas hospedeiras específicas. Quando em simbiose vivem no interior de nódulos existentes nas raízes das plantas leguminosas, fornecendo parte do nitrogênio que absorvem da atmosfera às plantas e recebendo açúcares e outros compostos orgânicos. A fixação biológica do nitrogênio é um processo de tamanha importância que só pode ser comparado à fotossíntese. As bactérias fixadoras de nitrogênio formam o único grupo de seres vivos que suplementam os ecossistemas naturais com novas quantidades de nitrogênio, utilizando-se da inesgotável reserva natural que é a atmosfera (HUNGRIA; VARGAS; ARAÚJO, 1997).

Os rizóbios atualmente estão distribuídos em cinco famílias e seis gêneros, sendo que as bactérias do gênero *Bradyrhizobium* (*bradus*, grego, lento; logo bactérias de crescimento lento), espécies *Bradyrhizobium japonicum* e *Bradyrhizobium elkanii* nodulam a soja. Essas espécies são divididas em estirpes, sendo que uma das características que as diferenciam é a quantidade de nitrogênio fixado. Algumas estirpes fixam todo o nitrogênio necessário à produção normal da planta e outras não fixam praticamente nenhum nitrogênio (MILAGRE, 2003).

A diversidade genética das espécies rizobiais tropicais é ainda pouco conhecida (GERMANO *et al.*, 2006) pois, a taxonomia dos rizóbios de crescimento lento (gênero *Bradyrhizobium*) é, ainda, pouco refinada devido, em parte, aos resultados conflitantes obtidos pelas análises fenotípicas e genotípicas (SO; LADHA; YOUNG, 1994; ROSSUM *et al.*, 1995). Desta forma, novos métodos baseados em técnicas computacionais (MILAGRE *et al.*, 2008; TALAVERA, 2007; CHRIST, 2007; LLERENA, 2008) têm sido desenvolvidos com o objetivo de, inicialmente, auxiliar a pesquisa com a base de dados de *Bradyrhi-*

zobium, mas que também poderão ser aplicados posteriormente a outras bases de dados genotípicas.

Assim, o objetivo da inclusão desta base de dados neste trabalho é verificar se resultados já obtidos para a mesma se confirmam após a aplicação do método aqui proposto.

A base de dados *Bradyrhizobium* consiste em uma coleção brasileira de 119 estirpes de *Bradyrhizobium* que nodula raízes e caules, isoladas de 33 espécies de leguminosas tropicais, representando nove tribos e as três sub-famílias da família Leguminosae (MILAGRE, 2003; GERMANO *et al.*, 2006). Esse banco de dados foi obtido utilizando-se as regiões ribossomais 16S rRNA, 23S rRNA e o espaço intergênico 16S-23S (IGS) e nove enzimas de restrição, que *cortam* o DNA produzindo fragmentos de diferentes tamanhos tornando possível a comparação entre diferentes indivíduos, neste caso bactérias de solo. Os dados foram fornecidos pelo laboratório de Biotecnologia do Solo da Empresa Brasileira de Pesquisa Agropecuária - Centro Nacional de Pesquisa de Soja (Embrapa Soja), em Warta, distrito de Londrina, Paraná. Detalhes da obtenção dessa base de dados podem ser obtidos em (GERMANO *et al.*, 2006).

Os dados utilizados neste trabalho, referentes a essa base de dados, foram obtidos de (LLERENA, 2008), ou seja são resultantes da aplicação da técnica MDS sobre os dados de (MILAGRE, 2003). Serão utilizadas somente duas partes do banco de dados completo, referentes à região ribossomal 16S.

Esta é uma base de dados real e permitirá verificar o comportamento do método proposto em dados que não tenham uma distribuição normal (Artificial 3 e 4) ou um padrão rígido de distribuição (Artificial 1 e 2).

Base de Dados Brady 1

Corresponde à região ribossomal 16S rRNA digerida com as enzimas de restrição *Cfo* I, *Dde* I, *Msp* I. Portanto essa base de dados contém dados referentes às três enzimas. Contém 119 dados com 49 atributos cada.

Neste trabalho foram utilizados 20 atributos que de acordo com a Figura 4.9 (LLERENA, 2008) somam cerca de 94% da variância total.

Na Figura 4.10 apresenta-se os dados do Brady 1 (LLERENA, 2008). Utilizou-se três atributos para construção desse gráfico.

Figura 4.9 – Análise de autovalores e variância da matriz de vetores resultantes de MDS (LLERENA, 2008) referentes à base de dados Brady 1.

Base de Dados Brady 2

Corresponde à região ribossomal 16S rRNA digerida com a enzima de restrição *Cfo* I. Portanto essa base de dados contém dados referentes a apenas uma enzima. Contém 119 dados com 38 atributos cada.

Figura 4.10 – Base de dados Brady 1 (LLERENA, 2008) contendo 119 dados e considerando os três primeiros atributos.

Neste trabalho foram utilizados 15 atributos que de acordo com a Figura 4.11 (LLERENA, 2008) somam cerca de 98% da variância total.

Na Figura 4.12 apresenta-se os dados do Brady 2 (LLERENA, 2008). Utilizou-se três atributos para construção desse gráfico.

4.2 O Método

A aplicação do método proposto neste trabalho foi dividida em duas etapas, sendo que não houve nenhuma alteração de parâmetros em relação à nenhuma delas:

- **Etapa 1:** Simulações utilizando bancos de dados de validação: Artificial 1, Artificial 2, Artificial 3, Artificial 4, *iris* e *wine*;
- **Etapa 2:** Simulações utilizando bancos de dados de aplicação: Brady 1 e Brady 2.

Figura 4.11 – Análise de autovalores e variância da matriz de vetores resultantes de MDS (LLERENA, 2008) referentes à base de dados Brady 2.

Figura 4.12 – Dados Brady 2 (LLERENA, 2008) contendo 119 dados e considerando os três primeiros atributos.

O método tem início com a normalização da base de dados (dados originais) para média zero e desvio padrão unitário, para remover os efeitos de escala.

A seguir, modificou-se a base de dados normalizada para obter-se uma base de dados incompleta. Esta modificação foi feita escolhendo-se, aleatoriamente, alguns atributos para serem valores incompletos, ou seja, um atributo que não contém nenhum valor (exemplificado nas Tabelas 4.4 e 4.5). Assim, gerou-se dados incompletos substituindo-se o valor original para o atributo pelo valor *not-a-number* (**NaN**). As simulações foram feitas inserindo-se percentuais de 2%, 5%, 10%, 20% e 30% de valores faltando na base de dados.

A seleção aleatória de valores de atributos candidatos a serem valores faltando tiveram as seguintes restrições:

- Cada dado retém pelo menos um valor de atributo presente;
- Cada atributo deve ter pelo menos um valor presente.

A partir dos dados contendo valores faltando para atributos gera-se o vetor de índices incompletos $\mathbf{I}_j$, de acordo com o que foi visto na Seção 3.4 (Equação 3.9), onde $I_{jk} = 1$ indica que o k -ésimo atributo do dado $\mathbf{x}_j$ é observado (possui valor) e $I_{jk} = 0$ indica que o k -ésimo atributo do dado $\mathbf{x}_j$ não é observado (não possui valor).

Desta forma, de acordo com a Tabela 4.4, para os dados 001, 003 e 349 os vetores de índices incompletos são, respectivamente: $\mathbf{I}_{001} = [0 \ 1]$, $\mathbf{I}_{003} = [1 \ 1]$ e $\mathbf{I}_{349} = [1 \ 0]$.

A partir desse ponto as bases de dados utilizadas serão as bases de dados modificadas, ou seja, normalizadas e possuindo dados incompletos e serão tratadas como *referência*.

O próximo passo consiste em realizar o agrupamento dos dados, utilizando-se o algoritmo *fuzzy c-means*. Mas o algoritmo *fuzzy c-means* não pode ser diretamente aplicado em bases de dados que possuem dados incompletos. Assim, foi utilizado o algoritmo

Tabela 4.4 – Amostras dos 350 dados que constituem a base de dados Artificial 2, já normalizada e contendo dados incompletos (30% de valores faltando). Onde o **NaN** indica que o atributo não possui valor observado.

<i>Dado</i>	<i>Atributo 1</i>	<i>Atributo 2</i>
001	NaN	-1.2770
002	-1.5013	-1.2185
003	-1.4113	-1.3818
⋮	⋮	⋮
100	0.4090	-0.3345
101	NaN	0.2793
⋮	⋮	⋮
348	NaN	-1.3060
349	1.4775	NaN
350	1.6682	NaN

Tabela 4.5 – Amostras dos 178 dados da *wine*(BALKE; MERZ, 1998), já normalizada e contendo dados incompletos (20% de valores faltando). Onde o **NaN** indica que o atributo não possui valor observado.

<i>Dado</i>	<i>Atributo 1</i>	<i>Atributo 2</i>	<i>Atributo 3</i>
001	1.0009	0.3524	1.0685
002	0.7129	NaN	1.0190
003	1.1737	0.3690	1.4644
⋮	⋮	⋮	⋮
104	NaN	-1.1381	NaN
105	NaN	NaN	-0.2281
⋮	⋮	⋮	⋮
176	-1.2741	2.2509	0.3097
177	-1.2837	1.8762	0.3262
178	NaN	1.8346	NaN

fuzzy c-means modificado (Seção 3.4), através das Equações 3.12 e 3.13, para realizar o agrupamento dos dados.

Os valores utilizados para a execução do algoritmo nebuloso *c-means* foram:

- número máximo de iterações: $t = 100$;
- fator de nebulosidade: $m = 2$;
- critério de parada: $\varepsilon \leq 0.00001$;
- número de grupos: $c = 2, 3, \dots, 8$;
- matriz de partições nebulosas inicial com c linhas (número de grupos) e n colunas (número de dados): $\mathbf{U}^{(0)}$ é obtida através da função *rand* do MATLAB[®] (MATHWORKS, 1998) que gera números aleatórios. Após ser gerada, a matriz $\mathbf{U}^{(0)}$ é normalizada de forma que a restrição dada pela Equação 3.3 seja satisfeita, ou seja,

$$\sum_{i=1}^c \mu_i(\mathbf{x}_j) = 1, \forall j \in \{1, 2, \dots, n\}$$

Esta normalização é feita toda vez que uma nova matriz de partição $\mathbf{U}$ é gerada.

Inicialmente, aplicou-se o algoritmo *fuzzy c-means* na referência, gerando-se a primeira matriz de comparação $\mathbf{U}^{(1)}$ ($\mathbf{U} = (\mu_{ij})$). Particionou-se cada referência em número de grupos variando de 2 a 8 ($c = 2, 3, \dots, 8$), para permitir uma melhor análise em torno do número de grupos existentes em cada base de dados utilizada (Tabela 4.1) que varia de 3 a 7.

A seguir, utilizando-se o método de reamostragem com reposição (*bootstrapping*) foi gerada a sub-amostra de dados contendo 90% dos dados da referência.

O vetor aleatório da sub-amostra foi gerado utilizando-se a função, do MATLAB® (MATHWORKS, 1998), *bootstrp*. Essa função gera, a partir de um número inteiro (número de dados n em cada referência) um vetor de permutações aleatórias, com reposição, de 1 até n . Do vetor aleatório gerado, contendo n números, foram selecionados os primeiros 90%, que passaram a compor o vetor da sub-amostra. A sub-amostra é então composta buscando-se o dado correspondente a cada índice, na referência, que contém todos os dados.

Criou-se, então, para a sub-amostra o vetor de índices incompletos $\mathbf{I}_j$, da mesma forma que foi criado para a referência.

