UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Novel visual approaches for attribute analysis, selection, and prediction

Erasmo Artur da Silva Júnior

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP SÃO CARLOS

Erasmo Artur da Silva Júnior

Novel visual approaches for attribute analysis, selection, and prediction

Doctoral dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Rosane Minghim

USP – São Carlos
August 2020

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

A792n

Artur, Erasmo S. Jr.
Novel visual approaches for attribute analysis, selection, and prediction / Erasmo S. Jr. Artur; orientadora Rosane Minghim. -- São Carlos, 2020.
87 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2020.

1. Visual Analytics. 2. Feature Space Analysis. 3. Visual Feature Selection. 4. Exploratory Data Visualization. 5. Regression and Correlation Analysis. I. Minghim, Rosane, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Erasmo Artur da Silva Júnior

Novas abordagens visuais para análise, seleção e predição de atributos

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Rosane Minghim

USP – São Carlos
Agosto de 2020

RESUMO

ARTUR, E. S. JR. **Novas abordagens visuais para análise, seleção e predição de atributos**. 2020. 87 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

Enquanto as capacidades de coleta e armazenamento de dados crescem extensamente hoje em dia, a capacidade geral de processar e analisar grande quantidade de dados cresce em uma taxa mais lenta. Essa assincronia introduz novos desafios impactando métodos que lidam com essa enorme quantidade de dados, como abordagens em mineração, estatística e aprendizado de máquina. Para ajudar a diminuir esta lacuna, abordagens visuais vem sendo propostas para combinar habilidades humanas com soluções consolidadas no desenvolvimento de ferramentas interativas que permitem uma investigação mais aprofundada dos dados. Uma quantidade substancial de abordagens visuais se concentra em técnicas baseadas em itens, onde os itens de dados representam os objetos de primeira ordem. Contudo, informações valiosas frequentemente aparecem a partir de observações de relacionamentos entre atributos, como os relacionamentos entre atributos categóricos e numéricos que frequentemente codificam informações relevantes. Nesse contexto, uma abordagem de análise visual para a exploração do espaço de atributos é fundamental, tanto quando há hipóteses de correlações que devem ser confirmadas, como também nos casos em que tais relações são desconhecidas ou imprevisíveis. Nesta Tese, propomos uma abordagem para análise de atributos com base na apresentação simultânea de múltiplas correlações por meio de uma visualização baseada em pontos, a qual visa construir mapas cognitivos desses relacionamentos para o usuário final. Além disso, o processo de análise oferece suporte a tarefas adicionais como seleção de atributos e criação de modelos de predição com base em um resultado alvo. Mostramos a eficiência das abordagens através de uma série de estudos de caso e cenários de uso que envolvem conjuntos de dados em contextos distintos.

Palavras-chave: Análise visual, Visualização de dados, Análise de espaço de atributos, Seleção de atributos, Análise visual preditiva.

ABSTRACT

ARTUR, E. S. JR. **Novel visual approaches for attribute analysis, selection, and prediction**. 2020. 87 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

While data collection and storage capabilities grow widely nowadays, the general ability to process and analyze large amounts of data increases at a slower rate. This asynchrony introduces new challenges touching methods for large amounts of data, such as the ones in data mining, statistics, and machine learning. To help addressing this gap, visual approaches have been proposed to combine human capabilities with consolidated solutions in the development of interactive tools that allow a more in-depth investigation of the data. A substantial amount of visual approaches has focused on items-based techniques, where the data items represent the first-order objects. Nevertheless, valuable knowledge frequently appears from observations of relationships between attributes of these data items, such as the relationships between numerical and categorical variables, which often encode relevant information. In this context, a visual analysis approach for attribute space exploration is paramount, both when there are hypotheses of correlations that must be confirmed, and also in cases where such relationships are unknown or unforeseen. In this Thesis, we propose an approach for attribute analysis based on the simultaneous presentation of multiple correlations through a point-based visualization aiming to build cognitive maps of these relationships to the end-user. Also, the analysis process then supports additional tasks such as feature selection and the development of prediction models based on a target outcome. We show the efficiency of the approaches through a series of case studies and usage scenarios involving real data sets in distinct contexts.

Keywords: Visual analytics, Data visualization, Attribute space analysis, Feature selection, Predictive visual analytics.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AIS	Abbreviated Injury Scale
AKI	Acute Kidney Injury
ARDS	Acute Respiratory Distress Syndrome
AUC	Area Under the Curve
COVID-19	Coronavirus Disease 2019
DA	Dimensional Anchor
DR	Dimensionality Reduction
EHRs	Electronic Health Records
FS	Feature Selection
GCS	Glasgow Coma Scale
ISS	Injury Severity Score
LDA	Linear Discriminant Analysis
LoRRViz	Logistic Regression Radial Visualization
LR	Logistic Regression
MODS	Multiple Organ Dysfunction Syndrome
ROC	Receiver Operating Characteristic
RR	Respiratory Rate
RSS	Residual Sum of Squares
RTS	Revised Trauma Score
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SBP	Systolic Blood Pressure
TBI	Traumatic Brain Injury
TF–IDF	Term Frequency–Inverse Document Frequency
TRISS	Trauma Revised Injury Severity Score
TSS	Total Sum of Squares
WHO	World Health Organization

CONTENTS

CHAPTER

1

INTRODUCTION

Over the past few decades, the increased computational efficiency and capacity for data collection and storage have made the availability of multidimensional data very high. However, dealing with large data sets is considerably challenging, especially in the context of demanding application domain. This scenario demands development of strategies to extract useful knowledge from large data sets; such strategies have been proposed in a diversity of fields such as statistics, data mining, and machine learning.

While a variety of solutions try to automatically extract information from multidimensional data sets, they struggle to aggregate the tacit knowledge of humans in the process. In this sense, the field of data visualization attempts to address this gap by inserting the user into the loop to combine flexibility, creativity, and background knowledge into the analysis tasks of multidimensional data sets (KEIM; MANSMANN; THOMAS, 2010; CUI *et al.*, 2019). Visual interactive approaches often allow users to gain insights about data sets in multiple fields as health, business, government, environment, weather and climate analysis, as well as in the prevention of undesired events, such as accidents, fraud, market instability and epidemics.

In general, the visualization community has made efforts in the development of approaches that visually explore data spaces, where the items in a data set are either explicitly or implicitly represented in the visualization (TURKAY; FILZMOSER; HAUSER, 2011). Yet, relevant information in the data set often lies in the interrelationships of attributes. For example, in medical research, analysts look for attributes that correlate with treatment or clinical results to prevent adverse outcomes and also to improve the quality of internal protocols. Parallel situations can be observed in most multidimensional data set applications.

Another advantage of attribute-based visualizations is the support for additional tasks such as feature selection (FS). Although many efficient automatic methods have been proposed for FS purposes, once again they are not currently able to aggregate prior human knowledge in the process. Interactive approaches that aim at building classification systems should allow users

who are familiar with the data to effectively apply their domain knowledge (WARE *et al.*, 2001; ZHENG *et al.*, 2016). Visual strategies that reveal the relationship of attributes are helpful to let users understand and recognize the various degrees of importance of attributes (frequently in isolation and also when combined) related to some target event or object of interest, as well as redundancy amongst them. Consequently, users should be able to perform potentially useful FS tasks to find the minimum subset of attributes capable of describing the data from the perspective of the target phenomenon.

As a consequence of enabling users to perform relevant FS tasks, a range of applications can take advantage of the selected representative subsets. An example is logistic regression analysis, which suffers from some limitations such as the multicollinearity phenomenon. It occurs when a large number of correlated predictor attributes are included in the model and can generate unreliable and unstable estimates of regression coefficients (CHATTERJEE; HADI, 2006). Thus, an accurate FS should be useful to eliminate or decrease the constraints of applying these methods combined with multidimensional data sets.

1.1 Motivation and Objective

Regardless the growing interest in the visualization community to represent attribute spaces, few results have been successful in adequately revealing relationships between all data attributes and target labels or features. For example, in a trauma registry data set, analysts may look for insights by investigating the relationship between attributes that encode relevant information of the data set (e.g., clinical outcomes, such as patients' final condition or cause of death). Other regular categorical attributes can also be employed as a target in searching for new insights, such as age ranges, length of care, or geographic region of the trauma event. These attributes can be interpreted as labels and be used to extract associated relevance and predictive power of other attributes. Our proposal is to present these relationships visually as a cognitive map that allows an in-depth investigation of attributes related to some phenomena encoded in a candidate target attribute.

The main objective of this Thesis is to provide visual approaches that allow users to investigate relationships between regular attributes and other candidate target attributes, in order to locate their relevance in coding valuable knowledge of the data set. Through a point-based radial visualization, we map attributes according to their correlations with user-chosen labels. Interacting with such a view, users can select attributes and evaluate the generated subset in a second view. We have also extended the applicability of the attribute analysis and selection task by providing an approach focused on attribute prediction based on logistic regression analysis.

1.2 Contributions

Among the main contributions of this Thesis, we highlight the development of a novel approach to analyze and select attributes, which include the Attribute-RadViz visualization. This interactive technique visually presents a panorama with multiple correlations of attributes related manly to target categorical attributes of interest. Since current data sets are full of relevant categorical attributes, each one revealing a particular phenomenon encoded in the data, users can examine distinct scenarios to gain insights into such data sets.

We also highlight the development of two fully operational web-based tools. The first one contemplates our attribute analysis and selection approach; whereby an interactive dual-view interface support users in selecting relevant attributes and then evaluating their capability in distinguishing patterns in the data. The second tool implements our approach to creating, evaluating, and applying both binary and multinomial logistic regression models.

As parallel achievements, we highlight the improvements for the RadViz technique, such as the arcs of the dimensional anchors (DAs), the greedy DAs sorting algorithm, and the novel interface model. We also highlight novel RadViz representations, which originally displays relationships by mapping items related to attributes; we include new designs showing labels as DAs and attributes as mapped elements, as well as labels as DAs and items as mapped elements.

These contributions have been documented mainly in two papers. The first one, called *A Novel Visual Approach for Enhanced Attribute Analysis and Selection* (ARTUR; MINGHIM, 2019), has been published in the Computer and Graphics Journal (Qualis A2[1]). The second, called *An Approach to Explore Logistic Regression Models and its Prediction Capabilities*, is, at the time of submission of this Thesis, being finished for submission.

1.3 Thesis Organization

This thesis is organized as follows:

- Chapter 2 presents the literature review centered on visual attribute analysis approaches;
- Chapter 3 describes our first approach focused on attribute analysis and selection;
- Chapter 4 presents our second approach, which explores logistic regression models aiming at the prediction of target attributes;
- Finally, in Chapter 5, we present the conclusions of this Thesis.

[1] Qualis is a combination of procedures used by the Coordination for the Improvement of Higher Education-Personnel of Brazil to stratify the quality of intellectual production in postgraduate programs. The classification of publications and events is carried out by each evaluation area and undergoes an annual update process. Strata levels classify the vehicles indicating quality, with the A1 being the highest one followed by; A2; B1; B2; B3; B4; B5; and finally C - with zero weight.

CHAPTER

2

RELATED WORK

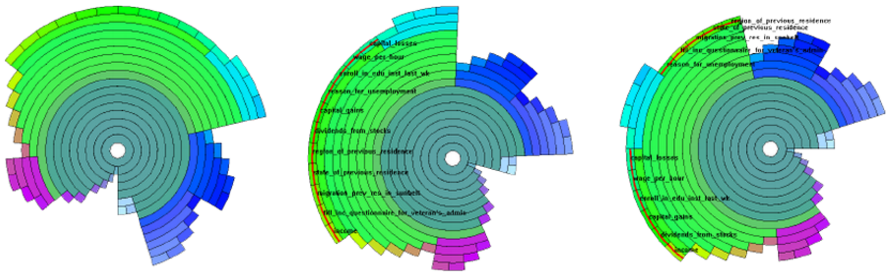
In this chapter, we present the previous works related to this Thesis focusing on approaches that visually support the attribute analysis task. Initially, we show techniques that interactively represent the attribute space by visual means, and in the following, we expose the methods which also support the selection task. Then we present approaches that promote the interactive construction of regression models by aggregating FS mechanisms.

2.1 Interactive Visual Exploration of Attribute Spaces

Traditional dimensionality reduction (DR) techniques transform multidimensional data into a meaningful representation with reduced dimensions amount. Several interactive approaches support DR tasks through the attribute subspaces exploration; therefore, users can find or combine relevant attributes. Yang *et al.* (2003) introduce the Visual Hierarchical Dimension Reduction (VHDR), an approach to handling and exploring multidimensional data. In VHDR, a hierarchical structure scheme arranges the dimensions, and a hierarchical radial visualization method, called InterRing (see Figure 1), displays the data. The central idea is to visualize data without losing inherent meaning by generating lower-dimensional spaces and allowing users to play an interactive role in the DR process by modifying the hierarchy and selecting interesting clusters. Cheng and Mueller (2016) propose an approach that provides visualization both from the point of view of items and from the attributes in the same layout. They employ four matrices; two similarity matrices (usually applied in isolation), one encoding the similarity between attributes, and another between items of the data. The remaining two matrices are assembled from the combination of the first two, thus forming the four sub-matrices required for the proposed multidimensional visualization model. The authors explain that the layout, called Data Context Map, allows observation of the relationships between items and attributes, as well as the relationship between themselves simultaneously.

Some approaches apply resources like operators or well-known DR techniques to find

Figure 1 – Example of subspaces exploration with the InterRing.

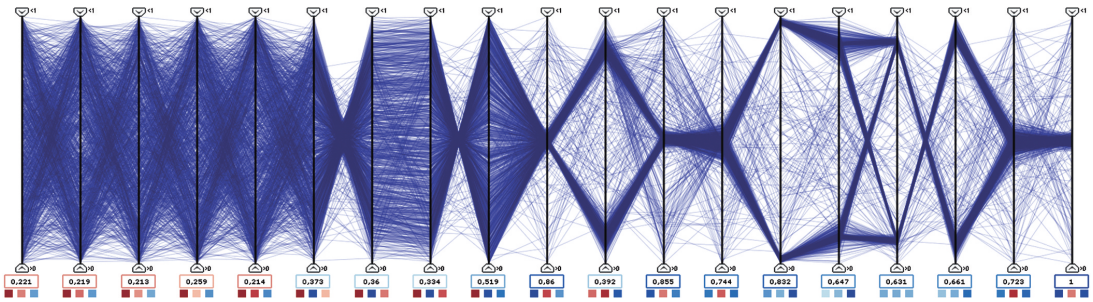


Source: Yang *et al.* (2003).

and evaluate subsets of attributes and usually show their results as two-dimensional scatter plots. Ingram *et al.* (2010) present the DimStiller, which allows users to proceed with analysis and reduction of dimensionality by creating and handling pipelines composed of operators capable of transforming data. Johansson and Johansson (2009) combine different user-defined quality metrics applying weight functions to reduce the data dimensionality, trying to preserve as many important structures within the original data set as possible. The authors aim to provide a quality-guided reduction of dimensions with a flexible user-controlled DR. Figure 2 shows through parallel coordinates the DR result in a synthetic data set from 100 to 18 attributes applying a combination of quality metrics. Choo *et al.* (2010) present the iVisClassifier, an interactive visual tool based on the Linear Discriminant Analysis (LDA). Users can interactively explore and understand the obtained dimensions by the LDA through views rendered with parallel coordinates and scatter plots. Additionally, a heat map shows an overview of the clusters' relationships regarding the pairwise distances between their centroids related to both original and obtained distances. Hence, iVisClassifier permits a user-driven classification by observing adjusted clusters, and also by mutual dimension filtering in the mentioned parallel coordinates and the scatter plot views.

Other works promote sequential subspaces navigation with some similarity criteria between adjacent ones to find representative subsets of attributes. Dy and Brodley (2000) introduce the Visual Feature Subset Selection using Expectation-Maximization Clustering (Visual-FSSEM), an interactive visual tool focused on selecting attributes in unsupervised data. Users can select any attributes subset as the starting point, perform a sequential exploratory search (forward or backward) on attribute subspaces, and visualize the results of the expectation-maximization clustering. Tatu *et al.* (2012) employ an interestingness-guided subspace search algorithm to find a candidate set of subspaces. Initially, an algorithm is used to identify a candidate set of interesting subspaces automatically. After that, a filtering step is employed to reduce the representations to a user-selectable amount. Lastly, a visual-interactive tool to explore the representations of subsets is provided to the user. DR can also be achieved by applying user-defined metrics and operators for data filtering. Other solutions try to expose insights through dynamic projections. The idea is to reveal information from the patterns transitions shown by the navigation between subspaces.

Figure 2 – A parallel coordinate plot showing a DR task result related to a synthetic data set with 100 attributes initially.



Source: Johansson and Johansson (2009).

In (LIU *et al.*, 2015), users can navigate between projection pairs animatedly through a transition view graph. Jäckle *et al.* (2017) present a framework that displays pairwise of projections linked by trails, making the transitions easily to compare. Additionally, the authors present a data-driven similarity measure for projections to group subspaces and avoid redundancy.

An alternative strategy to aid users in the subspace analysis task is providing mechanisms for finding combinations of attributes that reveal content, such as clusters or recurrent patterns. Seo and Shneiderman (2005) describe a conceptual framework where – according to a ranking generated from user-selected criteria – attributes are displayed graphically by several views. Features are presented in graphs, exposing their pairwise relationship, the intensity of the criterion value (by colors), a summary of the dimension distribution, amongst other information. Figure 3 shows the main interface of the framework. Tatu *et al.* (2011) also use rankings (by a specified user task) to show a potentially relevant set of visualizations for further interactive data analysis. The authors also present ranking quality measures for both class-based and non-class-based scatterplots and parallel coordinates visualizations. Zhou *et al.* (2016) present an approach that reconstructs new attributes by combining well-known multidimensional projections considering the preservation of interesting clusters. McKenna et al. McKenna *et al.* (2016) present the s-CorrPlot visualization, a highly scalable method to explore correlations in the attribute space in large data sets. Gleicher (2013) describes an approach to summary the data by creating projections functions that represent user-defined concepts.

These previously described approaches are very useful to reduce significant amounts of attributes interactively. As a matter of fact, they provide visual means to insert the user into DR processes. Still, they fail to provide insight into the relevance of isolated or combined attributes related to some phenomenon of interest, which is a relevant goal in our research. In general, these works try to find cluster structures and internal patterns exploring subspaces; therefore, they are not intended to use labels or target attributes when available. Our approach supports finding meaningful attributes and representative subspaces directly from the perspective of a target one to evaluate the choices for the prediction of such attributes. An example would be to define which sets of exam values (attribute subspaces) are related to a discharge condition from

Figure 3 – The main interface of the rank-by-feature framework.

Source: Seo and Shneiderman (2005).

the hospital (target attribute).

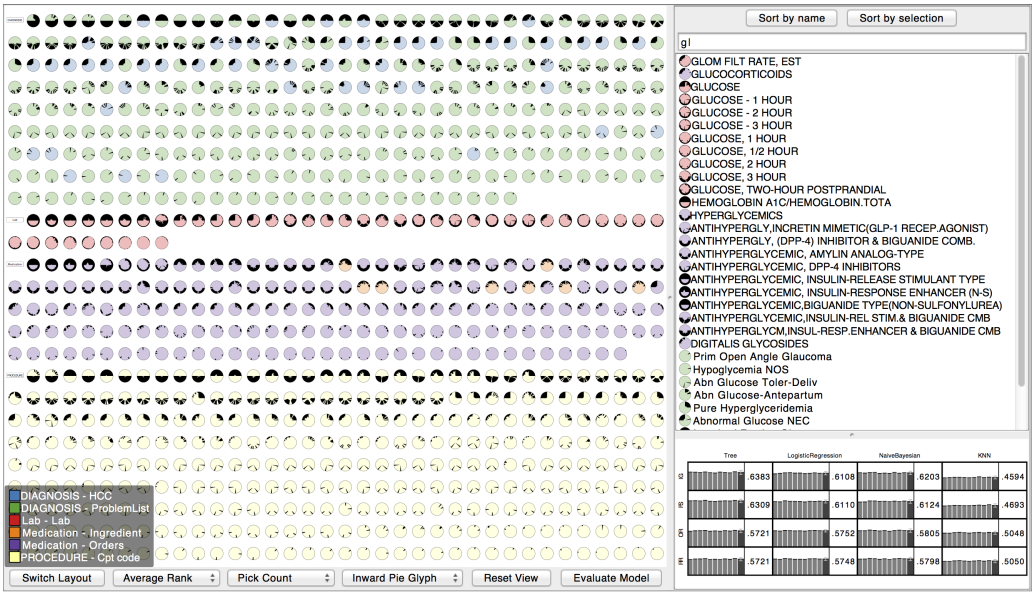
2.2 Visual Feature Selection

Several FS algorithms have been proposed to find relevant attribute subsets automatically. In general, they have successfully reached their goals. Nevertheless, they remove the analyst from this step, which can both reveal valuable insights and also take advantage of the user expertise in the process. Based on this, the visualization community has started to focus on the development of interactive FS methods.