A sub-amostra foi agrupada, utilizando-se também o algoritmo *fuzzy c-means* modificado (Seção 3.4) e o mesmo número de agrupamentos a que foi submetida a referência.

Neste ponto tem-se então:

- o agrupamento gerado para a referência e suas respectivas matriz de pertinência $\mathbf{U}^{(1)}$ e centros $\mathbf{v}^{(1)}$;
- o agrupamento gerado para a sub-amostra e suas respectivas matriz de pertinência $\mathbf{U}^{(2)}$ centros $\mathbf{v}^{(2)}$.

As matrizes $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$ são matrizes de dimensões $(c \times n)$ e $(c \times 0.9n)$, respectivamente. Desta forma, não podem ser comparadas diretamente. Assim, os centros gerados ($\mathbf{v}^{(2)}$) pelo algoritmo *fuzzy c-means* para a sub-amostra (Equação 3.12), são reaplicados na referência, ou seja, utiliza-se os centros da sub-amostra e a matriz de dados para a referência e recalcula-se a distância entre os dados da referência e o modelo de centros que

foram gerados para a sub-amostra. Dessa forma, obtém-se nova matriz de pertinência, a qual passa então, a ser a segunda matriz de comparação $\mathbf{U}^{(2)}$ ($c \times n$). Este processo é repetido para cada sub-amostra (são utilizadas 100 sub-amostras), ou seja, cada sub-amostra dará origem a uma matriz de pertinência $\mathbf{U}^{(2)}$ ($c \times n$).

De posse das matrizes de partição $\mathbf{U}^{(1)}$ (referência) e $\mathbf{U}^{(2)}$ (uma para cada sub-amostra), calcula-se, para cada uma das 100 sub-amostras, as medidas de comparação binária, descritas na Seção 3.5, que são: Diferença Média Quadrática (Diff.), F_1 , Classificação Cruzada (Acc.), Hubert, Jaccard, Índice Randômico (Rand.), Fowlkes & Mallows (F&M) e Índice Randômico Ajustado (Rand. Aj.). O valor final para cada medida é obtido aplicando-se a média entre os resultados das 100 sub-amostras.

O valor máximo obtido para cada medida, entre todos os particionamentos ($c = 2, 3, \dots, 8$), indica o número de grupos definido pela medida como sendo o mais correto para a base de dados em análise. Cabe ressaltar que no caso da medida Diff., que utiliza o valor mínimo, o valor obtido foi subtraído de 1. Portanto também utiliza-se o valor máximo para essa medida.

Para o cálculo das medidas de comparação entre matrizes de partição (Seção 3.5.1), aplicou-se a *t-norma* (Tabela 3.2) produto (Equação 4.7) e para comparação entre matrizes de coincidência (Seção 3.5.2) utilizou-se a combinação das *t-normas nil-potent* mínimo (Equação 4.8) e produto (Equação 4.7).

$$\top_{prod}(\mu_{ij}^{(1)}, \mu_{rj}^{(2)}) = \mu_{ij}^{(1)} \cdot \mu_{rj}^{(2)} \quad (4.7)$$

$$\tau_{npmin}(\mu_{ij}^{(1)}, \mu_{rj}^{(2)}) = \begin{cases} \min(\mu_{ij}^{(1)}, \mu_{rj}^{(2)}), & \text{se } (\mu_{ij}^{(1)} + \mu_{rj}^{(2)}) \geq 1 \\ 0, & \text{senão} \end{cases} \quad (4.8)$$

No Algoritmo 1 mostra-se todos os passos utilizados neste trabalho para o cálculo do número de grupos em bases de dados com dados incompletos.

4.2.1 Comparando Matrizes de Partição

Descreve-se aqui um exemplo para melhor entendimento dos cálculos. Será utilizada uma matriz de pertinência *crisp* (pertinências 0 ou 1) para tornar o exemplo mais simples.

Seja uma base de dados contendo 10 dados ($n = 10$), possuindo 6 atributos cada e um agrupamento em 3 grupos ($c = 3$).

Sejam também duas matrizes de pertinência ($c \times n$), referentes à referência, $\mathbf{U}^{(1)}$ e sub-amostra, $\mathbf{U}^{(2)}$, obtidas do algoritmo *fuzzy c-means* modificado, dadas por:

$$\mathbf{U}^{(1)} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

$$\mathbf{U}^{(2)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Nas matrizes $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$ as linhas representam os grupos ($c = 3$) e cada coluna

Algoritmo 1 - Algoritmo para cálculo do número de grupos em bases de dados incompletas, utilizando *bootstrapping* e algoritmo *fuzzy c-means* modificado.

```

1: Input: Base de dados,  $t = 100$ ,  $cmin = 2$ ,  $cmax = 8$ ,  $m = 2$ ,  $\varepsilon \leq 0.00001$ ,  $num\_sub\_amostras = 100$ ,  $perc\_valores\_faltando = (2\%, 5\%, 10\%, 20\%, 30\%)$ 
2: Output: Medidas de similaridade: média das similaridades (calculadas de acordo com as medidas descritas na Seção 3.5) entre a referência e as 100 sub-amostras.
3: Ler a base de dados (referência)
4: Obter o número de dados ( $n$ ) que a referência possui
5: Gerar  $num\_dados\_sub\_amostras = 0.90 * n$ 
6: Obter o número de atributos ( $n\_atrib$ ) que a referência possui
7: Gerar  $n\_dado\_atrib = n\_atrib * n$ 
8: for  $f = 1$  to 5 do
9:   Obter percentual de valores faltando:  $num\_perc = perc\_valores\_faltando[f]$ 
10:  Gerar  $num\_valores\_faltando = num\_perc * n\_dado\_atrib$ 
11:  for  $c = cmin$  to  $cmax$  do
12:    Inserir o percentual de valores faltando na base de dados (referência)
13:    Gerar o vetor de índices incompletos (I) para a referência
14:    Normalizar a base de dados (referência) para média zero e desvio padrão unitário
15:    Aplicar algoritmo fuzzy c-means na referência (obtendo-se  $\mathbf{U}^{(1)}$  e  $\mathbf{v}^{(1)}$ )
16:    for  $d = 1$  to  $num\_sub\_amostras$  do
17:      Gerar vetor_indice_bootstrap (função bootstrap (MATHWORKS, 1998)), contendo  $n$  itens
18:      for  $ind\_nsub = 1$  to  $num\_dados\_sub\_amostras$  do
19:        Gerar  $vetor\_ind\_sub\_amostras = vetor\_indice\_bootstrap[ind\_nsub]$ 
20:      end for
21:      for  $ind\_nsub = 1$  to  $num\_dados\_sub\_amostras$  do
22:        Gerar  $ind\_dado\_ref = vetor\_ind\_sub\_amostras[ind\_nsub]$ 
23:        Montar  $dados\_sub\_amostras = referencia[ind\_dado\_ref]$ 
24:      end for
25:      Gerar o vetor de índices incompletos (I) para a sub-amostra  $d$ 
26:      Aplicar algoritmo fuzzy c-means na sub-amostra  $d$  (obtendo-se  $\mathbf{U}^{(2)}$  e  $\mathbf{v}^{(2)}$ )
27:      Recalcular distância entre referência e os centros da sub-amostra  $d$  ( $\mathbf{v}^{(2)}$ )
28:      Gerar nova matriz  $\mathbf{U}^{(2)}$  ( $c \times n$ )
29:      Calcular a similaridade (Equação 3.27, 3.28 e 3.29) entre o agrupamento de referência ( $\mathbf{U}^{(1)}$ ) e o agrupamento da sub-amostra  $d$  ( $\mathbf{U}^{(2)}$ )
30:      Calcular a similaridade (Equação 3.36, 3.37, 3.39, 3.40 e 3.41) entre o agrupamento de referência ( $\mathbf{U}^{(1)}$ ) e o agrupamento da sub-amostra  $d$  ( $\mathbf{U}^{(2)}$ )
31:    end for
32:    Calcular a média entre as medidas de similaridade obtidas para as sub-amostras ( $num\_sub\_amostras$ )
33:  end for
34: end for

```

($n = 10$) representa a pertinência de cada dado dentro do grupo. Neste caso, *crisp*, se o dado pertence (1) ou não pertence (0) ao grupo.

A comparação entre matrizes de partição irá indicar o maior índice de similaridade entre os grupos da matriz de pertinência original (referência) e da sub-amostra, ou seja, *quantos* dados pertencem a um mesmo grupo, ou não.

- **Calculando erro médio quadrático (Diff.):**

Inicialmente calcula-se uma matriz de comparação entre cada linha das matrizes $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$. Desta forma, compara-se, por exemplo, cada elemento da primeira linha da matriz referência ($\mu_{1j}^{(1)}$) com cada elemento da primeira linha da matriz da sub-amostra ($\mu_{1j}^{(2)}$). A seguir repete-se o processo entre a primeira linha de $\mathbf{U}^{(1)}$ com a segunda linha de $\mathbf{U}^{(2)}$ e assim por diante até que sejam feitas todas as combinações de linhas entre as duas matrizes.

Obtém-se a matriz de comparação ($\mathbf{C}^{(1,2)}$) que é uma matriz $c \times c$, onde cada elemento representa o resultado da comparação entre as linhas, assim:

$$\mathbf{C}^{(1,2)} = \begin{pmatrix} (\mathbf{U}_1^{(1)} \mathbf{U}_1^{(2)}) & (\mathbf{U}_1^{(1)} \mathbf{U}_2^{(2)}) & (\mathbf{U}_1^{(1)} \mathbf{U}_3^{(2)}) \\ (\mathbf{U}_2^{(1)} \mathbf{U}_1^{(2)}) & (\mathbf{U}_2^{(1)} \mathbf{U}_2^{(2)}) & (\mathbf{U}_2^{(1)} \mathbf{U}_3^{(2)}) \\ (\mathbf{U}_3^{(1)} \mathbf{U}_1^{(2)}) & (\mathbf{U}_3^{(1)} \mathbf{U}_2^{(2)}) & (\mathbf{U}_3^{(1)} \mathbf{U}_3^{(2)}) \end{pmatrix}$$

que, para o exemplo, resulta em:

$$\mathbf{C}^{(1,2)} = \begin{pmatrix} \mathbf{3} & 6 & 3 \\ 6 & 1 & 8 \\ 2 & 9 & 2 \end{pmatrix}$$

Tomando-se como exemplo o elemento C_{11} igual a 3, resultado da comparação entre as linhas $(\mathbf{U}_1^{(1)})$ e $(\mathbf{U}_1^{(2)})$:

$$\mathbf{U}_1^{(1)} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{U}_1^{(2)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

De acordo com a Equação 3.29:

$$Diff.(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \min_{\zeta \in \Pi(c)} \frac{1}{cn} \sum_{i=1}^c \sum_{j=1}^n \left(\mu_{ij}^{(1)} - \mu_{\zeta(i)j}^{(2)} \right)^2$$

em relação ao primeiro somatório ($j = 1$ até n), faz-se o cálculo da diferença entre o valor dos elementos de $\mathbf{U}_1^{(1)}$ e seu correspondente em $\mathbf{U}_1^{(2)}$, eleva-se o mesmo ao quadrado e acumula-se o resultado:

$$\mathbf{C}_{11}^{(1,2)} = (1 - 0)^2 + (0 - 0)^2 + (0 - 0)^2 + \dots + (0 - 1)^2 + \dots + (0 - 0)^2 = 3$$

Obtém-se então o mínimo entre as comparações feitas entre as linhas das matrizes $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$, que corresponde a obter-se o menor valor em cada linha da matriz de comparação e efetuar a soma dos mesmos, que para o exemplo dado, é igual a 6 (soma dos valores 3, 1 e 2). Divide-se por $c * n$, obtendo-se o resultado

final para a medida Diff. igual a 0.2.

Para o caso nebuloso procede-se da mesma forma. A matriz $\mathbf{C}_{Diff}^{(1,2)}$ traz um resultado de comparação obtido para a base de dados *wine* (BALKE; MERZ, 1998) (onde $n = 178$) e uma sub-amostra, para um agrupamento $c = 3$. Neste caso a medida Diff. é igual a 0.0775. Os valores em negrito representam o valor mínimo referente a cada comparação.

$$\mathbf{C}_{Diff}^{(1,2)} = \begin{pmatrix} 56.9993 & \mathbf{12.1360} & 49.7055 \\ 49.2259 & 50.4032 & \mathbf{16.1202} \\ \mathbf{13.1122} & 46.0784 & 43.3801 \end{pmatrix}$$

- **Calculando classificação cruzada (Acc.):**

De acordo com a Equação 3.28:

$$Acc.(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \max_{\zeta \in \Pi(c)} \frac{1}{cn} \sum_{i=1}^c (n_{00}^{(i, \zeta(i))} + n_{11}^{(i, \zeta(i))})$$

calcula-se primeiramente n_{00} (Equação 3.18):

$$n_{00}^{(i,r)}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \sum_{j=1}^n (1 - \mu_{ij}^{(1)}) \cdot (1 - \mu_{rj}^{(2)})$$

calculando n_{00} entre $\mathbf{U}_1^{(1)}$ e $\mathbf{U}_1^{(2)}$:

$$n_{00}^{(1,1)}(\mathbf{U}_1^{(1)}, \mathbf{U}_1^{(2)}) = \sum_{j=1}^{10} (1 - \mu_{1j}^{(1)}) \cdot (1 - \mu_{1j}^{(2)})$$

utilizando-se a *t-norma* escolhida para este processo que foi a *t-norma* produto

(Equação 4.7):

$$\top_{prod}(\mu_{ij}^{(1)}, \mu_{rj}^{(2)}) = \mu_{ij}^{(1)} \cdot \mu_{rj}^{(2)}$$

$$n_{00}^{(1,1)} = (1-1).(1-0) + (1-0).(1-0) + \dots + (1-0).(1-0) = 7$$

calculando n_{11} (Equação 3.15):

$$n_{11}^{(i,r)}(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \sum_{j=1}^n \mu_{ij}^{(1)} \cdot \mu_{rj}^{(2)}$$

efetuando-se o cálculo de n_{11} entre $\mathbf{U}_1^{(1)}$ e $\mathbf{U}_1^{(2)}$:

$$n_{11}^{(1,1)}(\mathbf{U}_1^{(1)}, \mathbf{U}_1^{(2)}) = \sum_{j=1}^{10} \mu_{1j}^{(1)} \cdot \mu_{1j}^{(2)}$$

utilizando-se também a *t-norma* produto:

$$n_{11}^{(1),(1)} = (1).(0) + (0).(0) + (0).(0) + \dots + (0).(0) = 0$$

desta forma, $n_{00}^{(1,1)} + n_{11}^{(1,1)} = 7$, conforme mostra a matriz de comparação $\mathbf{C}^{(1,2)}$.

Cabe ressaltar que o resultado $n_{00} = 7$ significa que existem 7 dados que não se encontram nem no grupo $c = 1$ de $\mathbf{U}^{(1)}$ e nem no grupo $c = 1$ de $\mathbf{U}^{(2)}$ e o resultado $n_{11} = 0$ significa que não existe nenhum dado que esteja no grupo $c = 1$ de $\mathbf{U}^{(1)}$ e também no grupo $c = 1$ de $\mathbf{U}^{(2)}$.

Os demais valores são calculados para todas as combinações de linhas entre as matrizes $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$, gerando-se assim a matriz de comparação $\mathbf{C}^{(1,2)}$, que para o exemplo

fica da seguinte forma:

$$\mathbf{C}^{(1,2)} = \begin{pmatrix} 7 & 4 & 7 \\ 4 & 9 & 2 \\ 8 & 1 & 8 \end{pmatrix}$$

O próximo passo no cálculo de Acc. é obter o valor máximo entre a soma de todas as possíveis permutações de elementos da matriz de comparação $\mathbf{C}^{(1,2)}$ que corresponde a obter-se o maior valor em cada linha da matriz de comparação e efetuar a soma dos mesmos, que de acordo com exemplo é igual a 24 (soma dos valores 7, 9 e 8). Divide-se por $c * n$, obtendo-se o resultado final para a medida Acc. igual a 0.8.

A matriz $\mathbf{C}_{Acc.}^{(1,2)}$ é resultante da comparação entre a base de dados *wine* (BALKE; MERZ, 1998) (onde $n = 178$) e uma sub-amostra, para um agrupamento $c = 5$. Os valores em negrito representam o valor máximo em cada linha. Neste caso a medida Acc. é igual a 0.7337.

$$\mathbf{C}_{Acc.}^{(1,2)} = \begin{pmatrix} 121.7665 & 110.4763 & 109.9923 & \mathbf{125.5037} & 124.0678 \\ \mathbf{133.0242} & 131.9842 & 116.9316 & 108.0532 & 120.3607 \\ 113.2147 & 103.0739 & \mathbf{125.9744} & 113.8367 & 107.7775 \\ 127.1391 & 125.6279 & 117.0424 & 117.9521 & \mathbf{131.4628} \\ 135.6394 & \mathbf{137.4941} & 123.7853 & 117.6152 & 126.2039 \end{pmatrix}$$

- **Calculando F_1 :**

A medida de comparação F_1 é obtida através da Equação 3.27:

$$F_1(\mathbf{U}^{(1)}, \mathbf{U}^{(2)}) = \max_{\zeta \in \Pi(c)} \frac{1}{c} \sum_{i=1}^c \frac{2\pi_{i,\zeta(i)}\rho_{i,\zeta(i)}}{\pi_{i,\zeta(i)} + \rho_{i,\zeta(i)}}$$

Como os processos são similares ao do cálculo de Acc., os passos aqui serão descritos de forma sucinta. Assim, inicialmente calcula-se n_{11} , n_{01} , n_{10} e n_{00} (Equações 3.15 a 3.18), para cada uma das possíveis combinações entre as linhas de $\mathbf{U}^{(1)}$ e $\mathbf{U}^{(2)}$, de acordo com a *t-norma* escolhida para este processo que foi a *t-norma* produto.

A seguir são calculadas a precisão π (Equação 3.25) e recuperação ρ (Equação 3.26) substituindo-se, então na Equação 3.27, obtém-se, para o exemplo dado, a matriz de comparação $\mathbf{C}^{(1,2)}$:

$$\mathbf{C}^{(1,2)} = \begin{pmatrix} 0 & 0.25 & 0.4 \\ 0 & 0.9091 & 0 \\ 0.5000 & 0 & 0.6667 \end{pmatrix}$$

Calcula-se o máximo entre a soma de todas as permutações de elementos da matriz de comparação $\mathbf{C}^{(1,2)}$, que é igual a 1.9758 (soma dos valores 0.4, 0.9091 e 0.6667).

Divide-se por c , obtendo-se o resultado final para F_1 igual a 0.6586.

A matriz $\mathbf{C}_{F_1}^{(1,2)}$ é resultante da comparação entre a base de dados *wine* (BALKE; MERZ, 1998) (onde $n = 178$) e uma sub-amostra, para um agrupamento $c = 7$. Os valores em negrito representam o valor máximo em cada linha. Neste caso a medida F_1 é igual a 0.2886.

$$\mathbf{C}_{F_1}^{(1,2)} = \begin{pmatrix} 0.1483 & \mathbf{0.2835} & 0.1612 & 0.1330 & 0.1522 & 0.0611 & 0.0527 \\ 0.2842 & 0.1296 & 0.0828 & \mathbf{0.3438} & 0.0826 & 0.0946 & 0.0856 \\ 0.1140 & 0.1674 & \mathbf{0.3056} & 0.0826 & 0.1833 & 0.1246 & 0.0766 \\ 0.0746 & 0.0672 & 0.1306 & 0.1232 & 0.1275 & \mathbf{0.2670} & 0.2176 \\ \mathbf{0.2642} & 0.2094 & 0.1236 & 0.1643 & 0.1098 & 0.0709 & 0.0666 \\ 0.0660 & 0.0662 & 0.0760 & 0.0944 & 0.0743 & 0.1761 & \mathbf{0.3767} \\ 0.0879 & 0.1124 & 0.1584 & 0.0775 & \mathbf{0.1791} & 0.0884 & 0.1177 \end{pmatrix}$$

4.2.2 Comparando Matrizes de Coincidência

O objetivo da comparação entre matrizes de coincidência é verificar *quais* dados pertencem a um mesmo grupo, ou não.

Para comparar matrizes de coincidência será utilizado um exemplo com matrizes de pertinências nebulosas, devido ao fato de que a *t-norma nil-potent* mínimo (Equação 4.8), utilizada na primeira parte deste cálculo, não pode ser aplicada ao exemplo utilizando particionamento *crisp* (o resultado é sempre zero). Lembrando que no particionamento *crisp* o *and* é implementado pelo produto algébrico comum.

Seja, então, uma base de dados contendo 10 dados ($n = 10$) e um agrupamento em 3 grupos ($c = 3$) e as duas matrizes de pertinência ($c \times n$), referentes à referência, $\mathbf{U}^{(1)}$ e sub-amostra, $\mathbf{U}^{(2)}$, obtidas do algoritmo *fuzzy c-means* modificado, dadas por:

$$\mathbf{U}^{(1)} = \begin{pmatrix} 0.7 & 0.3 & 0.1 & 0.9 & 0.3 & 0.1 & 0.2 & 0.1 & 0.2 & 0.1 \\ 0.2 & 0.5 & 0.8 & 0.1 & 0.2 & 0.7 & 0.2 & 0.3 & \mathbf{0.7} & \mathbf{0.8} \\ 0.1 & 0.2 & 0.1 & 0.0 & 0.5 & 0.2 & 0.6 & 0.6 & 0.1 & 0.1 \end{pmatrix}$$

$$\mathbf{U}^{(2)} = \begin{pmatrix} 0.2 & 0.2 & 0.1 & 0.2 & 0.3 & 0.2 & 0.4 & 0.1 & 0.3 & 0.1 \\ 0.7 & 0.5 & 0.8 & 0.2 & 0.2 & 0.5 & 0.3 & 0.3 & 0.6 & 0.8 \\ 0.1 & 0.3 & 0.1 & 0.6 & 0.5 & 0.3 & 0.3 & 0.6 & 0.1 & 0.1 \end{pmatrix}$$

Iniciam-se os cálculos comparando-se cada elemento da coluna j ($j = 10, 9, \dots, 1$) com todos os elementos da coluna l ($l = 10, 9, \dots, 1$), para cada linha que representa o grupo (para o exemplo dado $i = 1, 2, 3$), através da Equação 3.30:

$$\psi_{jl} = \sum_{i=1}^c \mu_{ij} \cdot \mu_{il}$$

Para este processo escolheu-se a *t-norma nil-potent* mínimo dada pela Equação 4.8:

$$\top_{npmin}(\mu_{ij}^{(1)}, \mu_{rj}^{(2)}) = \begin{cases} \min(\mu_{ij}^{(1)}, \mu_{rj}^{(2)}), & \text{se } (\mu_{ij}^{(1)} + \mu_{rj}^{(2)}) \geq 1 \\ 0, & \text{senão} \end{cases}$$

Assim, para $j = 10$ e $l = 9$, tem-se:

- em $\mathbf{U}_1^{(1)}$:

$$\mu_{(1,10)} = 0.1 \text{ e } \mu_{(1,9)} = 0.2$$

então,

$$\mu_{(1,10)} = 0.1 + \mu_{(1,9)} = 0.2 = 0.3 < 1$$

logo,

$$\top_{npmin}(\mu_{(1,10)}, \mu_{(1,9)}) = 0$$

- em $\mathbf{U}_2^{(1)}$:

$$\mu_{(2,10)} = 0.8 \text{ e } \mu_{(2,9)} = 0.7$$

então,

$$\mu_{(2,10)} = 0.8 + \mu_{(2,9)} = 0.7 = 1.5 \geq 1$$

logo,

$$\top_{npmin}(\mu_{(2,10)}, \mu_{(2,9)}) = \min(0.7, 0.8) = 0.7$$

- em $\mathbf{U}_3^{(1)}$:

$$\mu_{(3,10)} = 0.1 \text{ e } \mu_{(3,9)} = 0.1$$

então,

$$\mu_{(3,10)} = 0.1 + \mu_{(3,9)} = 0.1 = 0.2 < 1$$

logo,

$$\top_{npmin}(\mu_{(3,10)}, \mu_{(3,9)}) = 0$$

Da Equação 3.30 obtém-se, então o resultado para $\psi_{(10,9)} = 0 + 0.7 + 0 = 0.7$. Este resultado significa que os dados $j = 9$ e $j = 10$ possuem sua maior pertinência no mesmo grupo (no caso $c = 2$), como pode-se verificar pelo valores em negrito na matriz $\mathbf{U}^{(1)}$.