To help users better understand how automated FS algorithms rank (or select) features, Krause, Perer and Bertini (2014) propose a visual analysis tool called INFUSE (INteractive FeatUre SElection). Its main screen initially has three components: feature view, list view, classifier view, and optionally, the interactive model builder, as shown in Figure 4. The first one shows attributes from sliced glyphs to represent the values obtained by FS algorithms. The second displays an ordered list of all dimensions according to some custom criterion. The third component shows the subsets evaluation by five classifiers. The customization of models for subsequent evaluation is the focus of the last component. Their tool presents an overview of automatic algorithms results, assisting the user in understanding how these methods work. Similarly to the DR methods, this approach is handy for finding and evaluating subsets of relevant attributes. However, it is complicated for analysts to understand the importance of attributes individually in relation to others, as well as in relation to potential data labels.

Since strong local correlations may be hidden between attributes in their global distribu-

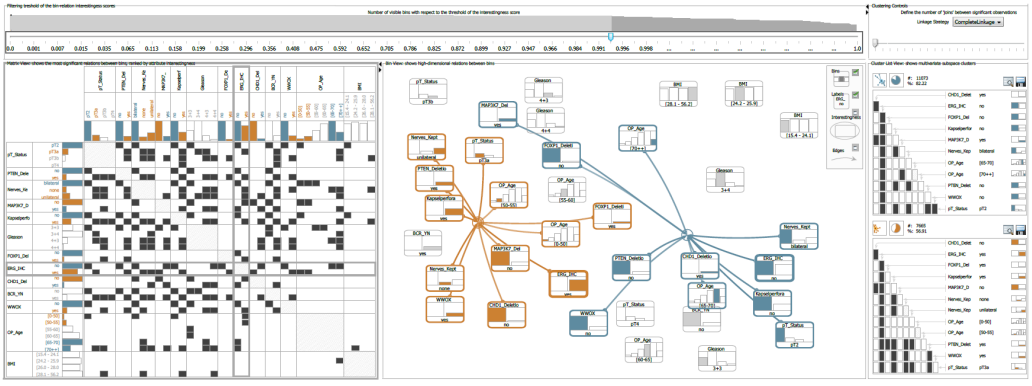
Figure 4 – The main interface of the INFUSE framework showing the feature view, the list view, and the classifier view.



Source: Krause, Perer and Bertini (2014).

tions, some approaches visually expose the relationship between partitions of the data. Bernard *et al.* (2014) present a tool that builds a relationship panorama between attributes and, more importantly, between their bins (see Figure 5). The tool is also able to find correlations between bins in mixed data sets. Another interactive FS approach is the SmartStripes, a technique proposed in (MAY *et al.*, 2011). It supports the investigation of interdependencies between different attributes and entity subsets defined by a selected one. A heat map shows the dependency intensity of attributes with each entity subset. Overall, these approaches have a high analytical capacity with the ability to find hidden relationships; However, in terms of FS, their application may be limited due to scalability issues, since they discretize attributes into several partitions, and the quantity of the attributes might be a limitation by itself.

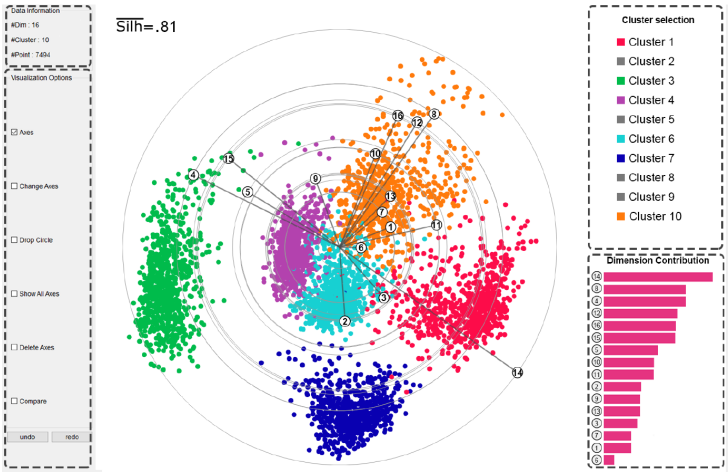
Figure 5 – The interactive prototype presented in (BERNARD *et al.*, 2014) that supports the exploration of interesting relations between aggregated bins of attributes in mixed data sets.



Source: Bernard *et al.* (2014).

Enhanced radial visualization techniques also have been applied to support FS tasks. Wang *et al.* (2017) propose an approach to formulate an optimal initial anchors disposition of a star coordinates visualization applying the concept of clusters and class separation from an LDA model (see Figure 6). The approach is useful for promoting weights to attributes. Users can handle a built linear discriminative star coordinates visualization in the investigation for the best separation between clusters or classes. Therefore, users can understand the influence of each attribute in the formation of such structures. Since users manage anchors as attributes in the main view, a limitation of this approach lies in the scalability of those manageable entities, when their number passes the hundred mark, the overlapping may cause the visualization and handling impracticable. The approach also suffers from redundancy problems, as it tends to give similar weights to correlated attributes and then misleading the FS tasks. Sanchez *et al.* (2018) present another radial solution, where attributes are exposed as improved scaled axes. Thus, analysts can interactively eliminate unimportant attributes since the axes represent the degree of importance of each attribute. However, backward FS strategies are not efficient for filtering large amounts of attributes; therefore, the approach may be coupled with another DR technique.

Figure 6 – The prototype interface of the FS based on linear discriminative star coordinates.

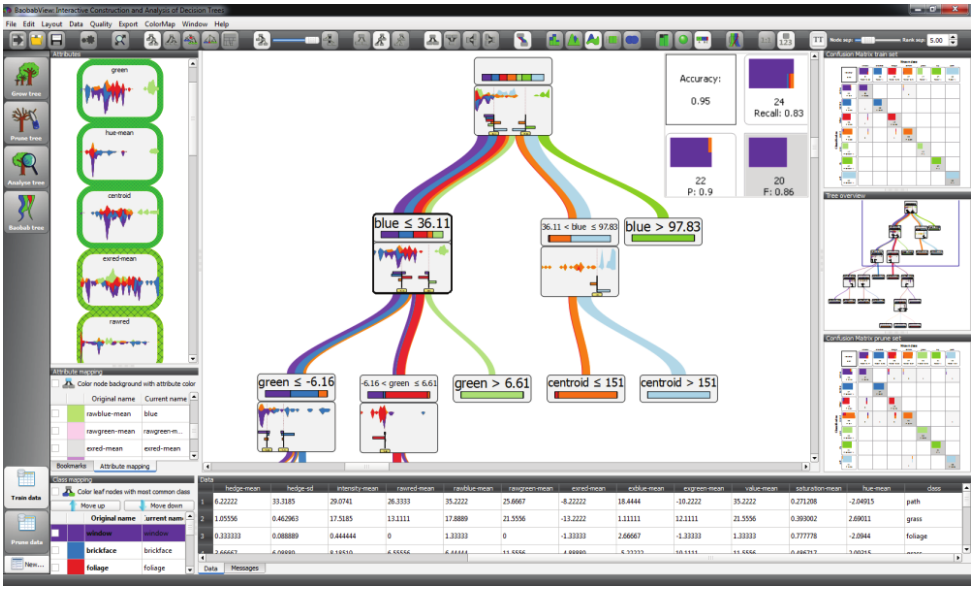


Source: Wang *et al.* (2017).

2.3 Dual-Visual Analysis

Aiming to provide a broader view in the exploration of data sets, many approaches promote the simultaneous visualization of the attribute space with another one representing data items (or their values). These provided views are often linked and allow for new possibilities in the interactive model of the analysis, as well as FS tasks. Turkay, Filzmoser and Hauser (2011) present an approach that displays an item view linked with another one showing the statistical properties of attributes. The dual-view promotes an interface that adopts the style *linking and brushing* to support the interactivity of one view, which subsequently updates the

Figure 7 – The multiple view interface of BaobabView, an interactive decision tree construction tool.



Source: Elzen and Wijk (2011).

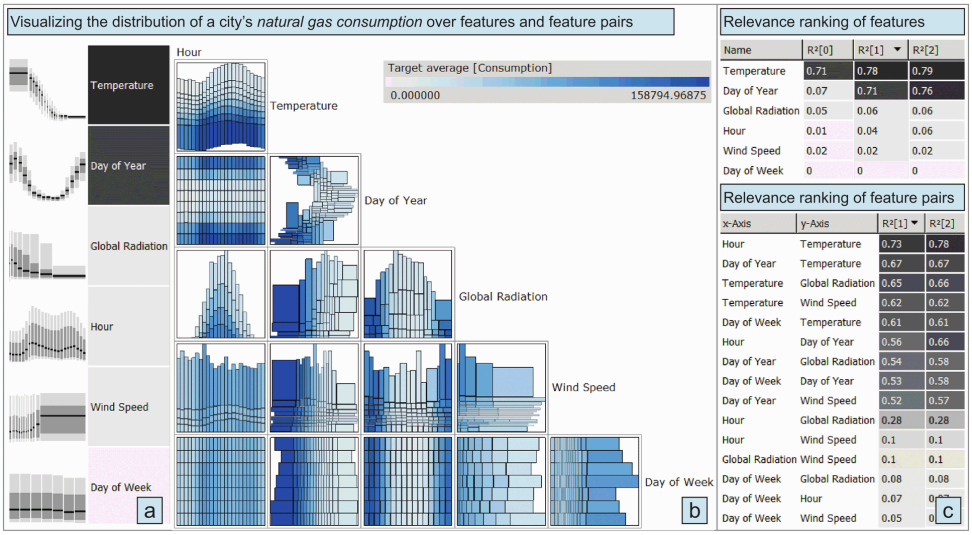
other in a *focus+context* fashion. Therefore, analysts can recognize, together, the structure of the attribute space, as well as the related distribution of data items. Later, the authors applied the dual-analysis for characterizing cancer subtypes in (TURKAY *et al.*, 2014). Similarly, Yuan *et al.* (2013) present a dimension projection tree/matrix visualization that simultaneously explores both attributes and items spaces. The user can perform the investigation by drilling down the data, restricting its range as well as pruning dimensions to examine different levels of the data. Another approach is the BaobabView (see Figure 7) proposed by Elzen and Wijk (2011), a multiple-view tool that allows users to build and analyze decision trees using their domain-specific knowledge.

Rauber, Falcao and Telea (2017) propose a projection-based visual analytics methodology to aid classification systems design coupled with the working tool. The tool displays multiple views, where users can perform FS tasks, generate projections according to the current FS, check the relevance of attributes concerning data labels, and so on. The approach aims at experienced users since they define the projections responsible for the exposure of attributes as well as data items, and this choice determines the cognitive map that will guide the FS tasks, as well as the learning process that may provide insights. Another aspect to be improved is the lack of resources to prune the data or to concentrate the analysis on specific subsets (such as mixed regions in the projection). This is important in FS methods to prevent relevant attributes with significant local correlations from being discarded.

2.4 Interactive Exploration of Regression Models

A variety of approaches combines visual analysis with regression modeling to investigate both the selection of representative attribute subsets as well as the quality of the generated

Figure 8 – The interface of a partition-based framework for building regression models presented in (MÜHLBACHER; PIRINGER, 2013).



Source: Mühlbacher and Piringer (2013).

models; this section presents state of the art related to interactive visual models for generating and exploring regression models. Mühlbacher and Piringer (2013) present an interactive framework for building regression models that aids the user in understanding the relationship of attributes related to a target one (see Figure 8). The tool provides two overviews, the first one showing the relationship of the target attribute to each attribute, and another exposing the relationship of the target attribute to each possible attribute pairs. Both display feature ranks to assist the user. A key point of the approach is the local approximation of the conditional target distribution by partitioning the feature domains into disjoint sections to enable a visual investigation of local patterns. Likewise, Klemm *et al.* (2016) present a regression analysis tool that exhaustively creates regression models between features and a target attribute. A three-dimensional heat map shows the results for the consequent user exploration. Users can also freely adjust the regression formulas.

A commonly used regression method for predicting an occurrence of a dichotomous event is the logistic regression (LR). Some visual tools apply LR models to check the prediction power of attributes and then formulate relevant attribute subsets. Zhang *et al.* (2016) present a visual analytics approach to multidimensional LR modeling for risk factor identification. The authors define three basic steps. In the first step, users perform FS tasks based on univariate indicators displayed by the tool. In the second step, users evaluate the relationships between the variables chosen in the first one to define good subsets for building the regression model. The last step deals with the evaluation of the regression models generated among the subsets chosen by the user. Similarly, Dingen *et al.* (2019) introduce the RegressionExplorer, a tool that allows users to find and evaluate subsets of attributes and then apply on regression models. A univariate analysis view shows individual attribute significance level, which aids the search for

proper combinations of relevant subsets of attributes for multivariate analysis.

LR analysis is widely used in many fields such as financial performance, consumer purchasing, ecology, medicine, epidemiology, and so forth (HARRELL, 2015), a fact that includes several non-experts machine learning users as analysts. Thus, the demand for user-friendly regression modeling tools persists where analysts should easily create, evaluate, and subsequently employ the generated regression models. It is precisely over this gap that we developed our approach for exploring regression models supported by interactive visualization techniques.

2.5 Final Remarks

In this chapter, we reviewed techniques that explore attribute space for a variety of purposes. The common ground between all of these works is some mechanism to reveal and investigate the attributes subspaces visually. Hence, analysts may gain insights and cooperate in the following tasks inserting their knowledge into the process.

We separated the review into four groups based mainly on the main goal of each work. Table 1 summarizes the main features of the reviewed works in this chapter. In the next chapter, we present our approach to analysis and selection of attributes based on correlation with visual support by radial technique. In the following chapter, we present another approach that aims the exploration of LR models, as well as the tooling for the later application of the generated models.

Table 1 – List of the reviewed work in this chapter and their main features.

	Interactivity			Learning mode		Data displaying							Major goal		
	Direct manipulation	Controls for subspaces exploration	Data focus or pruning	Clustering	Classification	Linked attribute and item views	Feature ranking	Point-based visualization	Attribute-to-item relationship	Attribute-to-attribute relationship	Attributes-to-label relationship	Categorical attributes	Dimensionality reduction	Feature selection	Regression modeling
Yang *et al.* (2003)	x	x	x	x						x			x		
Cheng and Mueller (2016)		x						x	x	x		x			
Ingram *et al.* (2010)		x	x					x		x			x		
Johansson and Johansson (2009)		x	x	x			x			x			x		
Choo *et al.* (2010)	x		x		x			x	x	x			x		
Dy and Brodley (2000)		x		x				x	x					x	
Tatu *et al.* (2012)		x	x	x				x	x	x					
Liu *et al.* (2015)	x	x		x				x							
Jäckle *et al.* (2017)		x						x		x					
Seo and Shneiderman (2005)		x	x			x	x	x	x	x					
Tatu *et al.* (2011)		x		x				x		x					
Zhou *et al.* (2016)	x	x		x		x		x		x					
McKenna *et al.* (2016)	x							x		x					
Gleicher (2013)	x			x	x						x			x	
Krause, Perer and Bertini (2014)		x			x		x							x	
Bernard *et al.* (2014)		x	x	x			x			x		x		x	
May *et al.* (2011)		x	x		x		x			x	x			x	
Wang *et al.* (2017)	x		x		x	x		x	x	x	x			x	
Sanchez *et al.* (2018)	x		x		x	x		x	x	x	x			x	
Turkay, Filzmoser and Hauser (2011)		x	x	x	x	x		x	x					x	
Yuan *et al.* (2013)		x	x	x		x		x	x	x				x	
Elzen and Wijk (2011)		x	x		x		x			x	x				
Rauber, Falcao and Telea (2017)		x			x	x	x	x	x	x	x			x	
Mühlbacher and Piringer (2013)		x			x		x			x	x			x	x
Klemm *et al.* (2016)		x	x		x			x		x	x				x
Zhang *et al.* (2016)		x	x		x		x			x	x				x
Dingen *et al.* (2019)		x	x		x		x			x	x			x	x

Source: Research data.

CHAPTER

3

INTERACTIVE ATTRIBUTE ANALYSIS AND SELECTION

This chapter is a modified version of the paper “A Novel Visual Approach for Enhanced Attribute Analysis and Selection”, presented in the 32nd Conference on Graphics, Patterns and Images (SIBGRAPI 2019) and published in the Elsevier Computers & Graphics Journal (ARTUR; MINGHIM, 2019).

As a consequence of the current capabilities of collecting and storing data, a data set of many attributes frequently reflects more than one phenomenon. Understanding the role of attribute subsets and their impact on the organization and structure of a data set under study is paramount to many exploratory and analytical tasks. Example applications range from medicine to financial markets, whereby one wishes to locate subsets of variables that impact the prediction of target categorical attributes. The user is essential in this context since automated techniques are not currently capable of embedding user knowledge in attribute selections. In this work, we propose an approach to deal with the analysis and selection of attributes in a data set based on three principles: firstly, we center the analysis of the relationships on categorical attributes or labels, because they usually summarize important state variables in the application; secondly, we express the relationship between target attributes and all others in the data set within a single visualization, providing understanding of a large number of correlations in the same visual frame; thirdly, we propose an interactive dual-visual approach whereby changes and selections in attribute space reflect visually on the configuration of data layouts, conceived to support immediate analysis of the impact of selected subsets of attributes in the organization of the data set. We validate our approach by means of a number of case studies, illustrating distinct scenarios of knowledge acquisition and feature selection.

3.1 Introduction

In most applications, data sets are created with a certain degree of uncertainty in the importance of different attributes and their impact in defining the object under analysis. Additionally, one single data set may encode more than one phenomenon that may be governed by distinct subsets of attributes. Other attribute related complexities can be observed in many real-life cases, such as redundancy, various degrees of relevance, and the need to reduce dimensions so as to take advantage of strategies that work better with fewer attributes as well as to reduce the size of the data set. It is, thus, of great importance to be able to select attributes and to analyze the impact of such selection. Although there are algorithms to perform attribute selection, they cannot currently replace the perspective of the user in finding relevant patterns and relationships.

Relevant information in a data set frequently comes from observations of interrelationships between certain attributes and other data components. For example, in medical research, analysts try to find the most relevant attributes correlated with causes of death, both to improve the quality of medical protocols and to predict survival. That parallel can be seen in most multidimensional analysis cases. In this context, a visual analysis approach for attribute space exploration is essential when hypotheses of correlations must be confirmed, but most importantly when previously unknown correlations are to be found. With proper tools at the disposal of users, new knowledge can be gained by understanding the relevance of attributes that may support other tasks, such as classification and clustering. The human insertion in the analysis and FS allows flexibility, creativity, and inclusion of tacit knowledge not possible by employing fully automatic methods (KEIM; MANSMANN; THOMAS, 2010).

Correlation analysis is a recurrent way of measuring complex relationships between variables. Traditionally, approaches that visually present such correlations expose attribute measures in pairs, regularly employing scatter plots or heat maps. However, hidden correlations between groups of attributes and partitions of other attributes (such as categories or labels) often remain hidden. In our approach we present the relationship between these labels or categories and the remaining attributes. The relationship between attributes is also indirectly revealed since similar attributes tend to correlate with the same labels.

We propose a correlation-based visual approach for data analysis focused on attribute space. The uniqueness of this work lies in the exposure not only of the relationship between attributes but also between them and each data label. We adapt the RadViz projection (HOFFMAN *et al.*, 1997; HOFFMAN; GRINSTEIN; PINKNEY, 1999), a technique originally built to express similarities between data items, for displaying attribute data similarity. That choice is due to its ease of interaction, speed of rendering, and the capability of showing groups of related attributes. The resulting technique, named Attribute-RadViz, allows the analyst to conduct explorations by interacting with a projection of attributes. A set of interactions and displays panels complete the analysis structure. We also provide a layout of the data set items by means of a data space projection allowing the observation of the impact of the selected attributes subsets

on the structure of the data. We demonstrate, through two case studies, how the attribute analysis afforded by the approach can lead to an improvement in data understanding and meaningful selection of subsets of attributes.

The main contributions of this approach are therefore:

- A correlation-based approach for feature analysis to identify relationships between all attributes and the available data labels or categories. We devise a specific correlation matrix for this purpose;
- Attribute-RadViz, an enhanced RadViz visualization that maps large amounts of attributes as elements with new visual and exploratory features to increase its analytic capabilities; and
- the description of two case studies; one example of usage for knowledge acquisition and hypotheses testing and another for attribute selection.