Repetem-se os cálculos comparando-se $j = 10$ com $l = 8$, depois $j = 10$ com $l = 7$ e assim por diante, até que todos os elementos sejam comparados entre si.

O resultado final, deste exemplo, é a matriz de coincidência $\Psi^{(1)}$. Os valores em negrito representam o resultado da comparação para $j = l$.

Para o cálculo da segunda matriz de comparação $\Psi^{(2)}$, repete-se o processo utilizando-se a sub-amostra.

Geradas as matrizes $\Psi^{(1)}$ e $\Psi^{(2)}$, calcula-se os números N_{11} , N_{01} , N_{10} e N_{00} (Equações 3.32 a 3.35), sendo que, nesta etapa, utiliza-se a *t-norma* produto.

$$\Psi^{(1)} = \begin{pmatrix} \mathbf{0.7} & 0.3 & 0.2 & 0.7 & 0.3 & 0 & 0 & 0 & 0 & 0.2 \\ 0.3 & \mathbf{0.5} & 0.5 & 0.3 & 0 & 0.5 & 0 & 0 & 0.5 & 0.5 \\ 0.2 & 0.5 & \mathbf{0.8} & 0.1 & 0.2 & 0.7 & 0.2 & 0.3 & 0.7 & 0.8 \\ 0.7 & 0.3 & 0.1 & \mathbf{0.9} & 0.3 & 0.1 & 0.2 & 0.1 & 0.2 & 0.1 \\ 0.3 & 0 & 0.2 & 0.3 & \mathbf{0.5} & 0 & 0.5 & 0.5 & 0 & 0.2 \\ 0 & 0.5 & 0.7 & 0.1 & 0 & \mathbf{0.7} & 0 & 0.3 & 0.7 & 0.7 \\ 0 & 0 & 0.2 & 0.2 & 0.5 & 0 & \mathbf{0.6} & 0.6 & 0 & 0.2 \\ 0 & 0 & 0.3 & 0.1 & 0.5 & 0.3 & 0.6 & \mathbf{0.6} & 0.3 & 0.3 \\ 0 & 0.5 & 0.7 & 0.2 & 0 & 0.7 & 0 & 0.3 & \mathbf{0.7} & 0.7 \\ 0.2 & 0.5 & 0.8 & 0.1 & 0.2 & 0.7 & 0.2 & 0.3 & 0.7 & \mathbf{0.8} \end{pmatrix}$$

$$\Psi^{(2)} = \begin{pmatrix} \mathbf{0.7} & 0.5 & 0.7 & 0 & 0 & 0.5 & 0.3 & 0.3 & 0.6 & 0.7 \\ 0.5 & \mathbf{0.5} & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0.5 & 0.5 \\ 0.7 & 0.5 & \mathbf{0.8} & 0.2 & 0.2 & 0.5 & 0.3 & 0.3 & 0.6 & 0.8 \\ 0 & 0 & 0.2 & \mathbf{0.6} & 0.5 & 0 & 0 & 0.6 & 0 & 0.2 \\ 0 & 0 & 0.2 & 0.5 & \mathbf{0.5} & 0 & 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0 & 0 & \mathbf{0.5} & 0 & 0 & 0.5 & 0.5 \\ 0.3 & 0 & 0.3 & 0 & 0 & 0 & \mathbf{0} & 0 & 0 & 0.3 \\ 0.3 & 0 & 0.3 & 0.6 & 0.5 & 0 & 0 & \mathbf{0.6} & 0.3 & 0.3 \\ 0.6 & 0.5 & 0.6 & 0 & 0 & 0.5 & 0 & 0 & \mathbf{0.6} & 0.6 \\ 0.7 & 0.5 & 0.8 & 0.2 & 0.2 & 0.5 & 0.3 & 0.3 & 0.6 & \mathbf{0.8} \end{pmatrix}$$

Exemplificando o cálculo de N_{11} (Equação 3.32):

$$N_{11}(\Psi^{(1)}, \Psi^{(2)}) = \sum_{j=2}^n \sum_{l=1}^{j-1} \psi_{jl}^{(1)} \cdot \psi_{jl}^{(2)}$$

$$\begin{aligned} N_{11} = & (0.15) + (0.14 + 0.25) + (0 + 0 + 0.02) + (0 + 0 + 0.04 + 0.15) + (0 + 0.25 + 0.35 + \\ & 0 + 0) + (0 + 0 + 0.06 + 0 + 0 + 0) + (0 + 0 + 0.09 + 0.06 + 0.25 + 0 + 0) + (0 + 0.25 + \\ & 0.42 + 0 + 0 + 0.35 + 0 + 0) + (0.14 + 0.25 + 0.64 + 0.02 + 0.04 + 0.35 + 0.06 + 0.09 + 0.42) \end{aligned}$$

$$N_{11} = (0.15) + (0.39) + (0.02) + (0.19) + (0.6) + (0.06) + (0.40) + (1.02) + (2.01)$$

$$N_{11} = 4.84$$

Prosseguindo-se com os cálculos tem-se:

$$N_{10} = 8.16$$

$$N_{01} = 7.86$$

$$N_{00} = 24.14$$

A seguir, obtém-se as medidas Jaccard, Índice Randômico, Fowlkes & Mallows, Hubert, e Índice Randômico Ajustado (Equações 3.36 a 3.41):

$$Jaccard = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} = 0.2320$$

$$Rand. = \frac{N_{11} + N_{00}}{N_{11} + N_{01} + N_{10} + N_{00}} = 0.644$$

$$Fowlkes-Mallows = \frac{N_{11}}{\sqrt{(N_{11} + N_{10})(N_{11} + N_{01})}} = 0.3767$$

$$Hubert = \frac{N_{00}N_{11} - N_{10}N_{01}}{\sqrt{(N_{11} + N_{10})(N_{11} + N_{01})(N_{01} + N_{00})(N_{10} + N_{00})}} = 0.1276$$

$$Rand. Aj. = \frac{2(N_{00}N_{11} - N_{01}N_{10})}{2(N_{11}N_{00}) + (N_{11}N_{00})(N_{01} + N_{10}) + N_{10}^2 + N_{01}^2} = 0.0470$$

O processo de cálculo das medidas é repetido para cada par referência/sub-amostra. Ao final faz-se a média entre os resultados obtidos para os 100 pares referência/sub-amostra.

Repete-se o processo completo para cada particionamento ($c = 2, 3, \dots, 8$).

Após efetuar-se os cálculos para cada particionamento, repete-se novamente todo o processo para cada percentual de valores faltando (2%, 5%, 10%, 20% e 30%).

Capítulo 5

Resultados e Discussão

Pesquisadores freqüentemente se vêem *face a face* com o problema de bases de dados incompletas, quer seja na medicina, biologia ou uma pesquisa de opinião. Vários são os métodos que se propõem a resolver o problema, ou pelo menos diminuir o impacto causado pelos mesmos no resultado final da análise. Métodos tradicionais como exclusão dos dados incompletos ou de inserção de valores podem afetar o tamanho da amostra ou levar a resultados tendenciosos. Assim, na busca de novas alternativas, as atenções têm se voltado também para métodos que suportem a presença dos dados incompletos.

Neste sentido, o objetivo deste trabalho foi identificar os grupos presentes em bases de dados incompletas aplicando medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, nos resultados obtidos pelo algoritmo de agrupamento nebuloso, através da análise entre referência e sub-amostras geradas pelo mecanismo de reamostragem com reposição (*bootstrapping*).

A validação do método baseou-se em seis conjuntos de dados com número de grupos conhecidos, sendo quatro criados artificialmente, dois (*wine* e *iris* (BALKE; MERZ, 1998)) que são bastante utilizadas pela comunidade de aprendizado de máquina para vali-

dação empírica de algoritmos e duas bases são formadas por uma coleção de bactérias de solo fixadoras de nitrogênio (MILAGRE *et al.*, 2008; MILAGRE; MACIEL, 2008b; GERMANO *et al.*, 2006), cujo número de grupos ainda encontra-se em estudo (MILAGRE, 2003; GERMANO *et al.*, 2006; TALAVERA, 2007; CHRIST, 2007; LLERENA, 2008).

Mostra-se aqui alguns resultados das variabilidades das partições para as bases de dados utilizadas, sendo que, nem todos foram mostrados por serem em grande número e na maioria das vezes repetitivos, apresentando apenas pequenas variações entre os percentuais de valores faltando inseridos (2%, 5%, 10%, 20% e 30%). Os valores em negrito nas tabelas representam o valor máximo obtido em cada medida de comparação.

Este capítulo está dividido em quatro seções. Na primeira seção mostra-se os tempos de processamento para as simulações executadas. A segunda parte traz os resultados obtidos e as discussões para as bases de dados de validação. Na terceira seção mostra-se os resultados e discussões para as bases de dados de aplicação. A quarta seção traz um resumo dos resultados obtidos e discussões sobre os mesmos.

5.1 Tempos de Processamento

Uma análise inicial a ser feita é com relação aos tempos de processamento que se encontram relacionados na Tabela 5.1. O tempo gasto para processar as bases de dados variou de 2 minutos a 3 horas, dependendo do tamanho da base de dados (valor de n), e também do número de atributos nas mesmas, como pode ser observado na Tabela 5.1. Para as bases de dados que possuem o mesmo tamanho e número de atributos diferentes (Brady 1 e Brady 2) houve um pequeno acréscimo de tempo de processamento, o mesmo ocorrendo entre as bases de dados com mesmo número de atributos e tamanhos diferentes

(Artificial 1 e Artificial 2). Deve-se notar também o tempo de processamento elevado para Artificial 3 e Artificial 4 devido ao número de dados que as mesmas possuem, sendo que o tempo para Artificial 3 é bem superior ao de Artificial 4. As Figuras 5.1 e 5.2 trazem a representação dos tempos médios de processamento gastos nas simulações realizadas, sendo que o objetivo da elaboração de duas figuras foi facilitar a visualização, devido aos tempos elevados referentes à base de dados Artificial 1 e Artificial 2.

Os tempos de execução elevados são justificados pelo fato de ter sido utilizado o método *bootstrap* de alto custo computacional, a execução ter sido feita para vários números de particionamentos ($c = 2, 3, \dots, 8$) e também pela linguagem utilizada, que foi o MATLAB®(MATHWORKS, 1998), a qual não tem como objetivo a velocidade de processamento, mas que foi utilizada neste trabalho por ser de rápida implementação, característica inicialmente requerida, visto que o primordial era a validação do método proposto.

Tabela 5.1 – Relação dos tempos de CPU gastos na execução do algoritmo completo (Algoritmo 4.1) para cada base de dados (descritas na Tabela 4.1), com número de grupos variando de 2 a 8 ($c = 2, 3, \dots, 8$).

Base de Dados	Número de Dados/Atrib.	Percentual de Valores Faltando					Tempo Médio
		2%	5%	10%	20%	30%	
Artificial 1	400/2	7.5 min.	7.2 min.	7.4 min.	8.0 min.	9.3 min.	7.88 min.
Artificial 2	350/2	6.1 min.	6.8 min.	6.9 min.	7.0 min.	7.5 min.	6.86 min.
Artificial 3	2000/4	3.2 horas	3.2 horas	3.0 horas	3.5 horas	3.4 horas	3.26 horas
Artificial 4	1000/3	49.8 min.	49.6 min.	40.8 min.	53.4 min.	49.9 min.	48.70 min.
wine	178/3	2.4 min.	2.6 min.	2.7 min.	2.7 min.	2.7 min.	2.62 min.
iris	150/4	2.1 min.	2.0 min.	2.2 min.	2.2 min.	2.5 min.	2.20 min.
Brady 1	119/20	5.7 min.	5.8 min.	5.7 min.	6.6 min.	7.2 min.	6.20 min.
Brady 2	119/15	4.4 min.	4.5 min.	4.6 min.	4.6 min.	4.4 min.	4.50 min.

5.2 Resultados nas Bases de Dados de Validação

A Tabela 5.2 e a Figura 5.3 mostram o resultado da análise referente à base de dados artificial com quatro grupos (Artificial 1) e possuindo 2% de valores faltando. Observa-

Figura 5.1 – Relação dos tempos médios de CPU, em minutos, gastos para as bases de dados Artificial 1, Artificial 2, iris, wine, Brady 1 e Brady 2 (descritas na Tabela 4.1), com número de grupos variando de 2 a 8 ($c = 2, 3, \dots, 8$).

Figura 5.2 – Relação dos tempos médios de CPU, em minutos, gastos para as bases de dados Artificial 3 e Artificial 4 (descritas na Tabela 4.1), com número de grupos variando de 2 a 8 ($c = 2, 3, \dots, 8$).

se que as medidas Classificação Cruzada (Acc.), Hubert, Fowlkes & Mallows (F&M) e Índice Randômico Ajustado (Rand. Aj.) apresentaram valor máximo no número de grupos correto, ou seja, 4 grupos.

Tabela 5.2 – Resultados das simulações para a base de dados Artificial 1, contendo 400 dados, 2 atributos e 4 grupos. O percentual de valores faltando é de 2%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.

Nro de Grupos	Matriz de Partição			Matriz de Coincidência				
	Diff.	Acc.	F_1	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
02	0.9600	0.7171	0.2751	0.5104	0.6410	0.6727	0.2751	0.0199
03	0.9884	0.8394	0.7611	0.4951	0.4940	0.7766	0.6615	0.4948
04	0.9872	0.8729	0.7441	0.6321	0.5573	0.8705	0.7157	0.6318
05	0.9729	0.8478	0.5809	0.6008	0.5062	0.8820	0.6723	0.5999
06	0.9637	0.8413	0.4601	0.5643	0.4568	0.8862	0.6297	0.5601
07	0.9409	0.8440	0.3706	0.5221	0.4054	0.8876	0.5836	0.5133
08	0.9474	0.8357	0.2959	0.4193	0.3098	0.8859	0.4802	0.4106

Figura 5.3 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Artificial 1, contendo 400 dados, 2 atributos e 4 grupos.

Na Tabela 5.3 e Figura 5.4 mostra-se o resultado da análise referente à base de dados Artificial 2, com sete grupos e possuindo 20% de valores faltando. Observa-se que as medidas Classificação Cruzada (Acc.), Hubert e Índice Randômico Ajustado (Rand. Aj.) apresentaram valor máximo em 7 grupos, que é o valor correto para essa base de dados.

Tabela 5.3 – Resultados das simulações para a base de dados Artificial 2, contendo 350 dados, 2 atributos e 7 grupos. O percentual de valores faltando é de 20%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.

Nro de Grupos	Matriz de Partição			Matriz de Coincidência				
	Diff.	Acc.	F ₁	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
02	0.9814	0.7557	0.7547	0.3295	0.4996	0.6647	0.6660	0.3295
03	0.9021	0.7057	0.5519	0.2984	0.3599	0.6398	0.5234	0.2981
04	0.9230	0.7633	0.5312	0.3300	0.3152	0.7707	0.4768	0.3299
05	0.9132	0.7789	0.4433	0.2980	0.2591	0.8005	0.4089	0.2888
06	0.9417	0.8408	0.5245	0.3891	0.3147	0.8495	0.4769	0.3889
07	0.9575	0.8702	0.5424	0.4643	0.3597	0.8855	0.5291	0.4633
08	0.9339	0.8518	0.3930	0.4594	0.3474	0.8971	0.5164	0.4578

Figura 5.4 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Artificial 2, contendo 350 dados, 2 atributos e 7 grupos.

A Tabela 5.4 e a Figura 5.5 trazem o resultado da análise referente à base de dados Artificial 3, com quatro grupos e possuindo 30% de valores faltando. Esta é uma base de dados com grupos não tão bem separados como os apresentados em Artificial 1 e Artificial 2, mas ainda assim esses grupos apresentam-se perfeitamente delineados. Observa-se que as medidas Classificação Cruzada (Acc.), Hubert, Jaccard, Fowlkes & Mallows (F&M) e Índice Randômico Ajustado (Rand. Aj.) apresentaram valor máximo no número de grupos correto, ou seja, 4 grupos.

Tabela 5.4 – Resultados das simulações para a base de dados Artificial 3, contendo 2000 dados, 4 atributos e 4 grupos. O percentual de valores faltando é de 30%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.

Nro de Grupos	Matriz de Partição			Matriz de Coincidência				
	Diff.	Acc.	F_1	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
02	0.9491	0.6737	0.6658	0.1923	0.4452	0.5969	0.6139	0.1923
03	0.9996	0.8367	0.7548	0.5314	0.5232	0.7922	0.6870	0.5314
04	0.9983	0.8735	0.7477	0.6284	0.5533	0.8699	0.7124	0.6284
05	0.9645	0.8274	0.5362	0.5212	0.4343	0.8612	0.6053	0.5210
06	0.9639	0.8222	0.4438	0.4428	0.3485	0.8724	0.5162	0.4427
07	0.9661	0.8294	0.3924	0.3488	0.2618	0.8813	0.4148	0.3487
08	0.9733	0.8459	0.3780	0.3501	0.2514	0.9047	0.4018	0.3499

As análises foram feitas também em bases de dados com grupos não tão claramente separados (Artificial 4, a *wine* e a *iris*) e que não possuem arranjo estrutural tão uniforme quanto no Artificial 1, Artificial 2 e Artificial 3.

Neste caso, os resultados apresentaram-se de uma forma um pouco mais complexa, no sentido que, para a *wine*, poucas medidas identificaram o número correto de grupos e com relação à *iris*, nenhuma das medidas indicou número de grupos igual a 3. Entretanto, para a *iris*, esse resultado era esperado, considerando-se que somente um dos grupos é linearmente separável (BALKE; MERZ, 1998). No entanto, esta dubiedade entre 2 e 3 grupos torna-se bem clara a partir da análise dos resultados para Hubert e Índice Randômico Ajustado, cuja diferença entre os valores para 2 e 3 grupos é bastante pequena.

Figura 5.5 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Artificial 3, contendo 2000 dados, 4 atributos e 4 grupos.

Mostra-se alguns dos resultados das simulações para essas bases de dados nas Tabelas 5.5 a 5.7 e nas Figuras 5.6 a 5.8.

A Tabela 5.5 e a Figura 5.6 mostram o resultado da análise referente à base de dados *iris* (BALKE; MERZ, 1998) com três grupos e possuindo 5% de valores faltando. Note-se que as medidas Diferença Média Quadrática (Diff.), F_1 , Hubert, Jaccard, Fowlkes & Mallows (F&M) e Índice Randômico Ajustado (Rand. Aj.) apresentaram valor máximo no número de grupos correto, ou seja, 2 grupos.

A Tabela 5.6 e a Figura 5.7 mostram o resultado da análise referente à base de dados Artificial 4, com cinco grupos e possuindo 30% de valores faltando. Observa-se que apenas as medidas Diferença Média Quadrática (Diff.), Classificação Cruzada (Acc.), Hubert e Índice Randômico Ajustado (Rand. Aj.) apresentaram valor máximo em 5 grupos, que é o valor correto para essa base de dados.

Tabela 5.5 – Resultados das simulações feitas para a base de dados *iris*(BALKE; MERZ, 1998) contendo 150 dados, 4 atributos e 3 grupos (dos quais somente um é linearmente separável). O percentual de valores faltando é de 5%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.

Nro de Grupos	Matriz de Partição			Matriz de Coincidência				
	Diff.	Acc.	F_1	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
02	0.9996	0.8319	0.8245	0.5194	0.6286	0.7597	0.7719	0.5193
03	0.9971	0.8174	0.7236	0.5181	0.5044	0.7928	0.6705	0.5181
04	0.9697	0.7779	0.5387	0.4366	0.3847	0.8133	0.5542	0.4359
05	0.9617	0.7847	0.4371	0.4116	0.3346	0.8514	0.4984	0.4109
06	0.9593	0.8087	0.4031	0.4226	0.3312	0.8726	0.4950	0.4213
07	0.9534	0.8070	0.3006	0.3572	0.2634	0.8918	0.4164	0.3560
08	0.9741	0.8411	0.3425	0.3737	0.2640	0.9199	0.4166	0.3733

Figura 5.6 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados *iris*(BALKE; MERZ, 1998) contendo 150 dados, 4 atributos e 3 grupos (dos quais somente um é linearmente separável).

Na Tabela 5.7 e Figura 5.8 mostra-se o resultado da análise referente à base de dados *wine* (BALKE; MERZ, 1998) com três grupos e possuindo 5% de valores faltando. Observa-se que somente as medidas Hubert e Índice Randômico Ajustado (Rand. Aj.) apresentaram valor máximo em 3 grupos, que é o valor correto para essa base de dados.

Tabela 5.6 – Resultados das simulações para a base de dados Artificial 4, contendo 1000 dados, 3 atributos e 5 grupos. O percentual de valores faltando é de 30%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.

Nro de Grupos	Matriz de Partição			Matriz de Coincidência				
	Diff.	Acc.	F_1	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
02	0.9972	0.7669	0.7666	0.3481	0.5106	0.6741	0.6760	0.3481
03	0.9948	0.8163	0.7135	0.4886	0.4922	0.7726	0.6593	0.4886
04	0.9974	0.8385	0.6806	0.4953	0.4355	0.8264	0.6066	0.4953
05	0.9978	0.8558	0.6414	0.5146	0.4226	0.8670	0.5941	0.5146
06	0.9282	0.7910	0.3688	0.4119	0.3227	0.8680	0.4876	0.4118
07	0.9650	0.8356	0.4202	0.3643	0.2719	0.8863	0.4273	0.3641
08	0.9651	0.8461	0.3799	0.3446	0.2493	0.9009	0.3985	0.3445

Figura 5.7 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Artificial 4, contendo 1000 dados, 3 atributos e 5 grupos.