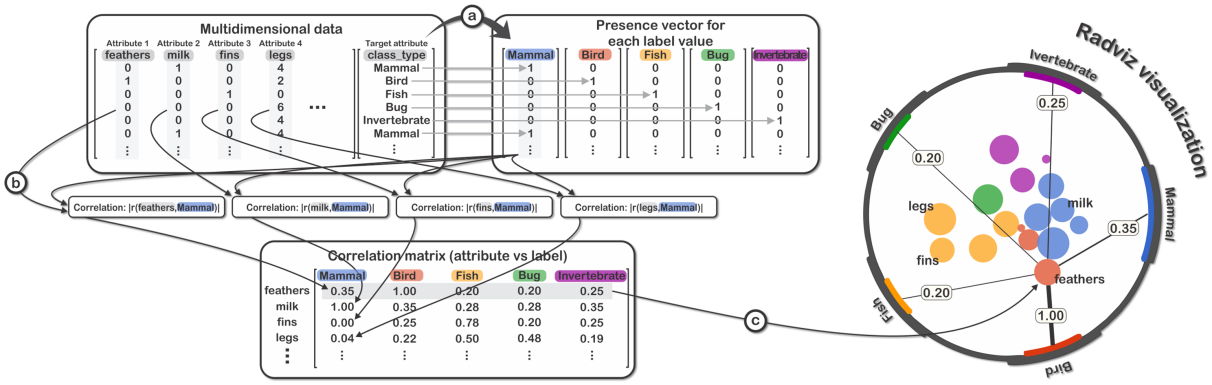
3.2 Overview of the Approach

The inspiration for this work is that, in many applications, data sets are full of categorical information, most of which are difficult to handle together with other attributes. But commonly, many of these attributes can be interpreted as labels and carry predictive potential if adequately targeted by classification models. To use as an example the health records data again, some attributes record different impacts of treatments or procedures; looking to other attributes from their perspective, it is possible to investigate how outcomes, such as the degree of severity or adverse reactions, are reached. This type of scenario occurs very often in weather prediction, environmental monitoring, epidemic control, and even in the prevention of undesired events, such as accidents, fraud, and market instability.

Our approach for attribute space exploration is named Attribute-RadViz, an improvement of the classic RadViz technique to make it capable of mapping attributes (instead of items) and highlight attribute correlation under the influence of a label set (or categorical attribute). The assembly of Attribute-RadViz is described in Figure 9 and detailed in Section 3.3.

A second view is provided to displays a multidimensional projection of the data set based on current attributes. Once the user finds interesting attributes towards some desired outcome, this second view is updated to consider only the selected attributes. For small to moderate quantities of selected attributes, the second view adopts a RadViz in its traditional construction. However, as the number of selected attributes grows, the user can change the visualization to the t-SNE technique, which is more scalable. The second view is described in Section 3.4.

Figure 9 – Graphical abstract of the Attribute-RadViz with the sample data set shown in Table 2 presenting the panorama of relationships between attributes and labels from the target attribute. (a) Decomposition of labels of target attribute into presence vectors (for each label). (b) Performing the calculation of correlation between data labels and other attributes. (c) Projection of the correlation matrix encoded into the RadViz visualization technique.



Source: Artur and Minghim (2019).

3.3 First View: Attribute-RadViz Layout

The classic RadViz is a radial visualization technique that presents attributes as points known as dimensional anchors, distributed initially equidistantly around the unit circle. The elements (data items) are mapped according to the approximation influence from each DA, similarly to a spring system. The attraction values for each DA are usually normalized, avoiding discrepancies in the positioning of the elements due to the different scales and ranges between attributes.

The reason why we choose RadViz is that its layout represents a map that helps the user to quickly find the objects of interest. This is due to its ability to jointly consider both aspects of a data matrix (rows and columns) and further rendering them in the final layout. Also, the RadViz computation is fast enough for implementation in an interactive context. For more detailed information about the RadViz and other radial visualization methods, the reader is referred to Draper, Livnat and Riesenfeld (2009) and Diehl, Beck and Burch (2010). Other interesting works regarding RadViz extensions are presented in (SHARKO; GRINSTEIN; MARX, 2008; ONO *et al.*, 2015; ZHOU *et al.*, 2015; CHENG; XU; MUELLER, 2017).

In Attribute-RadViz, the DAs are label values of a chosen categorical attribute. A correlation matrix gives the attraction forces from all other attributes to the DAs. We further made the correlation data to be encoded by element sizes, and the entire generated matrix coded on the same view and shown according to user's interaction. The search for specific correlations is also simplified because of the mentioned RadViz's simple mapping methodology.

As a practical example of the advantages of applying RadViz mapping to our correlation matrix, Figure 10 compares RadViz with well-known projection techniques in the task of

searching for strongly correlated attributes with the *"Venezuelan President"* label in the News data set. Inside the Attribute-RadViz, this task is straightforward; the user can examine the closest mapped attributes to the DA representing the desired label (see Figure 10b). The t-Distributed Stochastic Neighbor Embedding (t-SNE) (MAATEN; HINTON, 2008) clusters very well, but this does not make this task any easier; in this case, the user should investigate the clusters of interest to find the desired attributes (see Figure 10c). In the Principal Component Analysis (PCA) (JOLLIFFE, 2011), this task is even more challenging since the items mapping seems more complex (see Figure 10d).

However, RadViz approaches have limitations that need to be addressed. Firstly, the ambiguity of positions, where RadViz maps different items in the same location. Secondly, the overlapping, where RadViz maps several items in the same area. Finally, the DAs scalability issue, where their arrangement in the unit circle can become saturated. We minimize the first two drawbacks by the encoding of element sizes, the inclusion of item-to-item force layout adjustment, and smoothing the RadViz mapping adding to its equation a new parameter. Regarding the scalability problem, in our approach the DAs are label values in the attribute view, and its quantity is rarely more than 20 in most data sets. In the item view, DAs are the currently selected attributes; it is unusual for the user to select more than 20 without performing the pruning to refine the selection and start a new sub-selection. Figures 10a and 10b compare the traditional RadViz and our enhanced RadViz rendering 3,731 attributes.

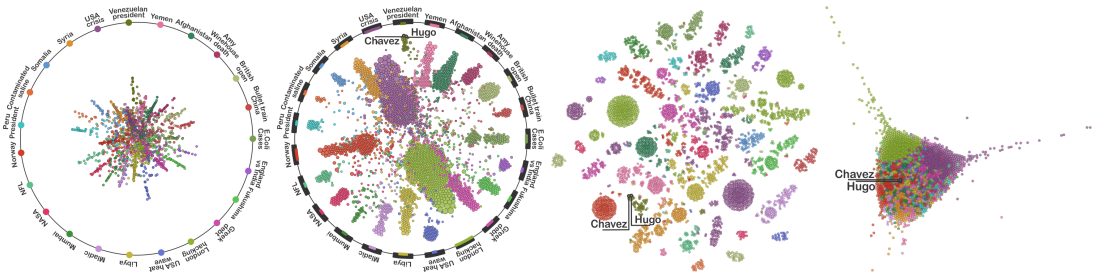
In the following, we present every step of the construction of Attribute-RadViz and some visual increments enhancing our RadViz. Firstly, we explain the correlation matrix construction and how we deal with categorical variables, and then we detail the visual encoding adopted in this approach. To demonstrate our methodology, we adopt a partition of the zoo data set available in UCI Machine Learning Repository (ASUNCION; NEWMAN, 2007). Originally, the data set has 18 attributes and 101 items. For our example, we reduce to 15 items, as presented in Table 2.

3.3.1 The Correlation Matrix

The core of Attribute-RadViz lies in the correlation matrix construction, which is different from most available methods. Given a target attribute, we want to quantify how related the remaining attributes are to each value inside the target. Conventional methods based on correlation compute the correlation coefficient for each pair of attributes and then assemble the matrix. We perform the correlation estimation between attributes and a boolean presence vector of each label value. Therefore, it is required to decompose the target attribute – which contains information for k label values – into k virtual presence vectors with binary values symbolizing the label presence for each item. Figures 9a and 9b illustrates the process.

To assemble the base matrix for subsequent visualization task, presence vectors are collected similarly *One-Against-All* (LIU; ZHENG, 2005) and *Binary Relevance* (TSOUMAKAS; KATAKIS, 2006) approaches. In them, each data label is represented by one vector where each

Figure 10 – Encoding the correlation matrix of the News data set containing 3,731 attributes in different visualization techniques. (a) The classic RadViz. (b) Our enhanced RadViz. (c) t-SNE Projection. (d) PCA Projection. Our Attribute-RadViz approach expresses a cognitive map where outer elements (close to the anchors) generally represent attributes with strong and exclusive correlations, and the internal ones commonly are shared (strong or not) correlations. For example, to find the two strongest correlated attributes with the label "Venezuelan President" (in this case "Hugo" and "Chavez"), the cognitive map provided by RadViz is of great help compared with the other projections, as we can see when highlighting the attributes in (b), (c), and (d).



(a) (b) (c) (d)

Source: Adapted from Artur and Minghim (2019).

element records the value one on the items where the label occurs. Their goal is to enable the classification of multi-label problems in separated stages by applying existing single-label solutions. In our approach, the presence vectors are instruments to measure the correlation intensity and direction between attributes and data labels.

Let L be the finite set of k possible labels values $L = \{l_1, l_2, ..., l_k\}$ and y_l be the target variable, which records the label information of a m-dimensional data set with n items x. To enable correlation calculation of labels individually, y_l is decomposed into k presence vectors $\{z_1, z_2, ..., z_k\}$, where for every $l \in L$, $z_i = 1$ if $y_{li} = l$ and $y_{li} = 0$ otherwise.

Pearson's correlation coefficient (TABACHNICK; FIDELL, 2006) is calculated between each attribute against the boolean presence vectors. Pearson's correlation is among the most popular metric to quantify the relationship between two variables. The Equation 3.1 obtains the coefficient, where x and y are vectors of the same size, $\bar{x}$ and $\bar{y}$ are their arithmetic means.

$$r(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}. \tag{3.1}$$

The resulting values range from -1 to +1, where the signal implies the direction of the relationship and the magnitude is related to the intensity of the correlation. The final matrix has dimensions $m \times k$.

Table 2 – Sample data set extracted from the zoo data set available in UCI Machine Learning Repository.

animal_name	feathers	milk	fins	legs	hair	eggs	airborne	aquatic	predator	toothed	backbone	breathes	venomous	tail	domestic	catsize	class_type
goat	0	1	0	4	1	0	0	0	0	1	1	1	0	1	1	1	Mammal
chicken	1	0	0	2	0	1	1	0	0	0	1	1	0	1	1	0	Bird
piranha	0	0	1	0	0	1	0	1	1	1	1	0	0	1	0	0	Fish
gnat	0	0	0	6	0	1	1	0	0	0	0	1	0	0	0	0	Bug
crab	0	0	0	4	0	1	0	1	1	0	0	0	0	0	0	0	Invertebrate
elephant	0	1	0	4	1	0	0	0	0	1	1	1	0	1	0	1	Mammal
bear	0	1	0	4	1	0	0	0	1	1	1	1	0	0	0	1	Mammal
clam	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	Invertebrate
dolphin	0	1	1	0	0	0	0	1	1	1	1	1	0	1	0	1	Mammal
penguin	1	0	0	2	0	1	0	1	1	0	1	1	0	1	0	1	Bird
leopard	0	1	0	4	1	0	0	0	1	1	1	1	0	1	0	1	Mammal
duck	1	0	0	2	0	1	1	1	0	0	1	1	0	1	0	0	Bird
seahorse	0	0	1	0	0	1	0	1	0	1	1	0	0	1	0	0	Fish
flea	0	0	0	6	0	1	0	0	0	0	0	1	0	0	0	0	Bug
scorpion	0	0	0	8	0	0	0	0	1	0	0	1	1	1	0	0	Invertebrate

Source: Asuncion and Newman (2007).

3.3.2 Handling Categorical Variables

Although Pearson's correlation remains an excellent way to quantify relationships between numerical variables, we still need some mechanism to build the correlation matrix with numerical and categorical data. An alternative solution is to convert each level (categorical value) into a numerical value and then apply the Pearson's correlation; however, the choice of those values influences the resulting coefficient value directly, since order and distance information is arbitrarily chosen. A strategy to solve this problem may be applying regression models, as the multinomial LR. Nevertheless, solving regression models for large amounts of data in interactive approaches may become too expensive.

Zhang et al. (ZHANG *et al.*, 2015) present a solution, adopted here, that avoids solving the entire model to find the transformation. They attempt to maximize the *coefficient of determination* r^2, and therefore the correlation coefficient r. Since $r^2 = 1 - RSS/TSS$, being RSS (*residual sum of squares*) the data variance unexplained by the regression model and TSS (*total sum of squares*) the sum of squared differences of the dependent variable from the overall mean. Then, minimization of RSS maximizes r^2.

They want to minimize RSS, being $RSS = \sum(y_i - \hat{y})^2$, where $\hat{y}$ is the predicted value of y given x. The Equation 3.2 describes RSS' which is the RSS of the desired transformation. Basically, it computes the RSS for each categorical value against every numerical value of the other variable that falls on it, where v_c is the categorical variable, v_n is the numerical variable, n is the number of points, m is the number of levels of v_c, m^i is the number of points related to the categorical value $v_c(i)$ and $v_n^i(j)$ representing the j^{th} numerical value that falls on the categorical level $v_c(i)$. So, they want to find numerical values $v'_c(i)$ to replace each categorical level $v_c(i)$

that maximize r.

$$RSS' = \sum_{i=1}^{m} \sum_{j=1}^{m^i} (v_n^i(j) - v'_c(i))^2. \tag{3.2}$$

By making $\mu(v_n^i)$ as the mean of all numeric values that fall under the categorical value $v_c(i)$ (the entire manipulation of the Equation 3.2 that allows this transformation is described in (ZHANG *et al.*, 2015)), the authors arrive at the expression 3.3, which must be minimized:

$$\sum_{i=1}^{m} \sum_{j=1}^{m^i} (\mu(v_n^i) - v'_c(i))^2. \tag{3.3}$$

This way, minimization occurs when $\mu(v_n^i) = v'_c(i)$. Thus, the values that replace each categorical level are computed from the averages of numeric values that affect them. This model of categorical data handling is efficient, given the need to compute only a set of means, and it confers good results in terms of cost and benefit. The disadvantage of this solution is that the categorical target attribute must have potentially different levels of ordering and distance measures associated with each other attribute. However, concerning our approach, this problem is mitigated, since our target attribute is decomposed into presence vectors that are essentially numerical, and then the correlation is performed between those presence vectors against the rest. Hence, this solution is suitable and brings fast results to our interactive approach.

3.3.3 Visual Encoding

A favorable arrangement of the visual elements is important for the correct perception of the information being coded (MACKINLAY, 1986). The various entities of the visualization can be managed to increase the interpretability of the data while at the same time minimizing the possible problems of the RadViz visualization. Here we present our visual encoding including the interactive mechanisms and widgets.

- **Elements position.** The positioning of the elements represents immediate hints of how and where users can find correlations of interest. The RadViz methodology defines the elements' mapping, and its application is straightforward since we are dealing with homogeneous data (correlation coefficients) in the same range of values (absolute correlation values between 0 and 1). The normalization step of the attributes made by the classic RadViz becomes unnecessary.

  Let us illustrate the positioning with an example. When constructing the correlation matrix in our sample data set (Table 2), we can observe that the attribute "feathers" has a powerful correlation with the "Bird" label value and mild correlation with the others. When applying the RadViz mapping, the attribute is placed next to the bird-DA since this anchor exerts

the most significant attraction force among all DAs, as shown in Figure 9. Therefore, the positioning scheme constructs a cognitive map that aids user exploration in the search for label targeted correlations.

Additionally, the user can adjust a force-directed layout approach (KOBOUROV, 2012) on the RadViz mapping intended to reduce overlapping. Each element exerts a user-selected repulsion force proportional to the size of its radius to the other elements (simulating a circle to circle collision detection). Also, we made a slight change in the original RadViz equation to make the elements mapping more flexible,

$$P_i = \sum_{j}^{m} \frac{a_{i,j}^{s}}{\sum_{j}^{m} a_{i,j}^{s}} v_j, \tag{3.4}$$

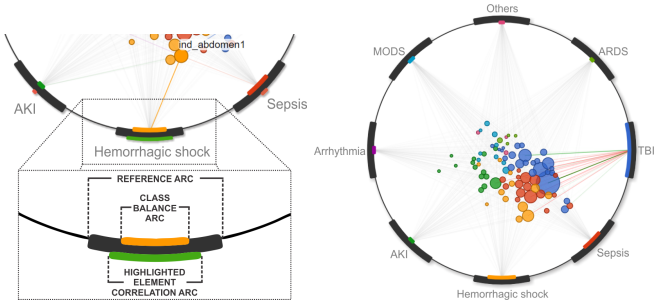
where P_i is the mapped point for the item i, v_j is the anchor j, a_{ij} is the item i in the j^{th} dimension among the m total dimensions. We included s as a user-defined parameter that changes the distribution of the values in the RadViz mapping; Figures 10a and 10b show how modifying these parameters can be useful for a better presentation of the attributes in the News data set, where the s value is 1 and the repel force is 0 in Figure 10a and the s value is 2 and the repel force is 1 in Figure 10b. The appropriate adjustment of these parameters is mainly dependent on the overlapping level of the rendering, which is generally proportional to the scale of the data sets.

- **Dimensional Anchors Ordering.** The order and arrangement of dimensions are essential for the effectiveness of several visualization techniques as it has a significant impact on the expressiveness of the visualization (ANKERST; BERCHTOLD; KEIM, 1998). In RadViz, it is not different, and we provide a greedy ordering algorithm to organize and consequently avoid or reduce problems clutering.

  The algorithm itself is quite simple and its pseudocode is outlined in Appendix A. A representative data item is chosen (the medoid) for each label value, and a pair to pair correlation test is performed between them. Label values with high positive correlation are placed in close slots (adjacent preferably) and those with high negative correlation in distant slots (opposite preferably), improving the distribution of the mapped attributes in that view. The algorithm is similar to the one applied in the data view with the traditional RadViz, and we give more details in the next subsection.

- **Colors.** In the attribute view, the colors of the elements symbolize their most correlated labels. For example, toward the sample data set shown in Figure 9c, the “feather” attribute has a stronger correlation with the “Bird” label; therefore, it assumes the corresponding label color. The same occurs with “milk” for “Mammals” (both are blue), or “fins” for “Fish” (both are yellow). In the data view, the color of the elements indicates their actual labels.

Figure 11 – Illustration of interaction when hovering the pointer over (a) the attribute "ind_abdomen1" and (b) the TBI-DA while investigating the data set. It is noticeable that "Hemorrhagic shock" is the most correlated label with the attribute "ind_abdomen1". In detail, interpretation of the arcs representing the DA; the inner arc refers to the balance of the data set; the outer arc denotes relevance for the attribute pointed at, and the middle one is a reference for both. When the pointer is hovering a DA (b), lines show the correlation direction for each attribute (green as positive and red as negative) whilst the size of each attribute encodes its correlation value for that DA-label.



(a) (b)

Source: Artur and Minghim (2019).

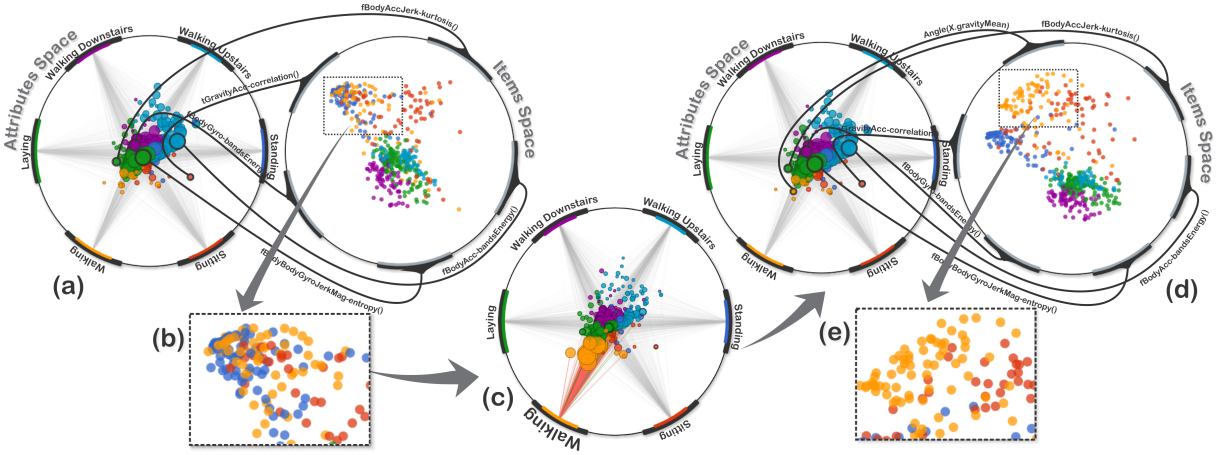
- **Elements' size.** As mentioned previously, the positioning of the elements gives an initial overview of the values present in the correlation matrix. However, the size of the elements also encodes the correlation matrix, and throught this property, users can investigate correlations more deeply.