5.3 Resultados nas Bases de Dados de Aplicação

Para a avaliação do número de grupos presentes nas bases de dados de aplicação (Brady 1 e Brady 2) considerou-se os resultados obtidos pelas medidas Hubert e Índice Randômico Ajustado, pois foram as medidas que apresentaram número de grupos corretos

Tabela 5.7 – Resultados das simulações para a base de dados *wine* (BALKE; MERZ, 1998) contendo 178 dados, 13 atributos (neste trabalho utilizou-se somente os atributos 7, 10 e 13) e 3 grupos. O percentual de valores faltando é de 5%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.

Nro de Grupos	Matriz de Partição			Matriz de Coincidência				
	Diff.	Acc.	F_1	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
02	0.9067	0.5986	0.5981	0.0580	0.3661	0.5290	0.5360	0.0579
03	0.9492	0.6891	0.5322	0.2261	0.2750	0.6865	0.4362	0.2215
04	0.9265	0.6870	0.3696	0.1226	0.1287	0.7600	0.2468	0.1101
05	0.9214	0.7042	0.2603	0.1011	0.0797	0.8314	0.1744	0.0828
06	0.9376	0.7422	0.2222	0.1044	0.0602	0.8813	0.1484	0.0774
07	0.9488	0.7737	0.2052	0.1061	0.0522	0.9062	0.1379	0.0742
08	0.9562	0.7966	0.1820	0.0774	0.0334	0.9240	0.1003	0.0486

Figura 5.8 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados *wine* (BALKE; MERZ, 1998) contendo 178 dados, 13 atributos (neste trabalho utilizou-se somente os atributos 7, 10 e 13) e 3 grupos.

em todas as simulações para todas as bases de dados de validação.

Na Tabela 5.8 e Figura 5.9 mostra-se os resultados das análises referentes à base de dados Brady 1 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006) com 2% de valores faltando. Observa-se que somente as medidas Hubert e Índice Randômico Ajustado (Rand.

Aj.) apresentaram valor máximo em 5 grupos.

Tabela 5.8 – Resultados das simulações feitas para a base de dados Brady 1 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006) contendo 119 dados e 49 atributos (neste trabalho utilizou-se somente os atributos de 1 a 20). O percentual de valores faltando é de 2%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.

Nro de Grupos	Matriz de Partição			Matriz de Coincidência				
	Diff.	Acc.	F ₁	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
02	0.9927	0.5512	0.5496	0.0381	0.3549	0.5191	0.5329	0.0381
03	0.9887	0.6002	0.3952	0.2847	0.2247	0.8360	0.3726	0.2785
04	0.9889	0.6544	0.3012	0.3496	0.2272	0.9434	0.3765	0.3395
05	0.9888	0.7011	0.2437	0.3597	0.2264	0.9665	0.3757	0.3495
06	0.9889	0.7371	0.2024	0.3064	0.1803	0.9754	0.3175	0.2911
07	0.9929	0.7683	0.1822	0.1760	0.0855	0.9846	0.1818	0.1493
08	0.9932	0.7927	0.1635	0.2122	0.1120	0.9877	0.2174	0.1927

Figura 5.9 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Brady 1 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006), com 119 dados e 49 atributos (neste trabalho utilizou-se somente os atributos de 1 a 20).

A Tabela 5.9 e a Figura 5.10 mostram os resultados das análises referentes à base de dados Brady 1 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006) com 5% de valores faltando. Observa-se que as medidas Diferença Média Quadrática (Diff.), Hubert e Índice

Randômico Ajustado (Rand. Aj.) apresentaram valor máximo em 6 grupos.

Tabela 5.9 – Resultados das simulações feitas para a base de dados Brady 1 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006) contendo 119 dados e 49 atributos (neste trabalho utilizou-se somente os atributos de 1 a 20). O percentual de valores faltando é de 5%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.

Nro de Grupos	Matriz de Partição			Matriz de Coincidência				
	Diff.	Acc.	F ₁	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
02	0.9905	0.5563	0.5533	0.0418	0.3576	0.5210	0.5269	0.0418
03	0.9833	0.5976	0.3895	0.2567	0.2100	0.8173	0.3541	0.2488
04	0.9887	0.6580	0.3067	0.3434	0.2247	0.9350	0.3746	0.3324
05	0.9891	0.7021	0.2462	0.3639	0.2321	0.9647	0.3808	0.3551
06	0.9870	0.7499	0.2429	0.4005	0.2610	0.9639	0.4178	0.3907
07	0.9928	0.7678	0.1806	0.1715	0.0842	0.9858	0.1769	0.1466
08	0.9920	0.7935	0.1655	0.2183	0.1185	0.9842	0.2251	0.2009

Figura 5.10 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Brady 1 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006), com 119 dados e 49 atributos (neste trabalho utilizou-se somente os atributos de 1 a 20).

As Tabelas 5.10 e 5.11 e as Figuras 5.11 e 5.12 mostram os resultados das análises referentes à base de dados Brady 2 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006) com 2% e 20% de valores faltando, respectivamente. Observa-se que somente as medidas

Hubert e Índice Randômico Ajustado (Rand. Aj.) apresentaram valor máximo em 6 grupos para a Brady 2.

Tabela 5.10 – Resultados das simulações feitas para a base de dados Brady 2 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006) contendo 119 dados e 38 atributos (neste trabalho utilizou-se somente os atributos de 1 a 15). O percentual de valores faltando é de 2%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.

Nro de Grupos	Matriz de Partição			Matriz de Coincidência				
	Diff.	Acc.	F_1	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
02	0.9971	0.6634	0.6621	0.1897	0.4273	0.5949	0.5987	0.1897
03	0.9957	0.7320	0.5616	0.3542	0.3829	0.7238	0.5539	0.3538
04	0.9877	0.7848	0.4978	0.4289	0.3966	0.7877	0.5686	0.4273
05	0.9867	0.7962	0.4383	0.4667	0.3759	0.8564	0.5499	0.4622
06	0.9893	0.8255	0.4282	0.5085	0.3988	0.8899	0.5711	0.5069
07	0.9837	0.8260	0.3603	0.4449	0.3250	0.9112	0.4928	0.4420
08	0.9852	0.8394	0.3295	0.4428	0.3120	0.9312	0.4788	0.4389

Figura 5.11 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Brady 2 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006), com 119 dados e 38 atributos (neste trabalho utilizou-se somente os atributos de 1 a 15).

Tabela 5.11 – Resultados das simulações feitas para a base de dados Brady 2 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006) contendo 119 dados e 38 atributos (neste trabalho utilizou-se somente os atributos de 1 a 15). O percentual de valores faltando é de 20%. Todas as medidas foram calculadas usando *t-norma* produto e *nil-potent* mínimo. Os valores em negrito correspondem aos valores máximos de cada medida.

Nro de Grupos	Matriz de Partição			Matriz de Coincidência				
	Diff.	Acc.	F_1	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
02	0.9961	0.6801	0.6781	0.2174	0.4413	0.6087	0.6124	0.2174
03	0.9930	0.7257	0.5452	0.3244	0.3658	0.7089	0.5359	0.3239
04	0.9859	0.7815	0.4917	0.4171	0.3865	0.7862	0.5577	0.4165
05	0.9861	0.7915	0.4174	0.4534	0.3674	0.8545	0.5388	0.4514
06	0.9888	0.8227	0.4193	0.4789	0.3741	0.8832	0.5453	0.4773
07	0.9800	0.8285	0.3598	0.4308	0.3187	0.9021	0.4843	0.4290
08	0.9851	0.8445	0.3431	0.4536	0.3262	0.9248	0.4936	0.4513

Figura 5.12 – Resultados das medidas de comparação, entre matrizes de partição e entre matrizes de pertinência, para a base de dados Brady 2 (MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006), com 119 dados e 38 atributos (neste trabalho utilizou-se somente os atributos de 1 a 15).

A Tabela 5.12 sumariza o número de agrupamentos obtidos para as simulações feitas em todos os percentuais de valores faltando para Brady 1 e Brady 2.

Estes resultados apresentam concordância, em relação a outras análises feitas, onde utilizou-se a base de dados *Bradyrhizobium* (MILAGRE, 2003; MILAGRE *et al.*, 2008;

Tabela 5.12 – Resumo do número de agrupamentos obtidos para as bases de dados Brady 1 e Brady 2, referentes às medidas Hubert e Índice Randômico Ajustado.

% Valores Faltando	Nro de Grupos Obtidos Brady 1	Nro de Grupos Obtidos Brady 2
2%	5	6
5%	6	6
10%	5	6
20%	6	6
30%	6	5

GERMANO *et al.*, 2006; CHRIST, 2007; TALAVERA, 2007; LLERENA, 2008), descrita na Seção 4.1.3. Mas, o método proposto aqui apresenta algumas vantagens em relação aos demais. Uma primeira consideração é que, nesses métodos, devido à retirada de algumas informações de forma aleatória e ao se realizar as análises, inclui-se variações nos dados que possivelmente estariam causando erros, pois os dados possuem variados graus de deformação causados por manchas no gel, arraste do DNA e bandas não definidas como pode ser visto na Figura 5.13. Desta forma, alguns dados deveriam ser definidos como incompletos devido à má definição no gel.

Estudos iniciais realizados em Milagre (2003) e Milagre *et al.* (2008), mostram que possivelmente os dados não se apresentavam em grupos bem definidos e, neste caso, os métodos baseados em lógica nebulosa seriam mais indicados. O uso de algoritmos hierárquicos (MILAGRE, 2003; MILAGRE *et al.*, 2008; GERMANO *et al.*, 2006), nestes casos, deveriam ser reavaliados pois, além de ser arbitrado o ponto de corte no dendograma, a própria representação dos dados na forma hierárquica seria muito dependente da qualidade dos dados. Sendo assim, no caso da presença de dados incompletos ou com baixa qualidade, o estudo sobre a presença das partições naturais existentes nos dados poderia ser afetada de forma significativa. Aliado a isso, esse tipo de análise sofre com a escolha da métrica adequada para avaliar a similaridade ou dissimilaridade pois freqüentemente

Figura 5.13 – Gel de eletroforese da região 16S rRNA das estirpes SEMIA (026 a 050) digeridas com enzima de restrição *Cfo* I (MILAGRE, 2003).

métricas distintas podem levar a agrupamentos distintos (TALAVERA, 2007).

O método proposto por Christ (2007) foi o de uma rede neural do tipo ART-2 cuja variação do parâmetro de vigilância indicava a formação dos grupos. Embora o particionamento fosse realizado de forma *hard*, a própria topologia da rede neural excluía os dados que não se enquadrassem em grupos bem estabelecidos. Posteriormente foram definidas métricas entre os grupos para definir o particionamento ótimo. Desta forma, neste trabalho, seria preciso definir-se o comportamento do sistema com a presença de dados com baixa qualidade ou mesmo errados e também, qual o limite de tolerância do sistema à presença desses tipo de dados.

Em Talavera (2007) considerou-se um modelo gráfico probabilístico/redes bayesianas para a formação dos grupos. Essa abordagem considera a formação dos grupos de forma nebulosa, mas as informações *a priori*, sobre as quais o estudo é realizado, baseia-se em qual a veracidade da distribuição uniforme dos dados nos grupos. Assim como em Christ

(2007), essa abordagem deve ser mais bem estudada para os casos onde exista a presença de dados com baixa qualidade.