  The initial sizes of the mapped elements encode the correlations of the attributes against the whole set of labels. Thus, large elements may imply good candidates for prediction purposes in the context of the entire data set. In contrast, small elements initially imply low predictive power for the whole data set; however, nothing can be stated about particular labels since strong correlations can be hidden, especially when considering poorly populated labels.

  In our sample data set, as shown in Figure 9c, the largest initial element is the "legs" attribute. This is because "legs" is the attribute of greater prediction power in the context associated with all the labels. However, if users are interested in strong correlations specifically with the "Fish" label, they can hover the pointer over the Fish-DA and all sizes of elements start to encode the correlations related to the 'Fish"label. Hence, the "fins" attribute will become the largest element of the view momentarily.

- **Arcs of Dimensional Anchors.** Another visual widget of our work is the set of arcs of DAs. Arcs for each DA are rendered in layers close to the model's unit circle; the inner ones represent the proportion of elements marked by the respective label value (thus data set balance regarding that label), and the outer ones represent the correlation amount for the attribute under the pointer. Figure 11a shows the arcs in detail when hovering an

Figure 12 – Improving discrimination in the Human Activity Recognition data set. (a) A dual-view containing the attribute space visualization as it relates to six label values (Standing, Sitting, Laying, Walking, Walking Downstairs and Walking Upstairs) and the data space view, where five attributes are selected and consequently influence the mapping of data items. (b) The user recognizes an area with a mixture of labels, the majority belonging to the yellow category "Walking". (c) In an attempt to improve separability between categories, the user hovers the pointer over the DA corresponding to "Walking" and investigates its relationships to all attributes. (d) An attribute strongly correlated to the label is found and selected becoming part of the data space view. (e) By the insertion of the new attribute, there is a considerable improvement in the visual organization and segregation of categories.



Source: Adapted from Artur and Minghim (2019).

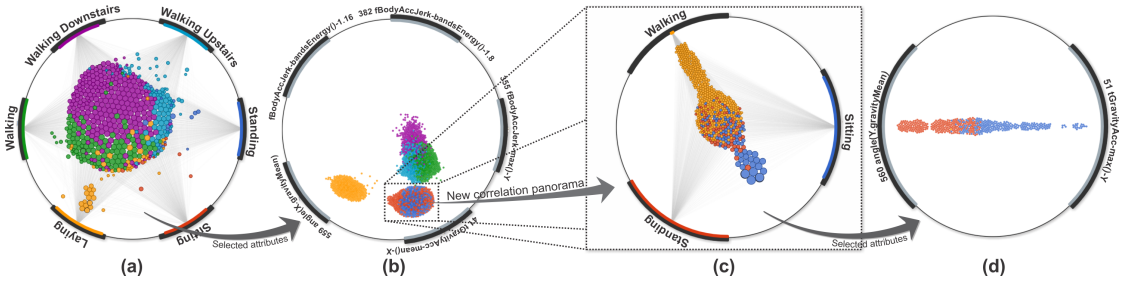
element. The label "TBI" (blue inner arc) has twice as many items concerning "Sepsis" (red inner arc) as shown in Figure 11b.

3.4 Second View: Visualization in Data Space

Once the approach presents a panorama of correlations between attributes and labels, a second linked view is provided to promote the analysis task jointly with the evaluation of current FS. For that purpose, users can choose the conventional RadViz or the t-SNE projection. Hence, users can adjust the model by adding or removing attributes and looking for useful insights. Additionally, they can request values of the silhouette coefficient as a feedback metric for evaluating the selected subspace. The silhouette is a prevalent measure to define class-based segregation. The approach returns three silhouette values: the original space, the selected space, and the mapped two-dimensional space coefficients.

Similarly to the first view, the linked RadViz from the second view also has its DAs ordered to avoid clutter. The algorithm is greedy, and it is somewhat similar to the DA positioning in attribute view. Initially, a pairwise attributes correlation matrix is built, and then the highest absolute value is picked and placed in close slots on the RadViz unit circle if they have a positive correlation; otherwise, they will be positioned in the most distant slots. From there, the algorithm chooses the highest correlated attribute in relation to those that have already been inserted and

Figure 13 – Pruning the data to focus on mixed and potentially hard to segregate areas. (a) Strongest globally correlated attributes selected in an FS task. (b) Pruning the data in an attempt to find correlated attributes to the picked subset. (c) A new attribute view is generated with a recalculated correlation matrix. (d) New useful attributes found and selected.



Source: Elaborated by the author.

places it in the nearest or farthest free slot, according to their correlation direction. In any case, the user can change positions of DAs at will.

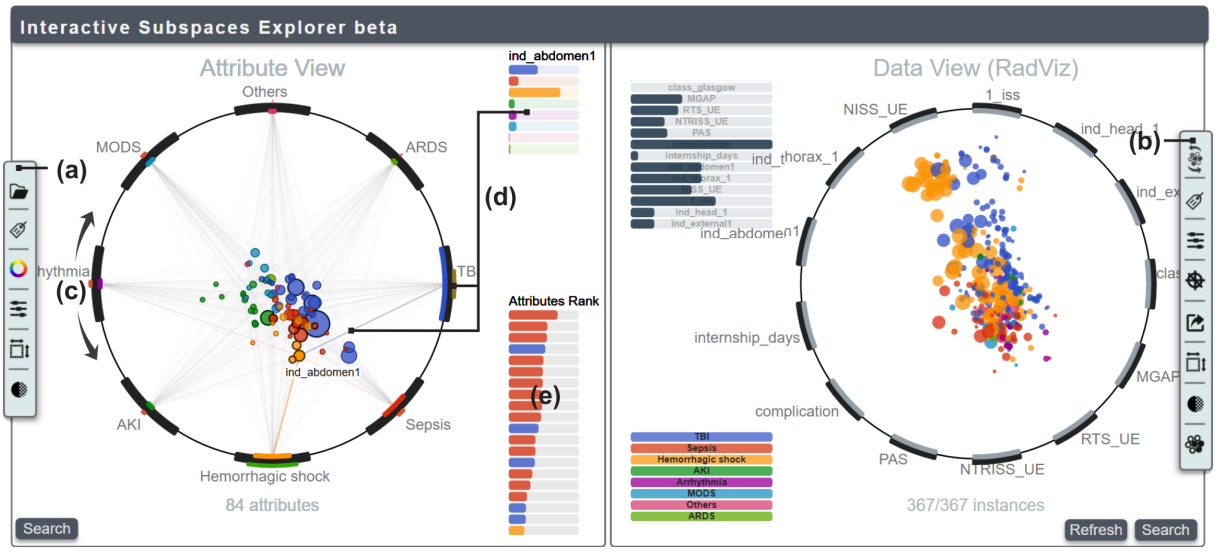
Figure 12 illustrates a scenario where the user evaluates and adjusts the FS aiming at the segregation of categories as well as reducing clutter. After a preliminary FS (see Figure 12a), the user examines the item space and realizes an area with multiple mixed items (see Figure 12b), most of them as yellow items (“Walking” label). Then, he or she returns to the attribute space and investigates attributes correlated with the yellow label searching potential attributes to segregate the “walking” cases (see Figure 12c). After locating and selecting an attribute (see Figure 12d), you can see the improvement in the label separation as well as the visual organization (see Figure 12e).

Another essential feature made possible by the dual-view is the FS refining mechanism through the pruning of elements and correlation matrix reconstruction. The primary goal of this functionality is to find locally correlated attributes, especially in scenarios where categories are difficult to segregate in the global analysis. In general, such situations are frequent when the data set has very similar categories. Hence, when users prune the data set, a parallel analysis is initiated with a new panorama of correlations containing possibly useful attributes for the segregation of the picked subset of elements. Figure 13 presents a scenario where the user applies the pruning mechanism to segregate two mixed categories (“Standing” and “Sitting”).

3.5 Prototype Implementation

In addition to the design elements described above, we have developed a tool with interactive resources to improve the user’s analytical process. Here we present a brief description of this implementation and its interactive functionality. Our implementation is a simple and portable web-based tool. It is developed based on HTML and Javascript languages combined with the visualization framework D3js (<https://d3js.org/>). The tool is made freely available at

Figure 14 – The prototype of a visual analysis tool implementing our approach. (a) Control panel for Attribute-RadViz. (b) Control panel for data space view. (c) Users can freely manipulate the DAs. (d) By hovering over an element, the correlation information can be observed from the information bars, arcs of DAs, or by the opacity of the influence lines. (e) The dynamic attributes rank shows the most correlated attributes for the last hovered DA-label. The rank helps users perceive the highest correlations when it is difficult to distinguish element sizes while interacting with DA-labels.



Source: Artur and Minghim (2019).

<https://github.com/erasmoartur/attribute-radviz> together with its manual and the sample data sets. Figure 14 shows the interface adopted in our prototype.

When starting the tool, the user must open the CSV data file and then select a target attribute. This allows the first view to render (attribute view). Some visualization parameters are configurable by the control panels (see Figures 14a and 14b). Users can enable or disable information bars, enable or disable element borders, adjust transparency levels, adjust the proportionality of element sizes, define the number of attributes simultaneously selected in the multi-select mechanism, and define the repulsion force intensity between the elements as well as force intensity between DAs to elements.

A variety of actions is possible inside the attribute view. In the search for patterns, the analyst can freely manipulate the DAs (see Figure 14c). By hovering the pointer over DAs, the correlation data between attributes and the current DA-label is encoded in element sizes, returning to regular sizes when removing the pointer (see Figures 12a and 12c). Correlation information between a particular attribute and all data labels is exposed by hovering the pointer over this attribute inside the attribute view. This information arises from arcs of DAs, information bars and influence lines (see Figure 14d). Additionally, the user can remove label values by dragging out DAs. Thus, the correlation matrix will be recalculated containing only the remaining label values. This mechanism is particularly useful when users notice some already segregated label

value (observing the second view); hence, users could remove this label to focus on interesting attributes of the remaining label values.

If the user wants to select a large number of attributes, he or she can use two distinct multi-selection mechanisms. The first one is the bounding box, which allows multiple selections inside the unit circle. The second is the multi-select click, where the user right-clicks on a DA-label and the P strongest correlated attributes (not yet selected) are included. The P value is defined by users in the control panel.

To start the second view, the user must choose an attribute in the right panel (see Figure 14b) to identify the items. Users can choose the visualization method between RadViz and t-SNE. Selecting RadViz, when hovering the pointer over the elements, arcs and information bars expose the actual data values of the selected item proportionally. Taking our sample data set again as an example, by hovering over the "chicken" element in the data view the arc over "eggs" DA becomes fully filled demonstrating the value of this element for that anchor is maximum. By hovering the pointer over DAs, the values of the attribute represented by that DA encodes new sizes of elements in the projection. It is possible to do this also by hovering over attributes in the attribute view, providing coordination between views during analysis. For example, by hovering over the "legs" attribute in the attribute view, the item "scorpion" assume the maximum size in the second view, while items "piranha" and "seahorse" become only dots with minimum size. This could give the user a sense of how each attribute affects the labels and items. Additionally, every attribute selected in the attribute view or in the list of attributes is added to the second view.

3.6 Case Studies

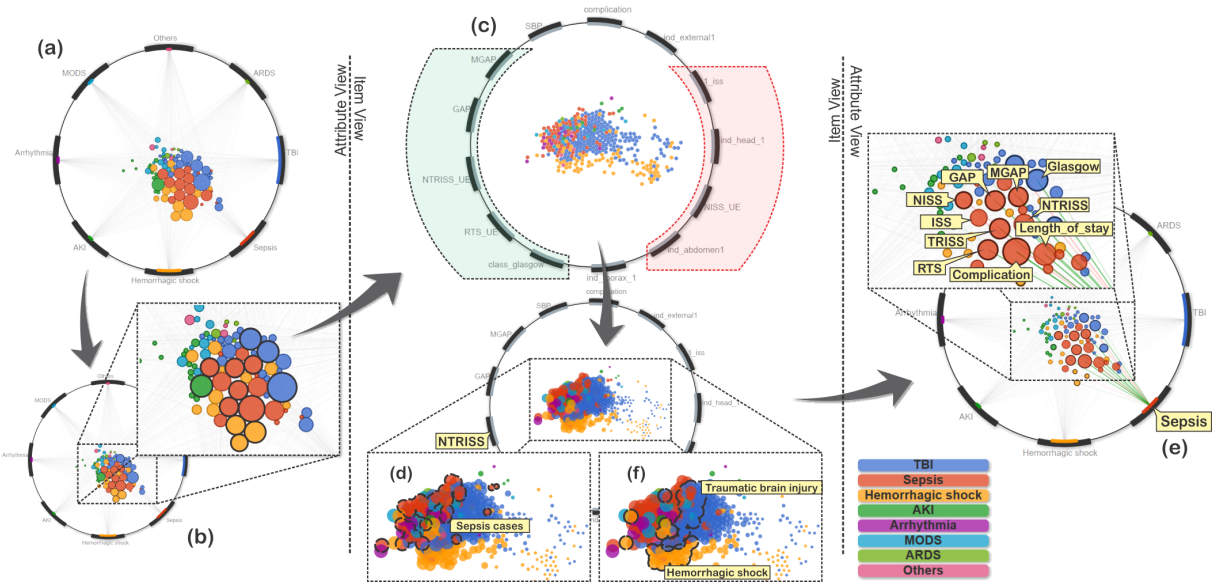
We present two case studies addressing different aspects of our approach. The first examines the analytical context, where the combined exploration between attribute and data views aids the user to formulate and test new hypotheses. The second case study investigates the ability of this approach to support finding representative subsets of attributes and compares the results with automated methods.

3.6.1 Case One: Understanding the Predictive Power of Attributes

Trauma scores represent an attempt to characterize and document traumatic injuries – as much as possible – by their severity levels. They are essentially mathematical or statistical values, quantified by numerical scores that vary according to intensity and types of trauma injuries. The estimation of a trauma score is carried out from the analysis of anatomical and physiological parameters (JÚNIOR *et al.*, 1999). These scores are mainly useful for evaluating the quality of trauma care and providing metrics for comparison for inter-hospital care protocols.

The data set for this case study comes from a collection of trauma information collected from the University Hospital (Hospital das Clínicas) of the Medical College of Ribeirão Preto,

Figure 15 – An analysis of health records data. This investigation considers only trauma cases with death outcome. The analyst aims to understand the behavior of trauma scores and how to improve their prediction accuracy. (a) Initial Attribute View. (b) The FS is performed; highlighted items represent the selected attributes. (c) Projection of data items considering the selected attributes. The DAs are rearranged to delineate a proper mental cognitive map. The red highlighted ones mean attributes for which high values imply in low survival probabilities. The green group describes the scores for which high values indicate a good survival probability. (d) By hovering the pointer over the "NTRISS" DA, the elements in the data space view assume sizes proportional to the actual attribute values (as explained in Section 3.5). The sepsis cases are among the ones with high recovery probability (large elements are approximately 100% of recovery probability cases while small elements (dots) are cases close to 0% of recovery probability), which implies affirming that the score has difficulties when predicting cases of this condition, since cases here resulted in death. (e) The analyst returns to the attribute space to investigate whether there are candidates correlated attributes with sepsis cases to adjust the trauma score. The attributes most correlated with "Sepsis" are length_of_stay (in days) and complication; both are dependent on the patient's evolution, not immediately available. Thus, to improve the score, it is necessary to model a dynamic solution, with monitoring of certain variables added to a time-varying score. Then, the analyst can proceed to investigate the reasons for the poor prediction for the cases highlighted in (f) to "hemorrhagic shock" and "traumatic brain injury".



Source: Adapted from Artur and Minghim (2019).

University of São Paulo (HC-FMRP-USP) for nine years (2006 to 2014). The data set has 21,294 records with 145 attributes. It includes patient profile data, trauma event information, clinical tests and observations, and calculated trauma scores.

We describe part of the analysis of trauma scores performed in the medical data set. Analysts attempt to adjust the trauma scores models to increase their efficiency in the characterization of patient care and their prediction capabilities. Aided by Attribute-RadViz, we have shown some gaps that can be explored to develop more descriptive scores.

This analysis considers only death cases of the data set, a total of 904 incidents in the nine years of collection. Due to the absence of some attribute values, the number of items is reduced to 367. We chose as target the attribute "death causes", the label values inside it are *Multiple Organ Dysfunction Syndrome (MODS)*, *Traumatic Brain Injury (TBI)*, *Acute Kidney Injury (AKI)*, *Hemorrhagic Shock*, *Sepsis*, *Arrhythmia*, *Acute Respiratory Distress Syndrome (ARDS)*, and *others*.

Attribute-RadViz initially displays the attribute space projection (see Figure 15a). Without any interaction, it is possible to make some observations. The most populous class label is the "TBI", followed by "Hemorrhagic Shock" and "Sepsis". The attribute of highest correlation with the whole class labels is "ind_head1", which is a measure that determines the damage level in the head region, including the face.

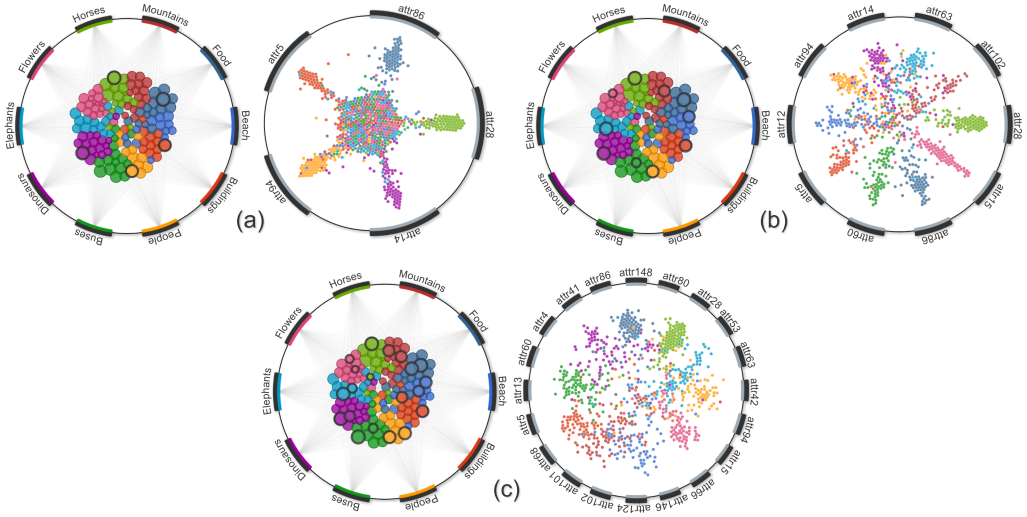
Making use of the available interactions, we perform a FS after examining label by label, trying to keep the choice as balanced as possible, as explained in subsection 3.5. This task is possible by investigating the DAs hovering the pointer over them. Figure 15b shows a selection of 13 attributes highlighting the selected ones.

Moving to data space (Figure15c), we organize the DAs to place the attributes whose high values indicate low survival probability in the right side. The attributes for which high values imply in a good recovery probability are arranged on the left side (as shown in Figure 15c). This attribute layout sets up a cognitive map that allows us to identify patterns but also separates items with different labels relatively well. In fact, the sorting algorithm organized the data similarly, but we rearranged the DAs to place semantic groups in the left and right sides.

As previously explained, when hovering over DAs in attribute view, it shows correlations between attributes to the current DA-label; but in data space, hovering DAs makes the size of elements proportional to the actual value of the current DA-attribute. The highest correlated score associated with death cases in this data set is *New Trauma and Injury Severity Score (NTRISS)* (DOMINGUES *et al.*, 2011). When we check their values (see Figure 15d), we observe that several cases have a good recovery estimation; however, all items in this context represent death cases, which opens a gap for an investigation over this inconsistency.

When examining the "NTRISS" values, the interesting point is that most cases of sepsis have estimation scores values indicating high recovery probability. Therefore, we may conclude that the "NTRISS" score is not a good predictor when sepsis cases occur. A question then emerges; is there any way to adjust the score to allow detection of such cases? The answer naturally is in attribute space. If we find an attribute with a strong correlation with that label value, it can then become part of the score modeling. Figure 15e shows the attributes exposed when hovering over the Sepsis's DA, and, besides other scores – which are not useful in this case – only two attributes have a good correlation level: "length_of_stay(in days)" and "complication". These two attributes are not instantly acquired, being determined only at the end of the process. That fact limits us from trying some new adjustment. Hence, this analysis has returned a recommendation

Figure 16 – Steps performing a FS in the Corel data set. The graphs are displayed in pairs, the left ones are the attribute views and the right ones are denoting the item views. Illustration of (a) 5, (b) 10, and (d) 20 attributes selected. As attributes are selected, we notice an improvement in label separation and reduction of items positioned in the central area (region with many mixed items).