Em Llerena (2008) trabalha-se com a redução dimensional do sistema seguido de sua representação gráfica. A definição do número de dados é realizada de forma subjetiva e pode variar de usuário para usuário.

A Tabela 5.13 traz o resumo dos métodos citados neste trabalho, que analisam a base de dados *Bradyrhizobium*, em relação às características de tipo da base de dados, tipo de classificação utilizada e participação do usuário na execução do método.

Tabela 5.13 – Resumo dos métodos de análise da base de dados *Bradyrhizobium*.

Método de Análise	Base de Dados		Classificação			Usuário	
	Completa	Incompleta	Hard	Nebulosa	Hierárquica	Sim	Não
Proposto	X	X		X			X
Milagre(2003)	X				X	X	
Germano(2006)	X				X	X	
Christ(2007)	X			X			X
Talavera(2007)	X			X			X
Llerena(2008)	X		X			X	

5.4 Resumo dos Resultados Obtidos

A Figura 5.14 traz representação dos percentuais de resultados corretos obtidos para cada medida de comparação, em relação às simulações feitas.

As Tabelas 5.14 e 5.15 sumarizam os resultados das simulações para todos os percentuais de valores faltando e todos os conjuntos de dados que foram utilizados neste trabalho.

Pode-se visualizar que somente as medidas Hubert e Índice Randômico Ajustado apresentaram o número correto de grupos para todas as bases de dados, ou seja, em 100%

Figura 5.14 – Relação dos percentuais de resultados corretos obtidos para cada medida de comparação em relação às simulações feitas, com número de grupos variando de 2 a 8 ($c = 2, 3, \dots, 8$) e percentuais de valores faltando iguais a 2%, 5%, 10%, 20% e 30%.

das simulações. É importante observar que essas medidas apresentam-se sobrepostas em quase todos os gráficos (não somente nos aqui apresentados, mas em todas as simulações feitas), possuindo assim, sempre valores muito próximos.

A medida Índice Randômico (Rand.) apresentou somente um resultado correto para a base de dados Artificial 2 com 10% de valores faltando. Portanto, não se mostrou adequada para o tipo de análise que apresentada neste estudo.

As medidas Classificação Cruzada (Acc.), Jaccard e Fowlkes & Mallows (F&M) também apresentaram resultados corretos em 47.5%, 40% e 42.5% das simulações, respectivamente.

Fazendo-se uma análise sobre os resultados obtidos para as medidas Hubert e Randômico Ajustado (Rand. Aj.), com relação ao formato, quantidade de atributos e quantidade

de dados em cada base de dados, pode-se afirmar que o método trabalhou de forma eficiente em todas as bases de dados utilizadas, desde as mais simples, com grupos bem separados, passando pelas bases com dados sobrepostos até bases de dados de difícil análise, como a Brady 1 e Brady 2.

Tabela 5.14 – Sumário dos resultados das simulações feitas neste trabalho para os percentuais de valores faltando de 2%, 5% e 10%. Os *pontos* em negrito representam onde a medida atingiu o seu valor máximo coincidindo com o número correto de grupos existente na base de dados. Pode-se observar que somente as medidas Hubert e Randômico Ajustado (Rand. Aj.) apresentaram o número correto de grupos em todas as simulações.

2% de valores faltando								
	Diff.	Acc.	F_1	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
Art. 1		•		•	•		•	•
Art. 2				•			•	•
Art. 3	•	•	•	•	•		•	•
Art. 4		•		•				•
<i>iris</i>	•		•	•	•		•	•
<i>wine</i>				•				•
Brady 1				•				•
Brady 2				•				•
5% de valores faltando								
	Diff.	Acc.	F_1	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
Art. 1	•	•		•	•		•	•
Art. 2				•	•		•	•
Art. 3	•	•	•	•	•		•	•
Art. 4		•		•				•
<i>iris</i>	•		•	•	•		•	•
<i>wine</i>				•				•
Brady 1	•			•				•
Brady 2				•				•
10% de valores faltando								
	Diff.	Acc.	F_1	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
Art. 1	•	•	•	•	•		•	•
Art. 2		•		•		•		•
Art. 3	•	•	•	•	•		•	•
Art. 4		•		•				•
<i>iris</i>	•		•	•	•		•	•
<i>wine</i>				•				•
Brady 1				•				•
Brady 2				•				•

Tabela 5.15 – Sumário dos resultados das simulações feitas neste trabalho para os percentuais de valores faltando de 20% e 30%. Os *pontos* em **negrito** representam onde a medida atingiu o seu valor máximo coincidindo com o número correto de grupos existente na base de dados. Pode-se observar que somente as medidas Hubert e Randômico Ajustado (Rand. Aj.) apresentaram o número correto de grupos em todas as simulações.

20% de valores faltando								
	Diff.	Acc.	F_1	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
Art. 1		●		●	●		●	●
Art. 2		●		●				●
Art. 3		●	●	●	●		●	●
Art. 4		●		●				●
<i>iris</i>	●		●	●	●		●	●
<i>wine</i>	●			●				●
Brady 1	●			●				●
Brady 2				●				●
30% de valores faltando								
	Diff.	Acc.	F_1	Hubert	Jaccard	Rand.	F&M	Rand. Aj.
Art. 1	●	●	●	●	●		●	●
Art. 2	●	●		●				●
Art. 3		●		●	●		●	●
Art. 4	●	●		●				●
<i>iris</i>	●	●	●	●	●		●	●
<i>wine</i>	●			●				●
Brady 1	●			●				●
Brady 2				●				●

Capítulo 6

Conclusão

O trabalho apresentou um método para identificar grupos existentes em bases de dados contendo dados incompletos, utilizando-se lógica nebulosa e reamostragem *bootstrap*.

O método apresentou-se eficiente na análise de todas as bases de dados. Até mesmo em presença de altos percentuais de valores faltando como 20% e 30% foi possível identificar o número de grupos em todas elas. Assim, os resultados mostram que nosso método é capaz de identificar partições relevantes até mesmo em presença de altos níveis de valores faltando.

Os resultados obtidos na análise das bases de dados de aplicação (Brady 1 e Brady 2) concordaram com outros métodos desenvolvidos por pesquisadores que utilizaram a base de dados *Bradyrhizobium* como estudo, o que novamente valida o método proposto. Cabe ressaltar que os estudos feitos em Christ (2007) e Talavera (2007) tiveram como objetivo o particionamento inicial dos dados e o trabalho atual e o descrito em Milagre (2003) foram no sentido de validar um determinado algoritmo de particionamento. Desta forma, devem ser mais bem estudadas as possíveis combinações entre as técnicas apresentadas em Christ (2007) e Talavera (2007) com Milagre (2003) e o trabalho atual.

Desta forma, as principais contribuições deste trabalho são:

- Analisar a base de dados como ela se apresenta, pois no método proposto, o agrupamento é feito sem alteração da mesma. Isto implica que nenhuma suposição é feita sobre os dados incompletos para a inserção de valores onde eles estão faltando e também que não são feitas exclusões de dados incompletos. Em outros métodos, o agrupamento é baseado na inclusão de valores que faltam ou na exclusão do dado incompleto, já que a maioria dos algoritmos de agrupamento só trabalham com bases de dados completas. Porém, em alguns casos, o valor faltando para um determinado atributo pode realmente não existir, e o processo de inclusão ou exclusão pode levar a resultados imprecisos;
- Realizar o agrupamento independentemente da forma de distribuição, quantidade de dados e número de atributos existentes na base de dados, visto que foram utilizadas oito bases de dados com variações de todas essas características para a validação do método;
- Realizar agrupamentos com altas taxas de valores faltando;
- Apresentar várias medidas de comparação entre matrizes de partição e coincidência.

As medidas de comparação de matrizes de partição F_1 , Classificação Cruzada (Acc.) e Erro Médio Quadrático (Diff.) não apresentaram bons resultados, sendo que entre as três a Classificação Cruzada foi a que apresentou o maior índice de resultados corretos (47.5%). As medidas de comparação entre matrizes de coincidência apresentaram os melhores resultados. Sendo que, as medidas Índice Hubert (Hubert) e Índice Randômico Ajustado (Rand. Aj.) foram as que encontraram o melhor resultado experimental, indicando valor máximo para o número correto de grupos em todas as simulações. As demais medidas de

comparação de matrizes de coincidência, coeficiente de Jaccard (Jaccard), Índice Randômico (Rand.) e Índice de Fowlkes & Mallows (F&M) não conseguiram identificar o número correto de grupos na maioria das simulações. A medida Rand. apresentou somente um resultado correto entre os experimentos.

Como sugestão para trabalhos futuros, citam-se os seguintes itens:

- Testar novas combinações de *t-normas* para verificar se existem combinações que possam melhorar os resultados obtidos, como por exemplo aumentando a eficiência de alguma medida;
- Realizar testes variando o número de amostras, ou seja, diminuir o número de amostras para verificar se ainda assim é possível obter bons resultados, pois isto diminuiria o tempo de processamento. Ou, se, aumentando o número de amostras seria possível aumentar a eficiência de alguma medida ou obter-se valores máximos ainda mais diferenciados dos demais;
- Buscar novas medidas para aplicar o método;
- Fazer um estudo do melhor percentual de sub-amostragem a ser aplicado (neste trabalho utilizou-se 90%). Estudos iniciais com 80% não deram resultado;
- Conversão dos programas para uma linguagem de processamento mais rápida, como a linguagem C;
- Estudo de combinações de novas técnicas com o método proposto.

Referências Bibliográficas

ACOCK, A. C. Working with missing values. **Journal of Marriage and Family**, Blackwell Publishing, Malden, Massachusetts, USA, v. 67, n. 4, p. 1012–1028, 2005.

BACKER, E. **Computer-Assisted Reasoning in Cluster Analysis**. 1st. ed. New York: Prentice Hall, 1995.

BALKE, C. L.; MERZ, C. J. **UCI Repository of Machine Learning Databases**. Irvine, CA, USA: University of California, 1998.

<http://mllearn.ics.uci.edu/MLRepository.html>. [Online, acessado em 02/08/2006].

BATISTA, G. E. A. P. A.; MONARD, M. C. An analysis of four missing data treatment methods for supervised learning. **Applied Artificial Intelligence**, v. 17, n. 5, p. 519–533, May 2003.

BENNETT, D. A. How can i deal with missing data in my study? **Australian and New Zealand Journal of Public Health**, v. 25, n. 5, p. 464–469, May 2001.

BEZDEK, J. C.; KELLER, J.; KRISHNAPURAM, R.; PAL, N. R. **Fuzzy Models and Algorithms for Pattern Recognition and Image Processing**. New York, USA: Springer, 2005.

BEZDEK, J. C.; PAL, S. K. **Fuzzy Models for Pattern Recognition: methods that search for structures in data**. New York, NY, USA: IEEE Press, 1992.

BONNISSONE, P.; TONG, R. Editorial: reasoning with uncertainty in expert systems. **International Journal of Man Machine Studies**, v. 2, n. 3, p. 241–250, 1985.

BORGELT, C. **Prototype-based Classification and Clustering**. Tese (Doutorado) — Otto-von-Guericke-Universität, Magdeburg, Germany, November 2005.

BORGELT, C. Resampling for fuzzy clustering. In: **Proc. Symposium on Fuzzy Systems in Computer Science**. Otto-von-Guericke-Universität: Magdeburg, Germany, 2006.

BOUTIN, F.; HASCOËT, M. Cluster validity indices for graph partitioning. In: **8th International Conference on Information Visualization (IV 2004)**. London, England, UK: [s.n.], 2004.