Source: Adapted from Artur and Minghim (2019).

to model a severity score dynamically, with updates in small periods according to the patient evolution. Analysts firmly agreed with that assessment and mentioned that they are currently working to perfect dynamic score.

Figure 15f shows that – particularly for this data set – a relatively large slice of the "hemorrhagic shock" and "TBI" cases have also score values indicating good recovery. Based on this, some hypotheses may be raised to explain the imprecision of the score. It may designate a certain error margin of the score's accuracy, or it may reflect different contexts in which the same scores are applied. In this case, there is a field for modeling custom scores that satisfy the local reality. Domingues et al. (DOMINGUES *et al.*, 2017) argue that Trauma and Trauma Revised Injury Severity Score (TRISS) has been developed based on data from high-income countries, such as USA and Canada, and when applied in low-income and middle-income countries it loses accuracy. Based on these analyses, authors have suggested adjustments to the score's coefficients based on LR model to increase the efficiency for these cases.

We demonstrate in this case study how to employ our approach in the analysis of attribute space combined with a collection of observations of the data space, particularly concerning the prediction potentialities of categorical attributes. It became clear that this type of analysis can be done even by non-experts in either the data set or the visualizations at hand, of course, with the help of experts in interpreting the observations.

3.6.2 Case Two: Finding Representative Subspaces

In the previous case study, health attributes would be well known and meaningful to the analyst, who has prior experience with the data. Non-experts can also gain insight since they understand the meaning of attributes in the data set. In some cases, though, such as in image collections, the attribute themselves are not individually meaningful to the final user. In this case study, we demonstrate how it is possible to deal also with data sets without prior user knowledge of attributes.

Here, we use the Corel (LI; WANG, 2003), the Human Activity Recognition Using Smartphones (ANGUITA *et al.*, 2013) and the News data sets. The first one has 150 attributes extracted from 1,000 images representing ten different image categories (African people and villages, beach, buildings, buses, dinosaurs, elephants, flowers, horses, mountains, and food). Each category is composed of 100 images, which makes the data set balanced. The second one includes data from an experiment with a group of 30 volunteers aged between 19 and 48 years of age. Each person performed six different activities; walking, walking upstairs, walking downstairs, sitting, standing, and laying. Through a smartphone attached to their waist, data from a gyroscope and accelerometer (3-axial linear acceleration and 3-axial angular velocity) is collected at a frequency of 50 Hz. It has 561 attributes extracted in a total of 10,299 items. Here we deal with a partition of this data set containing the first 1,000 items. Finally, the News data set contains 1,771 RSS news feeds from BBC, CNN, Reuters and Associated Press and its 3,731 attributes were created using term frequency–inverse document frequency (TF–IDF) with stemming and stopwords filtering. The data set is labeled between 23 categories.

We proceed with the FS highlighting interesting observations that might guide a good choice of attributes. In the following, we evaluate the subset selections illustrated here against automated FS methods, so we can discuss the practical advantages and disadvantages of performing FS tasks interactively aided by a visual support technique.

- **Feature Selection in the Corel Data Set.** Aiming to select the first ten attributes, we focus on the strongest correlations per label. Figure 16a presents both views with the first five chosen attributes. The star-like shape in the data space view reveals how attributes represented in DAs influence the position of elements concerning positively correlated labels. There is an apparent distinction related to the labels that already had a correlated attribute picked, but the central area remains considerably mixed. The separation between categories becomes more evident as we select ten attributes, one highly correlated attribute for each label (see Figure 16b).

  For the next picks, we adopt a different selection strategy. One of the primary concerns in FS methods is to filter out redundancies, avoiding the choice of attributes that might carry the same information. To check for redundancies, users can explore the visual and interactive resources of the tool. When the user hovers the pointer over elements in the

attribute view, the data view is updated with the attribute values as new elements sizes (as explained in section 3.5). Then, by hovering the pointer over the attributes, the user can investigate patterns revealed in the data view in the attempt to detect redundancy.

Figure 16c presents the selection of ten additional attributes, two related to each label. In the item view, the improvement in label separation is clear, although some mixed elements remain in the central area. In terms of silhouette, both the 10 attributes and the 20 attributes subsets improve over the 150 original attributes data set (in fact the 10 attributes subset improve even more, but for effect of classification and of projection the 20 attributes subset was better). Silhouette of the 150 attributes data set was 0.16; for 20 selected attributes it was 0.19, and for 10 selected attributes it was 0.24.

So far, we made a straight selection considering that the Corel is a balanced data set with strongly label-to-attribute correlations already apparent in the first rendering. The initial ten attributes were chosen directly as the top correlated for each label value. The second FS round was similar to the first; however, it required more care to avoid redundant attributes. Although it is hard to state precisely, we consider that in general FS tasks in data sets like this demand less than half an hour to be performed – not taking into account the time needed to learn the tool.

- **Feature Selection in the Human Activity Data Set.** In this data set, we had to make a more in-depth investigation by performing a tree-like exploration on the data set partitions through the refine/pruning mechanism. Figure 17 illustrates the FS task of the first ten attributes. We start by selecting the most significant correlation of the entire data set (the largest when investigating label-by-label). The “41 tGravityAcc-mean () - X” attribute has a correlation coefficient of 0.98 with the category “Laying”, which almost wholly segregate it in the data view. Figure 17a presents the attributes arrangement when hovering over the Laying-DA, where there are six strongly correlated attributes (we need to consult the rank to precisely distinguish the levels of correlation between them). Selecting another one from this group, purposely the strongest with the opposite correlation direction related to the first one (see the lines between attributes and DA in Figure 17a), we achieve the whole segregation of the category (see Figure 17b), which allow us to prune the data to focus on the remaining elements. Then, after selecting three more attributes (see Figure 17c), we notice that the categories are segregated into two clusters, which define our strategy of pruning them to perform two other FS tasks. Finally, in Figures 17d and 17e we select the attributes that promote proper visual segregation of the refined clusters.

  This time, we had to investigate further and look for particular correlations with the ability to segregate specific partitions. In data sets like this, the perception of the user is essential to determine what to select, where to prune, and when to stop. Even so, we consider this task to be accomplished in less than an hour for a simple FS without additional claims.

Again, this statement is quite subjective and depends a lot on what the user is actually looking for in the data set.

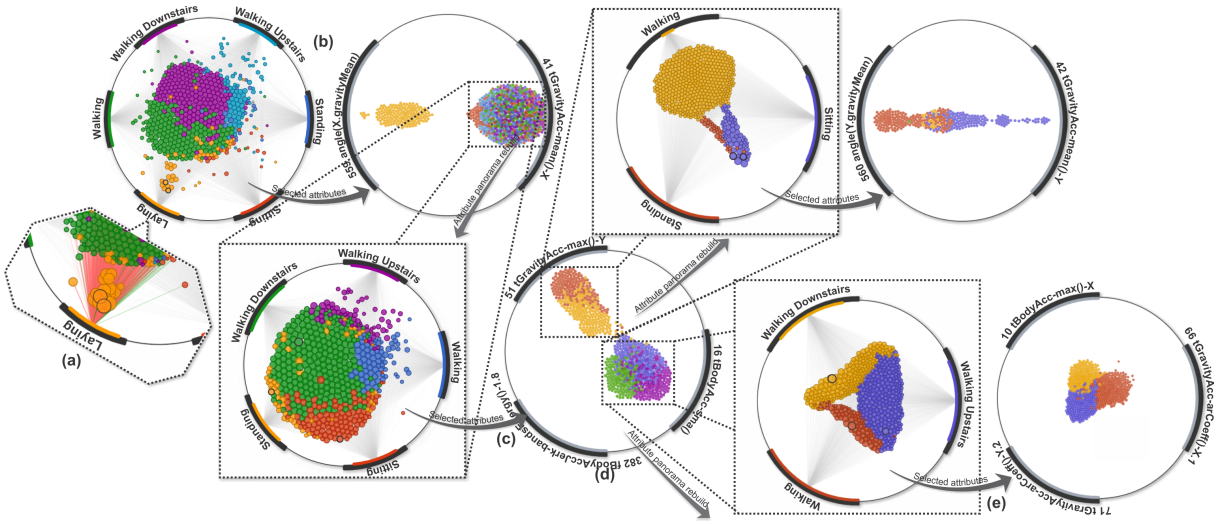
- **Feature Selection in the News Data Set.** The News is a sparse and unbalanced data set; following from this, we adopt the strategy of choosing attributes that segregate the most significant number of items in the current rendering state. For example, the first two selected attributes "Murdoch" and "debt" are correlated with the two most populous categories "London hacking scandal" and "USA crisis", respectively. In the data view, whenever we select new attributes, the elements not yet segregated remain mixed in the very central position of the unit circle, which helps the pruning process to start a newly refined sub-selection following this same strategy.

  We have also implemented mechanisms for multiple selection, which should help users when representative subsets are reached only with numerous attributes. In the Corel data set, after the FS previously described, we defined the multi-selection click mechanism to pick eight attributes, and then we right-clicked over each DA, thus achieving 100 selected attributes. For the Human Activity data set, instead of the selecting the five attributes presented in Figures 17d and 17e, we define the multi-selection click to select 19 attributes and then we right-clicked on each DA, also reaching the subset of 100 attributes. In the News data set, after the second pruning, eight categories remained, so we defined the multi-selection mechanism to pick eight attributes for each category (nine for the largest).

- **Classification Experiment.** To evaluate the quality of the selected subsets in the previous subsection, we confront it with subsets made by some of the most traditional automated algorithms. All chosen algorithms are of the filter type for supervised FS tasks; six are based on mutual information named: *Conditional Infomax Feature Extraction (CIFE)* (LIN; TANG, 2006), *Conditional Mutual Info Maximisation (CMIM)*(FLEURET, 2004), *Interaction Capping (ICAP)* (AKADI; OUARDIGHI; ABOUTAJDINE, 2008), *Joint Mutual Information (JMI)* (YANG; MOODY, 1999), *Mutual Information Feature Selection (MIFS)* (BATTITI, 1994), *Max-Relevance Min-Redundancy (MRMR)* (PENG; LONG; DING, 2005); three are similarity based: *Fisher Score* (DUDA; HART; STORK, 2000), *ReliefF* (ROBNIK-ŠIKONJA; KONONENKO, 2003), *Trace Ratio* (NIE *et al.*, 2008); and three are statistics based: *Chi Square* (LIU; SETIONO, 1995), *F-Score* (WRIGHT, 1965), and *Gini Index* (GINI, 1912).

  To proceed with the experiments, we select as classifier M the linear *SVM* method. The tool applied to perform the tests is the *FeatureMiner* (LI *et al.*, 2017), which includes all mentioned FS algorithms and the classifier implementation. The selection and evaluation are performed with 10, 20, and 100 attributes selected. For the whole experiment, we employed 5-Fold cross-validation. We define the MIFS algorithm parameter β equal to one. Our test/training ratio corresponds to 0.2, and our statistical measure is the classification accuracy.

Figure 17 – Steps while performing an FS in the Human Activity Recognition training set. Investigating each DA, we find an attribute group that assumes the largest sizes when (a) hovering over Laying-DA. After selecting two of these attributes, (b) there is a complete segregation of the "Laying" category in the data view. Then, through the refinement mechanism, the segregated elements are pruned to rebuild the correlation matrix focusing on the remaining data. In (c) three more attributes are selected witch reveals two large clusters that segregate the categories into two groups; therefore, the FS strategy is from now on to prune and select correlated attributes for each group. In (d) and (e) attributes that contribute to the segregation of the refined partitions in (c) are selected, totalizing the selection of ten attributes.



Source: Adapted from Artur and Minghim (2019).

Table 3 reports the classification accuracy results applying the FS described earlier in comparison with automatic methods. Given the characteristics of the tested data sets, we realize that it is advantageous to use our approach. By selecting a few attributes, we achieve robust representative subsets.

The custom FS strategy performed by the user should not always follow as outlined in this case study. The relationship between chosen categorical attributes or class labels and the remaining attributes is, naturally, very much dependent on the application. This is the proposal of the visual FS support, to show a panorama that leads the user to carry out a selection of representative subset from their insights.

3.6.3 Additional Experiments

Besides the controlled experiments above, we have set to find the answer to two particular questions: 1) "how does the approach perform when submitted to a pre-defined hypothesis on data attributes?", and 2) "how does the analyst view the utility of the approach?".

To answer the first question, two users have employed our tool to choose a subset with minimum attributes, from 27 available, that would keep or improve segregation of 4 labels in a

Table 3 – Classification accuracy of FS tasks with 10, 20, and 100 selected attributes among most traditional filter-based methods. We have left the trace ratio test for the News data set without results since we kept the algorithm running for days without any response.

	Corel			Human Activity			News		
	10	20	100	10	20	100	10	20	100
CIFE	0.41	0.58	0.89	0.69	0.81	0.92	0.68	0.84	0.88
CMIM	0.27	0.41	0.89	0.77	0.85	0.96	0.65	0.83	0.97
ICAP	0.27	0.42	0.88	0.83	0.89	0.98	0.64	0.83	0.97
JMI	0.27	0.42	0.87	0.86	0.94	0.98	0.64	0.74	0.96
MIFS	0.42	0.59	0.87	0.77	0.88	0.96	**0.69**	0.84	0.95
MRMR	0.46	0.60	0.87	0.71	0.87	**0.99**	0.64	0.74	0.96
ReliefF	0.37	0.58	0.88	0.70	0.84	0.94	0.22	0.31	0.94
Trace Ratio	0.42	0.63	**0.90**	0.78	0.85	0.90	–	–	–
F-Score	0.48	0.64	**0.90**	0.61	0.71	0.84	0.36	0.38	0.96
Fisher Score	0.46	0.63	0.89	0.68	0.82	0.90	0.36	0.38	0.96
Chi Square	0.41	0.61	0.89	0.59	0.73	0.82	0.28	0.31	0.77
Our approach	**0.76**	**0.84**	**0.90**	**0.96**	**0.98**	0.95	0.68	**0.86**	**0.98**

Source: Research data.

data set with $4,340$ data items (the application was indices in acoustic landscapes). Both users managed to achieve such a reduction, both keeping the same levels of segregation (as defined by various measurements over multidimensional projections of the data set, including original RadViz and t-SNE). One of the two users reduced the number of attributes to 4 and the other to 3. That first part of the job was performed in under half an hour. In order to segregate all labels at once more work was necessary since two of the labels were very hard to segregate. The successful strategy for this last trial was performed by one of the two users, who looked into attributes that were highly correlated to each class separately. In the end, a set was found that improved the segregation of the original four classes. That set was comprised of 10 attributes and the task took approximately 4 hours to perform. As a relevant note, that task of finding subsets of attributes for this particular application was being carried out in conventional ways for days with very little success without the help of our approach.

To answer the second question, regarding the user's opinion, the solution was shown remotely (with the use of a video) to one particular target user in the field of trauma medical records. We had already an idea of the utility of the tool in the case, since it had been demonstrated to other partners in the trauma application project before, who were very impressed with the capability of the tool. This analyst had no previous contact with the tool and found the system very useful for its purpose and is very excited to build that into their general system, with the purpose of visualizing the scenario of resources application and review of procedures and protocols in hospital emergency wards. All the suggestions for improvements presented by the analyst regarded the need to have documented reports of the findings in terms of actual written documents and charts to explain and register in more formal terms the solutions found by the user.

3.7 Final Remarks

In this work, we have presented an approach to analyze and select attributes; it includes the Attribute-RadViz, an interactive technique for visually displaying and exploring multiple correlations between attributes in a data set. The approach is based on relating all attributes at once to target categorical ones (or to labels) to find predictive subsets to the label values individually or in combination. Since categorical attributes are commonly vast in current data sets, this is expected to resonate within a large number of applications. It certainly has good results in at least two applications, to find patterns and verify hypotheses, as well as to perform useful FS tasks. Once users learn about these displayed relationships – some expected, others acquired – they can utilize their knowledge in the follow-up activities such as mining and machine learning tasks.

Among the contributions of our work, we highlight the development of a visualization technique coupled with proper data processing that exposes a cognitive map with the straight relationship between attributes and a chosen label set. This particular aspect is difficult to realize with any of the previous attribute analysis tools. Having the possibility of investigating the predictive capabilities of attributes towards categories is very useful to gain knowledge of the subject under analysis as well as finding ways towards predicting outcomes and target classes, sometimes identifying more than one phenomenon happening at the same time within a single data collection exercise.

Secondary achievements of our approach are the development of new visual widgets that would work in general to improve the analysis capabilities of RadViz (e.g., arcs providing extra information for each DA, the interaction model that changes visual attributes for features under analysis); Also, we demonstrate how to reveal valuable information from data sets by the combination of interactions between attribute and data views throughout our case studies.

Some limitations persist in our work. The tool is not effective when the data has only two label values, or when the number of target label values is too high and beyond RadViz capabilities, although that particular aspect is troublesome in its conception anyway. More than 20 categories, for instance, is challenging to handle under the same set of attributes. Concerning the FS capabilities, we show that there is significant potential in the subset selection performed by the user through our approach; however, for a substantial number of attributes, this task becomes challenging. We have implemented multi-selection mechanisms, but more sophisticated procedures (i.e., semi-automatic selections combining user choices with automated FS methods) would be ideal in applications where the size of attribute subsets reach the hundreds mark. Another limitation is the need to choose a useful and meaningful categorical target attribute. Poorly labeled data may bring inaccuracy in the correlation matrix. Nevertheless, punctual errors in the label values are tolerable and could be detected in the data view by observing outliers.

Another limitation relates to the correlation estimations regarding categorical and numeric

variables. In this approach, we adopted two distinct ways to perform the calculation, first by a regression model and also by binarization of the label values (one-against-all strategy). However, when applying regression, there are problems in the overestimation of correlations, especially in cases when dealing with sparse data. In binarization, there are problems in hiding useful correlations evident only in all-against-all strategies. Although the approach is conceptually independent of the metrics, we remain looking for more precise models for the hybrid calculation of correlations.

As future work, we intend to address aspects of scalability and support for more diversity in data types. For scalability, it is interesting to implement filters and introduce automatic methods to assist in the complementary FS, helping in situations with a more massive demand for selected attributes. The approach works with numerical and categorical data. Newer versions are being designed for textual data. We are also interested in studying how the proportion of numeric and categorical attributes impact the precision of the correlation matrix and how to develop new correlation methodologies to make it even more accurate accordingly to this balance.

CHAPTER

4

ATTRIBUTE ANALYSIS TO EXPLORE REGRESSION MODELS

In this chapter, we apply the potential of Attribute-RadViz in identifying correlations levels of attributes to explore LR models. We focus on reducing the limitations of applying those models in multidimensional data contexts. The developed methodology is based on scenarios derived from health records; however, the approach can be generalized in other broader contexts where one desires to predict phenomena of interest encoded into potential target attributes.

4.1 Introduction

The medical research studies can be highly benefited by multidimensional visualization and multivariate prediction tools. Nowadays, large amounts of medical data (e.g., clinical information and treatment outcomes) have been stored in structured electronic health records (EHRs). These EHRs can represent a valuable source of information in an attempt to extract insights that can shape healthcare methods (GOLDSTEIN *et al.*, 2017). A common way of using these data is by applying the predictive power of attributes to evaluate probable outcomes supporting the decision-making process of the healthcare team.

Regression analysis is a widely used method in scenarios of health records analysis (DREISEITL; OHNO-MACHADO, 2002; GOLDSTEIN *et al.*, 2017), where relevant attributes could potentially be used to build models for predicting outcomes very accurately. However, regression models generally do not deal well with multidimensional data, mainly because of the multicollinearity problem (MANLY; ALBERTO, 2016), where strongly correlated attributes can destabilize the model. Thus, an FS step is vital in the construction of a proper statistical model based on LR.

Another essential aspect of the regression analysis is the evaluation of the model's performance. Among many consolidated goodness-of-fit measures, a popular way to evaluate

LR efficacy is through receiver-operating characteristic (ROC) curves (METZ, 1978; HANLEY; MCNEIL, 1982), where the relationship between specificity and sensitivity is exposed among the range of cut-off values (GREINER; PFEIFFER; SMITH, 2000). However, this is efficient to evaluate binary LR, where the goal is to determine the occurrence or not of an event. For the multinomial LR, other solutions to evaluate the model's performance should be used, such as confusion matrices. We have developed a novel variant of the RadViz visualization technique to provide an overview of the model classification showing how the multinomial LR might be hitting or missing.