CHRIST, R. E. **Classificação de Bactérias do gênero *Bradyrhizobium* usando uma rede neural ART2 com dados de eletroforese de genes ribossomais**. Dissertação (Mestrado) — Universidade de São Paulo, São Carlos, SP, Brasil, 2007.

- CROWLEY, P. H. Resampling methods for computation-intensive data analysis in ecology and evolution. **Annual Review of Ecology and Systematics**, v. 23, p. 405–447, 1992.
- DEMPSTER, A. P.; RUBIN, D. B. Incomplete data in sample surveys. **Sample Surveys Vol. II: Theory and Annotated Bibliography**, Academic Press, New York, NY, USA, v. 2, p. 3–10, 1983.
- DIXON, J. K. Pattern recognition with partly missing data. **IEEE Transactions on System, Man and Cybernetics**, v. 9, n. 10, p. 617–621, October 1979.
- DÖRING, C.; LESOT, M. J.; KRUSE, R. Data analysis with fuzzy clustering methods. **Elsevier - Computational Statistics and Data Analysis**, v. 51, n. 1, p. 192–214, November 2006.
- EFRON, B.; TIBSHIRANI, R. J. **An Introduction to the Bootstrap**. New York, NY, USA: Chapman and Hall, 1993.
- EVERITT, B.; LANDAU, S.; LEESE, M. **Cluster Analysis**. 4th. ed. New York, NY, USA: A Hodder Arnold Publication, 2001.
- FOWLKES, E.; MALLOWS, C. A method for comparing two hierarchical clusterings. **ADAJ**, v. 78, n. 383, p. 553–569, 1983.
- GERMANO, M. G.; MENNA, P.; MOSTASSO, F. L.; HUNGRIA, M. Rflp analysis of the rna operon of a brazilian collection of bradyrhizobial strains from 33 legume species. **International Journal of Systematic and Evolutionary Microbiology**, The Society for General Microbiology, v. 56, p. 217–229, 2006.
- GORDON, A. D. **Classification : methods for the exploratory analysis of multivariate data**. 2nd. ed. London, New York, USA: Chapman and Hall, 1999.
- HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Cluster validity checking methods: Part ii. **ACM SIGMOD**, ACM Press, New York, NY, USA, v. 31, n. 3, p. 19–27, September 2002.
- HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Cluster validity methods: Part i. **ACM SIGMOD**, ACM Press, New York, NY, USA, v. 31, n. 2, p. 40–45, September 2002.
- HATHAWAY, R. J.; BEZDEK, J. C. Fuzzy c-means clustering of incomplete data. **IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics**, v. 31, n. 5, p. 735–744, October 2001.
- HOWELL, D. C. **Treatment of Missing Data**. Vermont, USA: University of Vermont, 2006. http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html. [Online, acessado em 15/10/2007].
- HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of Classification**, v. 2, p. 193–218, 1985.

HUNGRIA, M.; VARGAS, M. A. T.; ARAÚJO, R. S. Fixação biológica do nitrogênio em feijoeiro. **Biologia dos Solos dos Cerrados**, EMBRAPA-CPAC, p. 189–225, 1997.

HÖPPNER, F.; KLAWONN, F.; KRUSE, R.; RUNKLER, T. **Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition**. Chichester, England: John Wiley and Sons, 1999.

JAFELICE, R.; BARROS, L.; BASSANEZI, R. Teoria dos conjuntos fuzzy com aplicações. **SBMAC, Sociedade Brasileira de Matemática Aplicada e Computacional - Notas em Matemática Aplicada**, v. 17, 2005.

JAIN, A. K.; DUBES, R. C. **Algorithms for Clustering Data**. Upper Saddle River, NJ, USA: Prentice Hall, 1988.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: A review. **ACM Computing Surveys (CSUR)**, ACM Press, New York, NY, USA, v. 31, n. 3, p. 264–323, September 1999.

JOHNSON, R. A. **Applied Multivariate Statistical Analysis**. 3rd. ed. New Jersey, USA: Prentice Hall, 1992.

LAW, M. H.; JAIN, A. K. **Cluster Validity by Bootstrapping Partitions**. East Lansing, Michigan, USA, 2003.

LEVINE, E.; DOMANY, E. Resampling method for unsupervised estimation of cluster validity. **Neural Computation**, v. 13, n. 11, p. 2573–2593, November 2001.

LITTLE, R. J. A.; RUBIN, D. B. **Statistical analysis with missing data**. New York, USA: John Wiley and Sons, 1987.

LITTLE, R. J. A.; RUBIN, D. B. **Applied Multivariate Statistical Analysis**. 2nd. ed. New Jersey, USA: John Wiley and Sons, 2002.

LIU, W. Z.; WHITE, A. P.; THOMPSON, S. G.; BRAMER, M. A. Techniques for dealing with missing values in classification. In: **Second International Symposium on Intelligent Data Analysis**. London, England, UK: Springer-Verlag, 1997. p. 527–536.

LLERENA, S. E. **Mapeamento de Dados Genômicos Utilizando Multidimensional Scaling**. Dissertação (Mestrado) — Universidade de São Paulo, São Carlos, SP, Brasil, 2008.

MATHWORKS, I. S. M. T. **The Language of Technical Computing**. Natick, Massachusetts, USA: [s.n.], 1998.

MENGER, K. Statistical metrics. **National Academy of Science of the United States of America**, v. 28, n. 12, p. 535–537, December 1942.

MILAGRE, S. T. **Análise de Estabilidade de Cluster em uma Coleção Brasileira de Bactérias Diazotróficas do Gênero *Bradyrhizobium***. Dissertação (Mestrado) — Universidade Estadual de Londrina, Londrina, PR, Brasil, 2003.

MILAGRE, S. T.; MACIEL, C. D. Clustering analysis in datasets within missing values using fuzzy c-means and bootstrapping. **Elsevier**, a ser enviado, 2008.

MILAGRE, S. T.; MACIEL, C. D. Performance analysis of restriction enzymes based in results of cluster stability. **Ciência & Engenharia**, no prelo, 2008.

MILAGRE, S. T.; MACIEL, C. D.; PEREIRA, J. C.; PEREIRA, A. A. Estabilidade de classificadores usando a técnica de reamostragem. In: **6th Brazilian Conference on Dynamics, Control and Their Applications (Dincon 2007)**. São José do Rio Preto, SP, Brasil: [s.n.], 2007.

MILAGRE, S. T.; MACIEL, C. D.; PEREIRA, J. C.; PEREIRA, A. A. Fuzzy cluster stability analysis with missing values using resampling. In: **International Symposium on Mathematical and Computacional Biology (BIOMAT 2007)**. Armação de Búzios, RJ, Brasil: [s.n.], 2007.

MILAGRE, S. T.; MACIEL, C. D.; SHINODA, A. A.; HUNGRIA, M.; ALMEIDA, J. R. B. Multidimensional cluster stability analysis from a brazilian *Bradyrhizobium* sp. RFLP/PCR Data set. **Elsevier - Journal of Computational and Applied Mathematics**, in press, 2008.

MOTRO, A.; SMETS, P. Sources of uncertainty in information systems, proceedings of the 2nd workshop on uncertainty management and information systems: From needs to solutions. In: . Catalina Island, CA, USA: University of Vermont, 1993.

NUOVO, A. G. D.; CATANIA, V.; NUOVO, S. D.; BUONO, S. Envolving fuzzy c-means: An intelligent technique for efficient diagnosis of children mental retardation level from databases with missing values. In: **International Conference on Artificial Intelligence (ICAI '06)**. Las Vegas, USA: [s.n.], 2006.

PAL, N. R.; BEZDEK, J. C. On cluster validity for the fuzzy c-means model. **IEEE Transactions on Fuzzy Systems**, IEEE Computer Society, Washington, DC, USA, v. 3, n. 3, p. 370–379, August 1995.

PARSONS, S. Current approaches to handling imperfect information in data and knowledge bases. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 3, p. 353–372, June 1996.

RIJSBERGEN, C. J. van. **Information Retrieval**. London, England, UK: Butterworth, 1979.

ROSSUM, D. van; SCHUURMANS, F. P.; GILLIS, M.; MUYOTCHA, A.; VERSELD, H. W. van. Genetic and phenotypic analyses of *Bradyrhizobium* strains nodulating peanuts (*Arachis hypogea* L.). **Applied and Environmental Microbiology**, v. 61, p. 1599–1609, 1995.

ROTH, V.; LANGE, T.; BRAUN, M. L.; BUHMANN, J. M. A resampling approach to cluster validation. In: **15th Computational Statistics (COMPSTAT 2002)**. Berlin, Germany: Physica-Verlag Heidelberg, 2002.

SANDRI, S.; CORREA, C. Lógica nebulosa. In: **V Escola de Redes Neurais**. ITA, São José dos Campos, SP, Brasil: [s.n.], 1999.

SARKAR, M.; LEONG, T. Y. Fuzzy k-means clustering with missing values. In: **American Medical Informatics Association Annual Symposium (AMIA)**. Philadelphia, USA: Medical Publishers, 2001. p. 588–592.

SCHAFER, J. L. **Analysis of Incomplete Multivariate Data**. 2nd. ed. London, New York, USA: Chapman and Hall/CRC, 1999.

SCHAFER, J. L.; GRAHAM, J. W. Missing data: Our view of the state of the art. **Psychological Methods**, Methodology Center, Pennsylvania State University, USA, v. 7, n. 2, p. 147–177, November 2002.

SO, R. B.; LADHA, J. K.; YOUNG, J. P. W. Photosynthetic symbionts of *Aeschynomene* spp. form a cluster with bradyrhizobia on the basis of fatty acid and rrna analysis. **International Journal of Systematic Bacteriology**, v. 44, p. 392–403, 1994.

TALAVERA, E. R. V. **Métodos Bayesianos Aplicados em Taxonomia Molecular**. Dissertação (Mestrado) — Universidade de São Paulo, São Carlos, SP, Brasil, 2007.

TIMM, H.; DÖRING, C.; KRUSE, R. Different approaches to fuzzy clustering of incomplete datasets. **Elsevier - International Journal of Approximate Reasoning**, v. 35, n. 3, p. 239–249, March 2004.

TIMM, H.; KLAWONN, F. Classification of data with missing values. In: **6th European Congress on Intelligent Techniques and Soft Computing**. Aachen, Germany: [s.n.], 1998. p. 1304–1308.

TIMM, H.; KLAWONN, F. Different approaches for fuzzy cluster analysis with missing values. In: **7th European Congress on Intelligent Techniques and Soft Computing**. Aachen, Germany: [s.n.], 1999. p. 639–644.

XIE, X. L.; BENI, G. A validity measure for fuzzy clustering. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE Computer Society, Washington, DC, USA, v. 13, n. 8, p. 841–847, November 1991.

YU, J. General c-means clustering model. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 27, n. 8, p. 1197–1211, August 2005.

ZADEH, L. A. Fuzzy sets. **Information and Control**, v. 8, n. 1, p. 338–353, November 1965.

ZADEH, L. A. Fuzzy logic, neural networks and soft computing. **Comunnications of the ACM**, ACM Press, New York, NY, USA, v. 37, n. 3, p. 77–84, March 1994.

ZHANG, D.; CHEN, S. Clustering incomplete data using kernel-based fuzzy c-means algorithm. **Neural Processing Letters**, Kluwer Academic, Netherlands, NY, USA, v. 18, n. 12, p. 155–162, December 2003.

ZHOU, J. **A Different Way to Solve the Missing Value Problem: The Case of Equal Employment Opportunity Data**. Dissertação (Mestrado) — University of Maryland, 2006.