We describe in this chapter a comprehensive approach for LR models exploration, which includes: (i) feature selection, (ii) regression model construction, (iii) evaluating binary and multinomial regression, and (iv) constructing a panorama for queries over the model. The input of the approach is the data set containing the target attribute with the desired outcome labels. We employ a previously developed tool to provide an overview of attributes for the analyst to perform FS and evaluate combinations of attributes and their effect on target attributes. Optionally the analyst can create a query tool based on the generated regression models.

4.2 Interactive Logistic Regression Model Builder

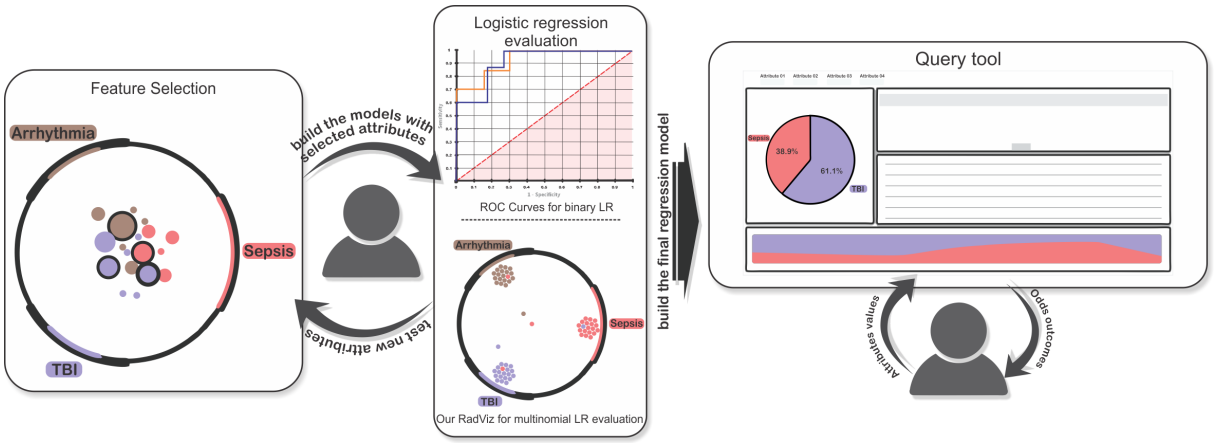
In this section, we describe the general methodology of the proposed approach. Firstly, the analyst performs an FS related to some chosen label set and then evaluates its performance as an LR model through ROC curves or RadViz visualization. The analyst is free to evaluate various combinations of subsets to define the model more accurately. After the analyst already knows or has gathered a considerable amount of information, he or she can create a query panorama based on a definitive LR model. The generated overview for queries can be accessed later without the need to perform the previous steps. When consulting some profile in the tool, the analyst should provide the associated values of the prior selected attributes, and the tool will show the probabilities of the outcomes as well as the cases most correlated with the queried one. Finally, it is also possible to visualize the evolution of outcomes through streamgraphs by submitting updated attributes values, assuming the existence of attributes that vary dynamically. Figure 18 shows the pipeline of the approach.

4.2.1 Interactive Feature Selection

After loading the data set, the first step in this approach is to perform an FS. As mentioned earlier, the LR models are sensitive to strongly correlated independent attributes. Thus, a good FS is required, which aims not only to select relevant attributes but also to select attributes that complement the selected subset in the potential of describing the target one.

Unlike automatic FS algorithms, an interactive selection allows analysts to use their prior knowledge as a criterion for choice. For example, in the trauma EHRs, several attributes correlate

Figure 18 – The pipeline of the approach for creating and evaluating an LR model. Initially, the analyst can create regression-based models and evaluate them interactively. After gaining more knowledge, the analyst can generate a definitive model for later use within a tool for quick queries.



Source: Elaborated by the author.

with the patient's final condition, especially the trauma scores. One of the well-known scores is the Revised Trauma Score (RTS) (CHAMPION *et al.*, 1989), which is a combination of the attributes: Glasgow coma scale (GCS), systolic blood pressure (SBP) and respiratory rate (RR). Once RTS is selected, the other three attributes should be preferentially not selected, or vice versa. Hence, user selection typically carries some tacit knowledge undetectable by automatic algorithms.

To accomplish this task, we use the Attribute-RadViz approach, described in chapter 3. Once the correlations between attributes are exposed to each label, users can analyze them individually and select their relevant attributes. This allows the creation of binary LR models fast and easily for any label inside the target attribute.

4.2.2 Logistic Regression

LR is widely used when one wants to predict the probability of some outcome that is regularly binary. It is a robust discriminative method that explicitly provides the user probabilities of classification (SHEVADE; KEERTHI, 2003). The model is generated from observations where one or more independent attributes (discrete or continuous) determine an outcome. In our approach, we employ both classical binary LR and its generalized alternative, the multinomial LR.

Binary LR is a special type of regression where a dependent attribute, which represents a binary outcome, is related to a set of independent attributes, also known as explanatory variables. Binary LR differs from other regression types in that it does not attempt to predict a value through the linear combination of the independent attributes. It tries to predict the odds and probabilities of a given event to occur or not occur.

For a given number of explanatory attributes, $y_1, y_2, ..., y_m$, the probability of the response variable occurs or not is a function p of the given attributes. The prediction function is defined by the equation below:

$$p = \sigma(w_0 + w_1 y_1 + w_2 y_2 + ... + w_m y_m), \tag{4.1}$$

where $w_0, w_1, ..., w_n$ are the coefficients (or weights) associated with different explanatory attributes, and $\sigma(z)$ is the logistic function which maps the prediction into a probability value between 0 and 1. The logistic function is usually calculated by the sigmoid function, defined below:

$$\sigma(z) = \frac{1}{1 + e^{(-z)}}. \tag{4.2}$$

We apply binary LR to generate predictive models related to isolated label values present in the dependent (target) attribute chosen by the user. However, often the user wants to examine the probability scenario between the various label values of the target attribute into unified charts. For this purpose, we also employ the multinomial LR.

The solution we adopt to solve the multinomial regression splits the problem in a set of $k-1$ binary logistic models, being k the number of label values inside the dependent variable. For each sub-problem, we find the coefficients of the model and apply in the logistic function, which is now the softmax function:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_1^k e^{z_k}}. \tag{4.3}$$

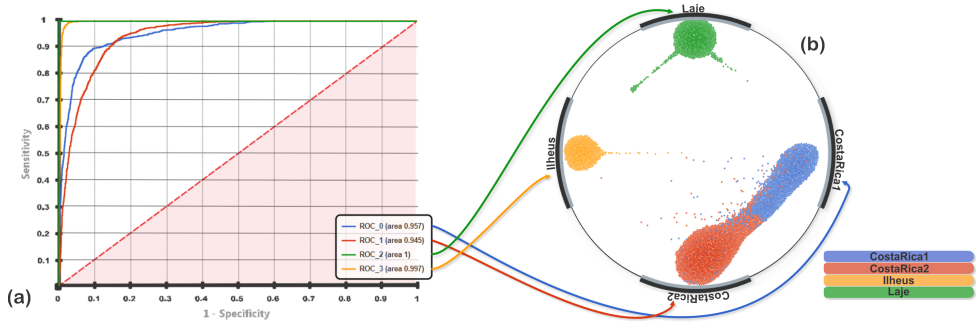
Hence, we can provide a complete scenario of outcomes probabilities to the user, both concerning the odds of isolated labels occur or not as well as multi-label situations.

4.2.3 Logistic Regression Evaluation

We provide visual interactive means to evaluate the generated LR models. The first one is the well-known ROC curve, which is generated as soon as the LR model is ready. The second method of evaluation is through a visual strategy, in which we adapt, again, the RadViz visualization technique. In the following, we describe details of implementation and the description of the interactive interface of each of the two tools.

ROC curves are widely adopted for performance analysis in classifiers (GONÇALVES *et al.*, 2014). It describes the relationship between the specificity (true negative rate) and sensibility (true positive rate) beyond the scope of a threshold assumed by some diagnostic test. A handy measure extracted from the overall performance of the classifier examined in the ROC curve is the area under the curve (AUC), and it can be interpreted as the average value of sensitivity

Figure 19 – Distinctive alternatives evaluating the same regression model generated by the approach. In this case, the identification of acoustic data origin from four different regions. (a) ROC curves identify the individual efficiency of the binary LR as well as the ideal cut-off values for each model. (b) The LoRRViz displays by the proximity for DAs how each item is classified by the multinomial model, giving a broader view of how the model might be missing.



Source: Elaborated by the author.

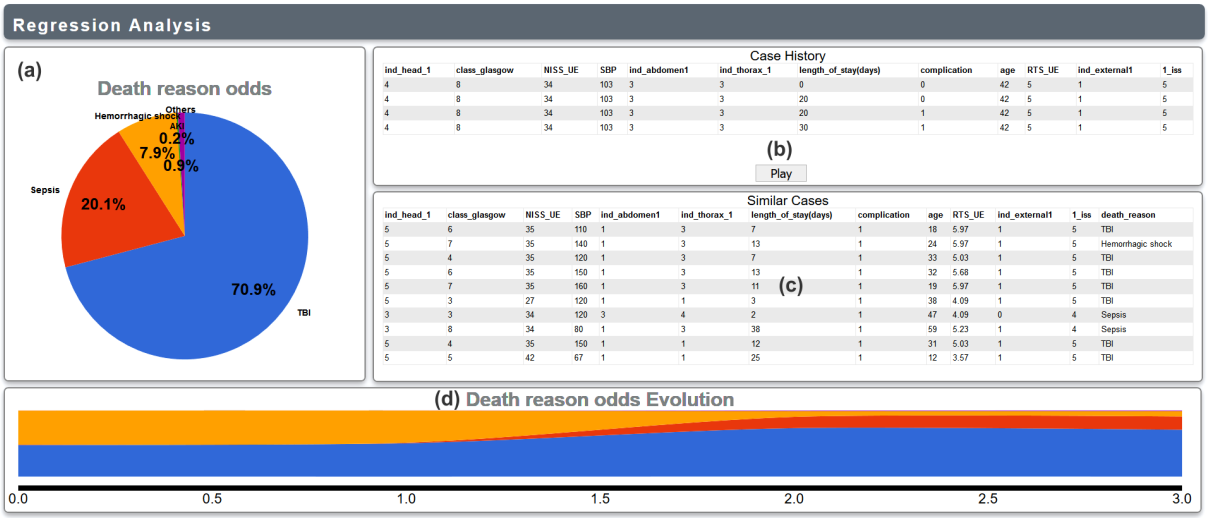
among all possible specificity values (ZHOU; MCCLISH, 2011.). In our approach, along any point in the path of the ROC curve, users can consult the sensitivity, specificity, AUC, and cutoff values. Various ROC curves can be plotted simultaneously to compare the generated models, as shown in Figure 19a. Also, the user can consult hidden information of each generated model interacting with the ROC view legends, as the overall model fit values (e.g., p-value), as well as the generated regression equation.

ROC curves are excellent for evaluating the performance of a classifier when the response is binary. Nevertheless, when the scenario includes a multi-label condition, that is, users want to hold more than two categories in the model; a multinomial LR modeling should be applied. For these cases, we can analogously evaluate the models separately by several ROC curves. However, some details of the model evaluation remain unclear, such as how the models are missing the mark.

We have developed Logistic Regression Radial Visualization (LoRRViz) to reveal the efficiency of the multinomial regression model visually (see Figure 19b). Each DA represents a label value and exerts attraction force according to the probability value for each item defined by the LR model. If an item is placed very near some label-DA, it implies that the probability value of this item regarding that label is high (and low regarding the other labels). In contrast, if an item is placed equidistantly between two labels, it may mean that the probability defined by the model is the same for both labels represented by the DAs. Interactively, the user can further investigate these cases of inconsistent probabilities generated by the model.

In practice, the entire matrix of n items versus k labels is assembled through the Equation 4.3 (or Equation 4.2 if only one label is chosen), resulting in a n x k matrix. The interpretation is that each cell holds the probability that item x_i belongs to the label l_j. Finally, the RadViz mapping is applied (Equation 3.4), where the interactive resources described earlier for the RadViz such as the force scheme improvement, DAs management, and positioning distortion

Figure 20 – The query tool interface composed of four panels. (a) Pie chart presenting the probabilities calculated by the regression model for the consulted profile. (b) The historic containing submitted profile states allowing the user to perform comparisons. (c) Cases most correlated with the last consulted profile. (d) Streamgraph containing all submitted states exposing trends in the evolution of probabilities.



Source: Elaborated by the author.

adjustment are also available in LoRRViz.

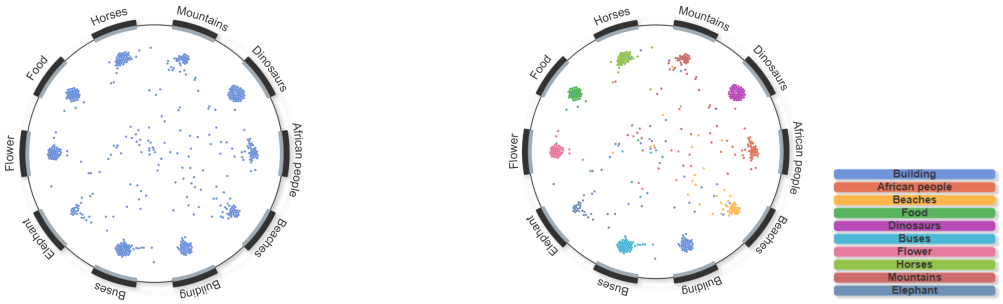
Figure 19 shows the evaluation of a multinomial regression model generated from an acoustic data set. The data has been collected in four different regions; their attributes have been extracted, and then we have created an LR model able to identify the origin of the samples by our approach. The color represents the actual origin, and the position represents the way in which the model classified the samples. We can note that, for the “Ilheus” and “Laje” categories, the model classified with quite a significant accuracy. However, in the “Costa Rica 1” and “Costa Rica 2” categories, we notice a considerable mix of samples, and it is up to the user to investigate whether the model needs adjustments focusing on these categories (such as new selection of attributes), or whether the samples are just intrinsically indiscernible regarding the available attributes.

4.2.4 A Regression Query Tool

Our approach allows the generation of regression models according to the user’s interest in predicting some single or multiple events. Also, we include resources to enable model evaluation and reveal potential adjustment demands. However, aiming for an all-around solution, we have developed a query tool to take advantage of the previously generated learned data for later queries related to any desired profile.

The tool usage is quite simple; the user restores previously saved regression data and then inserts the information about some profile that he or she wants to query; hence, the query

Figure 21 – Viewing estimated results by the multinomial LR model with the test partition (half of all samples) of the Corel data set. In addition to individual profile queries, users can query entire data sets using CSV files, where the tool automatically catches the attributes associated with the models by the performed FS tasks. The visualization can be rendered (a) without the label information or (b) applying the labels (if available) to encode the colors of mapped elements.



(a) (b)

Source: Elaborated by the author.

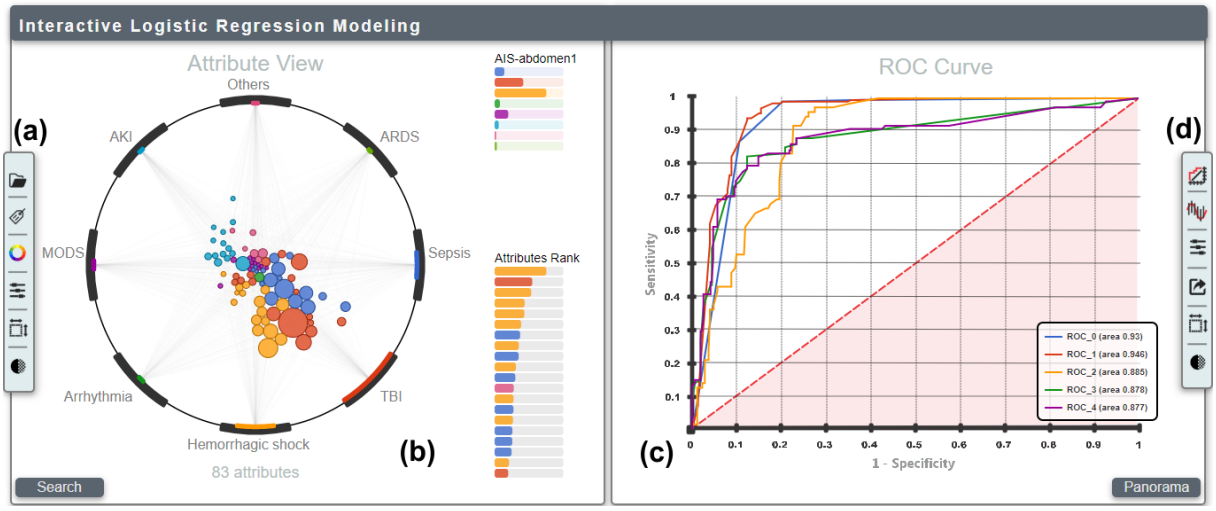
tool will return the probabilities of this profile belonging to any of the labels. He or she can also submit several states of this profile; in this way, the tool will expose evolution trend patterns. For example, in the case of a patient who varies his clinical condition over time, the user can submit his or her states and observe the variation of probabilities through the outcome trends exposed by the model.

Figure 20 shows the query tool interface. There are four panels per generated model. The first one (see Figure 20a) presents a pie chart with the probability scenario of the current consulted profile. The second (see Figure 20b) presents the history of submitted states. The next panel (see Figure 20c) shows the most correlated cases with the last consulted profile. To make this panel available, the user must, in addition to opening the regression model file, to load the original data set. Finally, the last panel (see Figure 20d) shows, through a streamgraph, the evolution of the probabilities according to the profile changes submitted by the user. This panel is useful when the analyzed object has attributes that change dynamically over time and reveals trends in the probability scenario.

The tool also allows multiple models queries. For example, in a trauma data set, we can find more than one target attribute in which we would like to create prediction models, such as the “condition of discharge” and “cause of death” attributes. Hence, we can imagine three scenarios that can be queried simultaneously: “probability of survival”, “condition of discharge (in case of survival)”, and “cause of death (in case of obit)”. Thus, the user can set up a more comprehensive odds scenario provided by the regression models generated from the same data set.

There is also the possibility of querying not only individual profiles but entire lists of items. As soon as the data is ready, the tool scans for the selected explanatory attributes of the built regression models, and them an overview of the classification is displayed through the

Figure 22 – Snapshot of the main interface of our prototype. (a) The attribute view control panel. (b) Visualization of attributes through RadViz, which constructs a cognitive map showing the relevant attributes for each label value of the data set. (c) Panel for evaluation and review mechanisms of generated regression models. (d) Control panel for evaluation settings.



Source: Elaborated by the author.

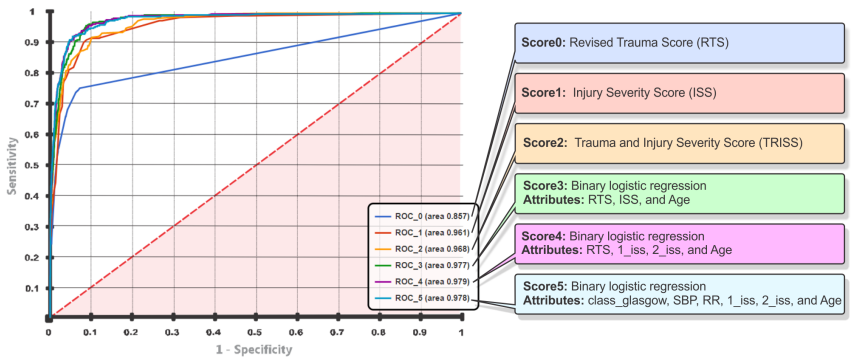
LoRRViz. Optionally the user can encode the color of the mapped elements based on some chosen attribute (usually the one containing the labels); this is useful for visualizing the efficiency of classification generated over supervised data (see Figure 21).

4.2.5 Prototype Implementation

We have developed the approach as a web-based tool predominantly encoded in Javascript language. Also, we use the D3js library to handle visual elements and HTML with CSS for the construction of the front-end available to the user. The regression module is also encoded in Javascript, and it is originally obtained from The Interactive Statistical Pages (<http://statpages.info/>). The prototype and its full source code are made freely available at <https://github.com/erasmoartur/lrxptool>.

Figure 22 presents the main interface of the prototype. On the left side, a control panel (see Figure 22a) allows the user to open the data set, define the target attribute, and change visual attributes panorama settings. The right side contemplates the evaluation and check mechanisms for generated regression models (see Figure 22c), employing either LoRRViz or ROC curves. The user selects the evaluation mode from the right control panel (see Figure 22d), as well as the settings for that view. In the following section, we present a step-by-step procedure applying our prototype to create and explore LR models. The experiment sequence follows the pipeline shown in Figure 18, focusing on a practical process over a real data set.

Figure 23 – A comparison of the performance between well-known trauma scores and generated regression models with the training data set. Score 4 was the most accurate for this data, where its concept is derived from the traditional TRISS, but it includes the values of the two worst injury segments instead of the ISS.



Source: Elaborated by the author.

4.3 Usage Scenarios

We present three usage scenarios investigating EHR data sets to reach different purposes. In the first two scenarios, we employ the same trauma records data set previously presented (see Subsection 3.6.1). Firstly, we load the data set to explore combinations of relevant attributes to generate LR models and further confront these models with well-known trauma scores. Then we show how to take advantage of the generated models to build a query tool and also how to apply it to gain insights into the presented predictions. In the last scenario, we report a brief analysis and the construction of LR models for a data set referring to the novel COVID-19 disease.

4.3.1 Scenario One: Predicting Mortality with Trauma Scores

As previously explained, trauma scores are mathematical and statistical models that try to characterize and document traumatic injuries levels. It can help in predicting the outcome of the patient and aid in the triage of the trauma patients. We compare the effectiveness of well-known trauma scores related to the patient risk of morbidity. The scores are TRISS, RTS, and Injury Severity Score (ISS); then we create and test variations of these scores to check and increase the prediction efficiency.

Our test/training ratio corresponds to 0.2; the metric for comparison is the AUC; which is automatically displayed by the approach as soon as the model are created. Our intention here is not to develop a new improved score; we would like to show the usefulness of the approach for analysts to investigate and test hypothesis creating their own regression-based models in a practical and rapid manner.

Figure 23 shows the ROC curves generated by applying the scores to predict mortality. The ROC curves of RTS, ISS, and TRISS scores are rendered directly as their values are present

Figure 24 – Attribute selection for three different contexts: survival, final condition, and cause of death odds; For this purpose, two different target attributes representing outcomes are selected: final condition and cause of death. (a) Selected attributes for the target attribute “final condition”: RTS, Age, ISS, AIS-head_1, surgery, previous_pathology, and complication. (b) Selected attributes for the target attribute “death cause”: AIS-head_1, complication, lengh_of_stay, AIS-abdomen, AIS-thorax, AIS-external_1, SBP, and RTS.

(a) (b)

Source: Elaborated by the author.

in the source data set. Then, we have generated scores 3, 4, and 5. The score 3 employs the TRISS concept, which is essentially an LR involving RTS, ISS, and age; the efficiency increase (AUC equal to 0.977 against 0.968 of the original TRISS) is due to the new regression training, which adjusts its coefficients according to the local reality of the data under analysis. Domingues *et al.* (2017) discuss the applicability of TRISS in different contexts explaining this effect.

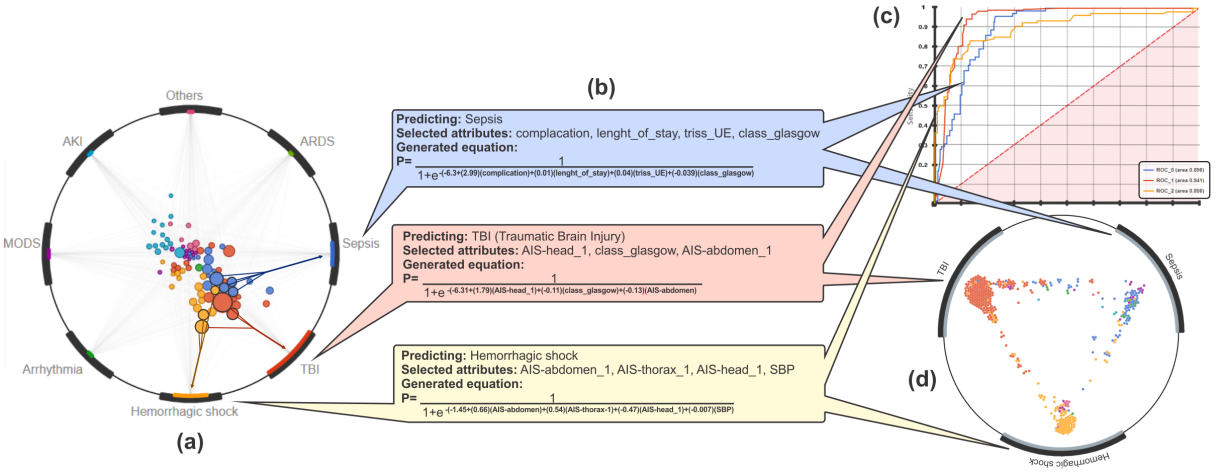
Scores 4 and 5 are hypothesis tests that we raised, where we may improve the TRISS score by changing attributes of its original equation, which is an LR with RTS, ISS, and age, by their primitive ones. In score 4, we replaced the ISS values – which represent the sum of squares of the abbreviated injury scale (AIS) values of the three segments with the most severe injuries – by the two most severe injury segments. Hence, the new LR model contains RTS, AIS-1, AIS-2, and age. Therefore, in this data set, the efficiency of the score has improved (with an AUC value of 0.979). The same idea has been extended to Score 5, where instead of RTS – which is the weighted sum of GCS, SBP, and RR – we insert those values directly to the regression model. However, no improvement noted in this score compared to the previous ones (AUC value of 0.978).

This scenario has shown how the agile creation and evaluation of regression models enable analysts to raise and test hypotheses about data efficiently. Additionally, they can model scores for prediction of events of interest encoded into categorical or numerical attributes in their own data sets.

4.3.2 Scenario Two: Building a Prediction Interface for Trauma Events

In this scenario, we present how to take advantage of the knowledge gained from the performed tests and then build a regression model for later queries. We follow each step described

Figure 25 – Generating and testing binary LR models. Before creating the definitive model for further queries, we have evaluated the performance of the highest correlated attributes for each label through ROC curves. (a) Selected attributes for the labels Sepsis, TBI, and Hemorrhagic shock. (b) Logistic model created after selecting attributes and clicking over the desired DA. (c) The individual model evaluation through ROC curves. (d) Evaluating the model by the LoRRViz.



Source: Elaborated by the author.

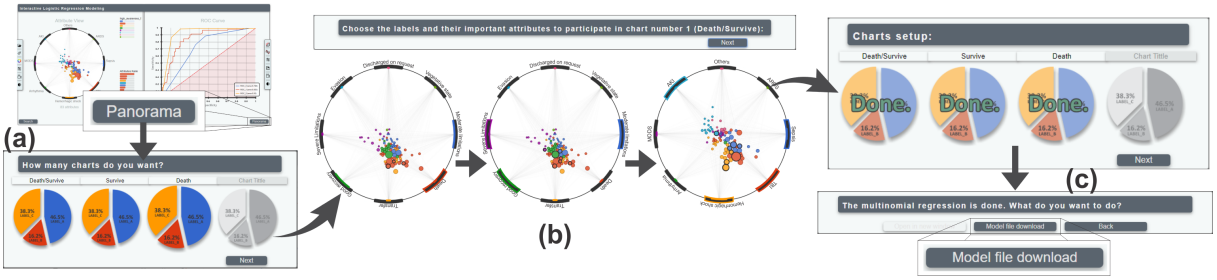
in Figure 18, showing how the resources available in the approach allow users to create a query mechanism quickly and effectively.

- **Feature Selection Step.** In the FS step, we try to choose the most relevant attributes for each label that we would like to predict. Thus, we must interact with the DAs (hovering the pointer) to discover correlated attributes and then generate its prediction model. We want to construct a scenario that returns odds of survival, discharge condition (in case of survival), and cause of death (in case of death). Figure 24 shows the attributes selected for each of these contexts. We have picked the most correlated attributes for each label, for example, toward the "hemorrhagic shock" label, the highest correlated ones are the "AIS-abdomen", "AIS-thorax", and "blood pressure".

- **Evaluating the Logistic Regression.** Selecting attributes and choosing a target label (by clicking on the DA) generates an LR model. Then its performance as a label predictor under the already loaded data is immediately presented through ROC curves on the right side of the prototype. Figure 25c shows ROC curves plotted by the generated models for prediction of labels "Sepsis", "TBI", and "Hemorrhagic shock". Since we want a scenario with multinomial regression models, we can use the LoRRViz to evaluate the generated models. To do so, in the right control panel, we chose LoRRViz. Figure 25d shows the classification of the models generated so far, including its inconsistencies; the analyst can investigate such cases and proceed, if necessary, with adjustments in the model.

- **Multinomial Logistic Regression Set Up.** Once we gain a better understanding of the

Figure 26 – Step by step illustration when generating the definitive model for this scenario. (a) After clicking in “Panorama”, the prototype prompts for the number of models/graphs to be created (in this case three), and their respective names. (b) Then, we enter the labels and attributes for each model. (c) Finally, the prototype returns the file with regression learning.



Source: Elaborated by the author.

data set, and we test the predictive power of the attribute’s subsets related to the labels in regression models, we can then generate a definitive and generic model. This model should allow future queries without the need to perform again the steps described previously.

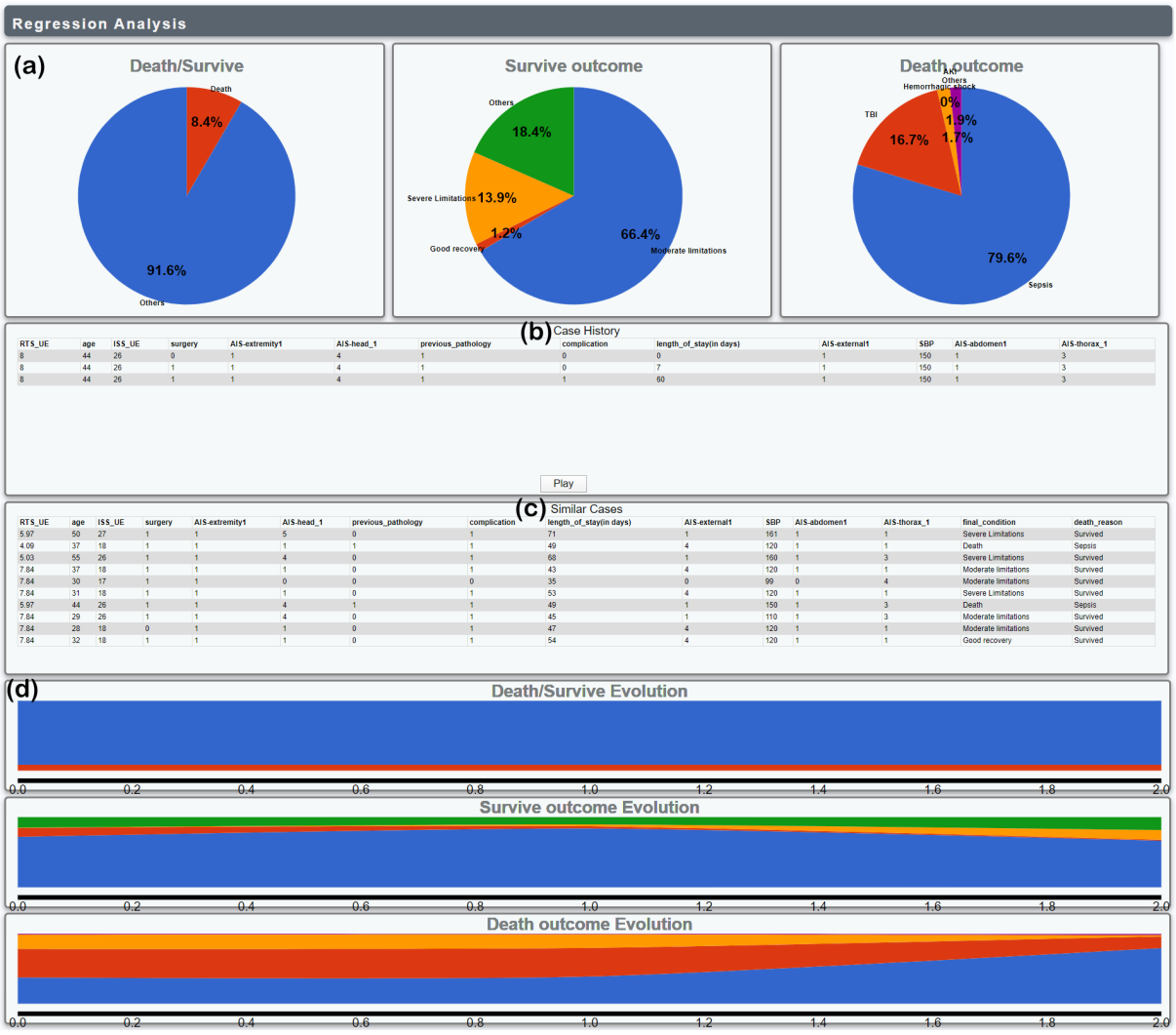
Inside the prototype, we click on the button “panorama” and it requests the number of graphs we want and their titles, as shown in Figure 26a. Users can choose more than one graph to make the tool able to display the probabilities of more than one outcome simultaneously, creating a more comprehensive overview. For example, in our case, we chose three graphs; the first presenting the probabilities of survival versus death; the second presenting the most probable discharge condition in case of survival; and the third presenting the most likely cause of death if the patient does not survive.

Then, we determine which labels and attributes will be part of each model (see Figure 26b). When only one label is chosen for some model, it is assumed that the user wants the binary LR scenario, where the odds are given for the label’s chance to occur or not. Here we choose the label death, that is death versus all, consequently death versus survival. The selected attributes are “RTS”, “Age”, and “ISS”. We click next to proceed to the second graph set up. Then we exclude the death label (only survival labels left) and choose the following labels that represent a discharge condition: “Good recovery”, “Moderate limitations”, and “Severe Limitations.” The attributes selected here are “RTS”, “Age”, “ISS”, “AIS-head_1”, “surgery”, “previous_pathology”, and “complication”.

For the last graph, we change the target attribute to “cause death”, and we chose the following labels: “TBI”, “Hemorrhagic shock”, “Sepsis”, and “AKI”. Then, the attributes “AIS-head_1”, “complication”, “length_of_stay”, “AIS-abdomen”, “AIS-thorax”, “AIS-external_1”, “SBP”, and “RTS” are selected. Finally, we click next to proceed with the creation of the model and download of the file containing the regression learning data (see Figure 26c).

- **Interacting with the Logistic Regression Query Tool.** After the generation of the mod-

Figure 27 – The query tool loaded with the previously described regression learning data. (a) The graphs are showing the calculated odds for the queried profile. (b) The submitted profiles (generally over time) for analysis of outcomes trends. (c) List containing the most similar cases to the last submitted profile. (d) The streamgraphs are displaying trends related to the states submitted in (b).



Source: Elaborated by the author.

els, we can perform queries to check the odds of the desired profiles. Inside the tool, we put the pointer on the top bar to make the input menu appear. Inside it, we insert the previously downloaded learning data and, optionally, we can add the original data so that the tool is also able to show the most correlated cases with the queried profile, this is useful for the user to verify if the results indicated by the models are similar to the cases inside the original data set.

Since we have created three models, when we restore the file containing the learning data, the tool requests values of the previously selected attributes of these models to set up the profile to be queried. As we also inserted the original data set, the average values of each attribute are automatically filled inside each input box. We then simulate a situation

witch a patient is admitted in the hospital in a particular condition, but his or her attributes change dynamically over time, so we submitted three more condition states of the patient. Figure 27a shows the odds for the last queried state. Figure 27b shows all submitted sates, where we simulate that the patient underwent a surgical procedure, changed the state of complication, and naturally spend more few weeks hospitalized. In Figure 27c, a list presents the cases most correlated with the last submitted state. Finally, in Figure 27d, we see the streamgraph of the submitted states for each model. In the third streamgraph, it is possible to observe an increase in the chances of Sepsis, and it may represent a tendency for the patient to suffer from this condition, which could serve as support to shape the decisions of the healthcare team.

This scenario has shown that users can apply the tool to jointly identify outcomes as well as trends in the evolution of the characteristics of the object of interest. This feature assists the analyst not only in predicting current status but also in supporting complex decision-making situations. This aspect is difficult to find with the previous tools designed to explore regression models.

4.3.3 Scenario Three: Investigating the Novel COVID-19 Disease

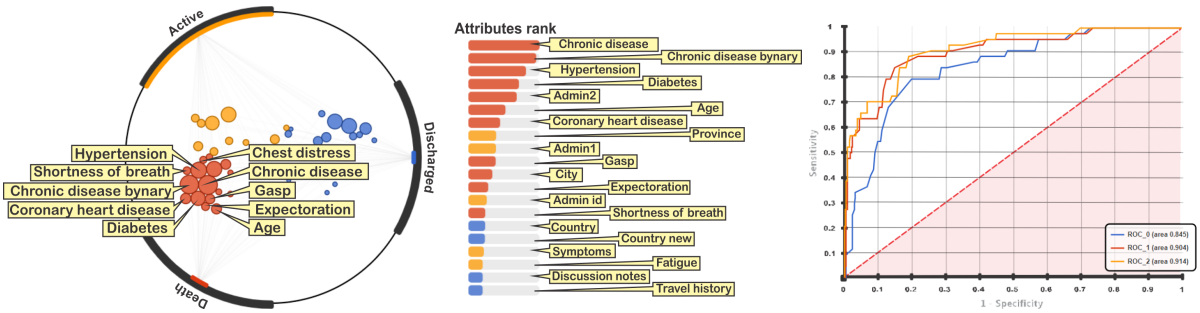
A recent outbreak of a virus, initially identified in the region of Wuhan, China, at the end of 2019, has spread and generated turbulence in the global community never seen before in modern human history. The World Health Organization (WHO) has officially named the virus as SARS-CoV-2, which means severe acute respiratory syndrome coronavirus 2. The infection caused by the virus has been named as coronavirus disease 2019 (COVID-19) also by the WHO (HE; DENG; LI, 2020). Since the spread outside Chinese borders, the global scientific community has been dedicating efforts to understand the virus and the infection that it causes.

In this scenario, we employed one of the freely available COVID-19 data sets to investigate the correlations of attributes and generate insights about the data. The focus is to understand how the onset symptoms, patient profile information, and chronic disease historic relate to the outcome of the treatment. Then, we built an LR model for the prediction of severe cases, and we compare results with the literature that outlines statistics of different patient profiles.

The data set is available from Kaggle[1] and obtained on March 31, 2020. It contains more than 33,000 reported cases. However, given our objective in this scenario, we have filtered out cases that do not specify age and symptoms. It reduced the data set to 1,586 occurrences, of which 56 cases resulted in death, and the others were either still active or cured until the date of download. We also decompose the symptoms attribute, which describes the set of symptoms for each report, into individual attributes for each of the most frequent symptoms.

[1] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

Figure 28 – Investigating the COVID-19 data set. (a) Attribute-RadViz shows correlations to the outcome labels; in addition, users can consult the dynamic rank. (b) Three binary LR models have been created; the blue one with only the "age" as an explainable attribute, the red one adds the set of attributes related to chronic diseases, and the yellow one includes all the previous ones plus the set of attributes related to the symptoms.



(a) Initial attribute view (b) ROC curves

Source: Elaborated by the author.

Initially, we want to understand the correlations of attributes in relation to the results of treatments; hence, we chose the "outcome" attribute as the target variable to start the tool. Figure 28a shows the initial rendering. We can quickly observe a group of attributes (the red ones) with potential to present high correlations with cases of death.

Investigating further, we realize our first relevant observation regarding this data set; the attributes that represent the history of chronic diseases have a significant correlation with cases of death, such as hypertension with a 0.38 estimated coefficient. On the other hand, the attributes that represent the initial symptoms of the disease have a medium-to-low or null correlation, even the symptoms considered more severe, such as shortness of breath, which has a 0.19 estimated coefficient.

Given the findings about the correlations of attributes, we have generated three LR models. The first model predicts death with only the "age" information. The second one contains the attributes: "Chronic disease binary", "Hypertension", "Diabetes", "Coronary heart disease", and "age". The last model includes all the attributes of the previous one plus the attributes that represent the most severe (for this data set) symptoms of the disease, which are: "Gasp", "Expectoration", "Shortness of breath", "Fatigue", and "Chest distress". Figure 28b shows the ROC curves generated by the models, wherewith only age information, the model has shown good predictive potential. The inclusion of attributes related to chronic diseases has shown a significant increase in the prediction capacity; however, the further inclusion of the symptoms attributes produces only a modest increase in the efficiency of the model for this data set. It emphasizes our first observation, in which the attributes related to the history of chronic diseases are relevant and have significant predictive power in comparison with attributes that represent symptoms of the disease.

Although most of the correlations found correspond to the patient's history and present

Table 4 – The estimated mortality rate caused by the COVID-19 disease distributed by age groups. The first record comprises the values determined by the analysis in (SURVEILLANCES, 2020). In the following, we have estimated rates simulating the absence and presence of comorbidities and severe symptoms.

	Age groups rate, %								
	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	≥ 80
Surveillances (2020)	–	0.2	0.2	0.2	0.4	1.3	3.6	8.0	14.8
Our LR model	–	0.1	0.1	0.3	0.6	1.3	2.9	6.3	13.1
Hypertension=1	–	0.1	0.1	0.3	0.8	1.8	4.2	9.5	19.8
Diabetes=1	0.2	0.5	1.1	2.3	5.1	10.7	21.1	37.3	57.0
Shortness of breath=1	0.1	0.1	0.3	0.7	1.5	3.3	7.2	14.8	28.1
Cough=1	0.1	0.1	0.3	0.6	1.4	3.0	6.6	13.6	26.1
Chest Distress=1	0.1	0.2	0.4	0.9	1.9	4.3	9.0	18.2	33.1

Source: Research data.

state, other significant correlations are also exposed by the tool such as “city”, “province”, and extra attributes related to the place of treatment. Factors as the local demographic characteristics, climatic condition, and medical care quality may generate some correlation of death to the geo-location of the patient (WANG *et al.*, 2020). In general, it is up to the analyst to investigate the validity of these findings. In the context of this scenario, the data set has been initially collected at the peak of the pandemic event in China, so most of the death cases come from the Wuhan region, but the data set includes many other active cases from different areas and countries, such as Japan and Italy. This condition generates a false correlation of death with the patient’s geo-location. Circumstances where hidden confounding attributes (such as the length of contagion of a region) create spurious correlations often can be detected by analysts, which highlights the importance of tools that insert the human into the process.

Inside the query tool, after building the LR model, we can observe the prediction of several profiles according to what the model learned from the data set. For comparison purposes, we attached results from a reference work in the analysis of reported cases collected in China on February 11, 2020 and presented by Surveillances (2020). Table 4 shows the acquired rates by age group. The rates exposed by the annexed work are very similar to those delivered by the LR model when ignoring comorbidities and severe symptoms. We also attached results by simulating the presence of two of the most frequent chronic diseases in the data set, as well as by simulating the presence of three different symptoms. All data we worked in this scenario, as well as the learning files, are freely available on GitHub [2]. If the reader is interested in simulating other situations, he or she can download the tool and load it on any web browser (we recommend Chrome or Firefox); there is no need for any installation process.

Much desired information is unavailable in the current freely available COVID-19 data sets; general data about the patient (such as smoke habits) and data about the provided treatment (such as the adopted drug administration) could substantially expand the investigation. However,

[2] https://github.com/erasmoartur/lrxptool

this scenario has shown that the approach is already capable of generating relevant insights and supporting possible decision-making tasks based on predictions using the resulting FS.

4.4 Final Remarks

In this chapter, we have presented an approach coupled with its portable and straightforward prototype for generating, applying , and evaluating regression models. The goal is to allow expert and non-expert users to create LR models with their data sets and make use of them in a unified tool. The approach works by initially exposing a panorama of correlations between attributes and a target events (usually outcomes), thus allowing users to choose good attributes for creating regression models. The approach also disposes of methods to visually evaluate the generated models; hence, the users can gain insights about the data and proceed with the production of the final model for later use.

Among the contributions of our work in this type of prediction, we highlight the development of a freely available web-based prototype for the transparent exploration of regression models. Also, report on the development of LoRRViz, an adapted RadViz tool to serve as a visual evaluation approach for multinomial LR models. It comes as an alternative to the classical ROC curves, or even to confusion matrices, since it generates a visual overview of how the model is classifying items.

At present, some limitations of our work still stand. The approach deals only with numerical attributes as independent attributes. Given the current high availability of categorical data, it would be interesting to make them available so that the user can perform regressions and consult the impact of these attributes on predictions. Another limitation is about the target attribute (dependent variable); it must encode meaningful states of the data set. Inadequately labeling may destabilize and produce imprecision to the generated models. Finally, the approach has great scalability potential on the number of attributes, since such attributes are plotted in a point-based visualization during the initial exploration phase; however, our prototype is implemented in Javascript with the D3js library, which does not imply the best performance option. Yet, in our tests, we were able to interactively generate models with hundreds of attributes with items in the ten thousand mark.

CHAPTER

5

CONCLUSIONS

In this Thesis, we have introduced interactive approaches to the analysis and selection of attributes as well as to the prediction of target ones. Such tasks are important since they allow analysts to explore data sets from the perspective of attributes and, furthermore, key attributes (often categorical ones) that encode relevant information about the data. Also, since these target attributes can be frequent in current data sets, a single data set can generate many data viewpoints. Hence, our proposed approaches hold attributes as first-order elements. The analysis core lies in the assembly of a correlation matrix, which is based on relationships of attributes and labels extracted from the target ones.

In chapter 3, we have presented a visual approach for attribute analysis and selection, including Attribute-RadViz, a correlation-based visualization built over the classic RadViz. The uniqueness of this work lies in the condensed design in which the relationships of the attributes and other data entities are presented, including the correlations with the potential data labels extracted from categorical attributes. The analyst can interactively conduct explorations from the projection of attributes combined with a set of interactions designed to enhance the RadViz technique. We also have developed a dual-view model to increase knowledge gain and provide an immediate evaluation of user-selected attributes. We have demonstrated through two case studies the performance of the approach in FS tasks and illustrated how the approach can lead to pertinent observations in the data set.

Taking advantage of the Attribute-RadViz to filter out redundant as well as irrelevant attributes, we have developed a second approach, presented in chapter 4, for creating, evaluating, and applying LR models. In the LR modeling, the excess of attributes increases the chances of de-stabilization of the models, mainly due to the problem of multicollinearity. LR models also tend to face overfitting problems when combined with multidimensional data. In our approach, users can explore the many possibilities of creating LR models guided by Attribute-RadViz and evaluate those models for later use. We have illustrated through three usage scenarios how analysts can gain insights about their data using the generated prediction models inside the

available implementation of the approach as a web-based tool.

5.1 Contributions

In summary, the main contributions of this Thesis are:

- An attribute visual analysis and selection approach and its fully working web-based prototype coupled with two case studies to validate and verify its effectiveness; It is shown that visual analysis can lead to more adequate matching between chosen attributes and actual patterns in the data effectively;
- Attribute-RadViz, an adaptation of the traditional RadViz to make it able to map attributes under the influence of labels; We have demonstrated how some of the original drawbacks of the standard RadViz can be overcome and also its applicability in comparing attributes rather than individual data items;
- An approach to explore LR models thought visual attribute analysis and its fully working web-based prototype coupled with three usage scenarios with real-world contexts showing its effectiveness;
- The LoRRViz, another RadViz adaptation that works jointly with a prediction model making it able to map data items according to their probability of belonging to each label rendered as DAs. We have shown its application as a visual evaluation mechanism of multinomial LR;
- Also, we have developed visual and interactive resources to enhance the traditional RadViz, such as dimensional arcs, positioning distortions control mechanisms, and its integration with a force scheme model.

5.2 Future Work

Three main aspects shape our intentions for future research; issues of scalability, support for a larger diversity of data types, and accuracy in the correlation estimates between mixed data. In the following, we list some topics that can delineate a further investigation:

- The FS tasks, despite allowing multi-selection, can be enhanced by the aid of automatic FS algorithms since users tend to select relevant subsets right at the beginning of the process. Automated methods can participate by suggesting selections or even complementing the FS made by the users to the subset size desired by them;
- Attribute-RadViz is capable of handling numerical and categorical data. Textual data in the TF–IDF format can also be manipulated in the most recent versions; however, there

is a lack of particular functions for this type of data such as topic detection through the observed clusters showing strongly correlated attributes;

- The correlation estimation between mixed data remains a challenge — the two models adopted in this Thesis try to determine the appropriate correlations in these mixed data. However, the investigation by more precise methods and how the proportion of each data type affects these estimates is still an open problem.

BIBLIOGRAPHY

AKADI, A. E.; OUARDIGHI, A. E.; ABOUTAJDINE, D. A powerful feature selection approach based on mutual information. **International Journal of Computer Science and Network Security**, v. 8, n. 4, p. 116–121, 2008. Citation on page 50.

ANGUITA, D.; GHIO, A.; ONETO, L.; PARRA, X.; REYES-ORTIZ, J. L. A public domain dataset for human activity recognition using smartphones. In: **ESANN**. [S.l.: s.n.], 2013. p. 437–442. ISBN 978-2-87419-081-0. Citation on page 48.

ANKERST, M.; BERCHTOLD, S.; KEIM, D. A. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In: **Proceedings IEEE Symposium on Information Visualization (Cat. No.98TB100258)**. [S.l.: s.n.], 1998. p. 52–60. ISSN null. Citation on page 39.

ARTUR, E.; MINGHIM, R. A novel visual approach for enhanced attribute analysis and selection. **Computers and Graphics**, v. 84, p. 160 – 172, 2019. ISSN 0097-8493. Citations on pages 19, 31, 34, 36, 40, 41, 43, 45, 47, and 51.

ASUNCION, A.; NEWMAN, D. **UCI machine learning repository**. 2007. Citations on pages 35 and 37.

BATTITI, R. Using mutual information for selecting features in supervised neural net learning. **IEEE Transactions on Neural Networks**, v. 5, n. 4, p. 537–550, July 1994. ISSN 1045-9227. Citation on page 50.

BERNARD, J.; STEIGER, M.; WIDMER, S.; LÜCKE-TIEKE, H.; MAY, T.; KOHLHAMMER, J. Visual-interactive exploration of interesting multivariate relations in mixed research data sets. **Computer Graphics Forum**, v. 33, n. 3, p. 291–300, 2014. Citations on pages 9, 25, and 30.

CHAMPION, H. R.; SACCO, W. J.; COPES, W. S.; GANN, D. S.; GENNARELLI, T. A.; FLANAGAN, M. E. A revision of the trauma score. **The Journal of trauma**, v. 29, n. 5, p. 623–629, 1989. Citation on page 57.

CHATTERJEE, S.; HADI, A. **Regression Analysis by Example**. Wiley, 2006. 234 p. (Wiley Series in Probability and Statistics). ISBN 9780470055458. Available: <https://books.google.com.br/books?id=uiu5XsAA9kYC>. Citation on page 18.

CHENG, S.; MUELLER, K. The data context map: Fusing data and attributes into a unified display. **IEEE Transactions on Visualization and Computer Graphics**, v. 22, n. 1, p. 121–130, Jan 2016. ISSN 1077-2626. Citations on pages 21 and 30.

CHENG, S.; XU, W.; MUELLER, K. Radviz deluxe: An attribute-aware display for multivariate data. **Processes**, Multidisciplinary Digital Publishing Institute, v. 5, n. 4, p. 75, 2017. Citation on page 34.

CHOO, J.; LEE, H.; KIHM, J.; PARK, H. ivisclassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In: **2010 IEEE Symposium on Visual Analytics Science and Technology**. [S.l.: s.n.], 2010. p. 27–34. Citations on pages 22 and 30.

CUI, Z.; BADAM, S. K.; YALçIN, M. A.; ELMQVIST, N. Datasite: Proactive visual data exploration with computation of insight-based recommendations. **Information Visualization**, v. 18, n. 2, p. 251–267, 2019. Available: <https://doi.org/10.1177/1473871618806555>. Citation on page 17.

DIEHL, S.; BECK, F.; BURCH, M. Uncovering strengths and weaknesses of radial visualizations—an empirical approach. **IEEE Transactions on Visualization and Computer Graphics**, v. 16, n. 6, p. 935–942, Nov 2010. ISSN 1077-2626. Citation on page 34.

DINGEN, D.; VEER, M. van't; HOUTHUIZEN, P.; MESTROM, E. H. J.; KORSTEN, E. H. H. M.; BOUWMAN, A. R. A.; WIJK, J. van. Regressionexplorer: Interactive exploration of logistic regression models with subgroup analysis. **IEEE Transactions on Visualization and Computer Graphics**, v. 25, n. 1, p. 246–255, Jan 2019. Citations on pages 28 and 30.

DOMINGUES, C.; COIMBRA, R.; POGGETTI, R. S.; NOGUEIRA, L. de S.; SOUSA, R. M. C. Performance of new adjustments to the triss equation model in developed and developing countries. **World journal of emergency surgery**, BioMed Central, v. 12, n. 1, p. 17, 2017. ISSN 1749-7922. Citations on pages 47 and 64.

DOMINGUES, C.; SOUSA, R. M. C. d.; NOGUEIRA, L. d. S.; POGGETTI, R. S.; FONTES, B.; MUÑOZ, D. The role of the new trauma and injury severity score (ntriss) for survival prediction. **Revista da Escola de Enfermagem da USP**, SciELO Brasil, v. 45, n. 6, p. 1353–1358, 2011. Citation on page 46.

DRAPER, G. M.; LIVNAT, Y.; RIESENFELD, R. F. A survey of radial methods for information visualization. **IEEE Transactions on Visualization and Computer Graphics**, v. 15, n. 5, p. 759–776, Sep. 2009. ISSN 1077-2626. Citation on page 34.

DREISEITL, S.; OHNO-MACHADO, L. Logistic regression and artificial neural network classification models: a methodology review. **Journal of Biomedical Informatics**, v. 35, n. 5, p. 352 – 359, 2002. ISSN 1532-0464. Available: <http://www.sciencedirect.com/science/article/pii/S1532046403000340>. Citation on page 55.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification (2Nd Edition)**. New York, NY, USA: Wiley-Interscience, 2000. ISBN 0471056693. Citation on page 50.

DY, J. G.; BRODLEY, C. E. Visualization and interactive feature selection for unsupervised data. In: **Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2000. (KDD '00), p. 360–364. ISBN 1-58113-233-6. Citations on pages 22 and 30.

ELZEN, S. V. D.; WIJK, J. J. van. Baobabview: Interactive construction and analysis of decision trees. In: IEEE. **Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on**. [S.l.], 2011. p. 151–160. Citations on pages 27 and 30.

FLEURET, F. Fast binary feature selection with conditional mutual information. **J. Mach. Learn. Res.**, JMLR.org, v. 5, p. 1531–1555, Dec. 2004. ISSN 1532-4435. Citation on page 50.

GINI, C. **Variabilità e mutabilità**. [S.l.]: Libreria Eredi Virgilio Veschi, Rome, 1912. Reprinted in Memorie di metodologia statistica (Edited by Pizetti, E. Salvemini, T.). Citation on page 50.

GLEICHER, M. Explainers: Expert explorations with crafted projections. **IEEE Transactions on Visualization and Computer Graphics**, v. 19, n. 12, p. 2042–2051, Dec 2013. ISSN 1077-2626. Citations on pages 23 and 30.

GOLDSTEIN, B. A.; NAVAR, A. M.; PENCINA, M. J.; IOANNIDIS, J. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. **Journal of the American Medical Informatics Association**, Oxford University Press, v. 24, n. 1, p. 198–208, 2017. Citation on page 55.

GONÇALVES, L.; SUBTIL, A.; OLIVEIRA, M. R.; BERMUDEZ, P. Roc curve estimation: An overview. **REVSTAT–Statistical Journal**, v. 12, n. 1, p. 1–20, 2014. Citation on page 58.

GREINER, M.; PFEIFFER, D.; SMITH, R. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. **Preventive veterinary medicine**, Elsevier, v. 45, n. 1-2, p. 23–41, 2000. Citation on page 56.

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. **Radiology**, v. 143, n. 1, p. 29–36, 1982. Citation on page 56.

HARRELL, F. **Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis**. Springer International Publishing, 2015. (Springer Series in Statistics). ISBN 9783319194257. Available: <https://books.google.com.br/books?id=94RgCgAAQBAJ>. Citation on page 29.

HE, F.; DENG, Y.; LI, W. Coronavirus disease 2019: What we know? **Journal of Medical Virology**, 2020. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jmv.25766>. Citation on page 68.

HOFFMAN, P.; GRINSTEIN, G.; MARX, K.; GROSSE, I.; STANLEY, E. Dna visual and analytic data mining. In: **Proceedings. Visualization '97 (Cat. No. 97CB36155)**. [S.l.: s.n.], 1997. p. 437–441. Citation on page 32.

HOFFMAN, P.; GRINSTEIN, G.; PINKNEY, D. Dimensional anchors: A graphic primitive for multidimensional multivariate information visualizations. In: **Proceedings of the 1999 Workshop on New Paradigms in Information Visualization and Manipulation in Conjunction with the Eighth ACM Internation Conference on Information and Knowledge Management**. New York, NY, USA: ACM, 1999. (NPIVM '99), p. 9–16. ISBN 1-58113-254-9. Citation on page 32.

INGRAM, S.; MUNZNER, T.; IRVINE, V.; TORY, M.; S., B.; MÖLLER, T. Dimstiller: Workflows for dimensional analysis and reduction. In: **2010 IEEE Symposium on Visual Analytics Science and Technology**. [S.l.: s.n.], 2010. p. 3–10. Citations on pages 22 and 30.

JÄCKLE, D.; HUND, M.; BEHRISCH, M.; KEIM, D. A.; SCHRECK, T. Pattern trails: Visual analysis of pattern transitions in subspaces. In: **2017 IEEE Conference on Visual Analytics Science and Technology (VAST)**. [S.l.: s.n.], 2017. p. 1–12. Citations on pages 23 and 30.

JOHANSSON, S.; JOHANSSON, J. Interactive dimensionality reduction through user-defined combinations of quality metrics. **IEEE Transactions on Visualization and Computer Graphics**, v. 15, n. 6, p. 993–1000, Nov 2009. ISSN 1077-2626. Citations on pages 22, 23, and 30.

JOLLIFFE, I. Principal component analysis. In: ____. **International Encyclopedia of Statistical Science**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 1094–1096. ISBN 978-3-642-04898-2. Citation on page 35.

JÚNIOR, G. A. P.; SCARPELINI, S.; BASILE-FILHO, A.; ANDRADE, J. I. de. Índices de trauma. **Medicina (Ribeirao Preto. Online)**, v. 32, n. 3, p. 237–250, 1999. Citation on page 44.

KEIM, D. A.; MANSMANN, F.; THOMAS, J. Visual analytics: How much visualization and how much analytics? **SIGKDD Explor. Newsl.**, Association for Computing Machinery, New York, NY, USA, v. 11, n. 2, p. 5–8, May 2010. ISSN 1931-0145. Available: <https://doi.org/10.1145/1809400.1809403>. Citations on pages 17 and 32.

KLEMM, P.; LAWONN, K.; GLASSER, S.; NIEMANN, U.; HEGENSCHEID, K.; VÖLZKE, H.; PREIM, B. 3d regression heat map analysis of population study data. **IEEE Transactions on Visualization and Computer Graphics**, v. 22, n. 1, p. 81–90, Jan 2016. ISSN 1077-2626. Citations on pages 28 and 30.

KOBOUROV, S. G. Spring embedders and force directed graph drawing algorithms. **CoRR**, abs/1201.3011, 2012. Citation on page 39.

KRAUSE, J.; PERER, A.; BERTINI, E. Infuse: Interactive feature selection for predictive modeling of high dimensional data. **IEEE Transactions on Visualization and Computer Graphics**, v. 20, n. 12, p. 1614–1623, Dec 2014. ISSN 1077-2626. Citations on pages 24, 25, and 30.

LI, J.; CHENG, K.; WANG, S.; MORSTATTER, F.; TREVINO, R. P.; TANG, J.; LIU, H. Feature selection: A data perspective. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 50, n. 6, p. 94:1–94:45, Dec. 2017. ISSN 0360-0300. Citation on page 50.

LI, J.; WANG, J. Z. Automatic linguistic indexing of pictures by a statistical modeling approach. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 25, n. 9, p. 1075–1088, Sep. 2003. ISSN 0162-8828. Citation on page 48.

LIN, D.; TANG, X. Conditional infomax learning: An integrated framework for feature extraction and fusion. In: **Computer Vision – ECCV 2006**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 68–82. ISBN 978-3-540-33833-8. Citation on page 50.

LIU, H.; SETIONO, R. Chi2: feature selection and discretization of numeric attributes. In: **Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence**. [S.l.: s.n.], 1995. p. 388–391. ISSN 1082-3409. Citation on page 50.

LIU, S.; WANG, B.; THIAGARAJAN, J. J.; BREMER, P.-T.; PASCUCCI, V. Visual exploration of high-dimensional data through subspace analysis and dynamic projections. **Computer Graphics Forum**, v. 34, n. 3, p. 271–280, 2015. Citations on pages 23 and 30.

LIU, Y.; ZHENG, Y. F. One-against-all multi-class svm classification using reliability measures. In: **Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.** [S.l.: s.n.], 2005. v. 2, p. 849–854 vol. 2. ISSN 2161-4393. Citation on page 35.

MAATEN, L. v. d.; HINTON, G. Visualizing data using t-sne. **Journal of machine learning research**, v. 9, n. Nov, p. 2579–2605, 2008. Citation on page 35.

MACKINLAY, J. Automating the design of graphical presentations of relational information. **ACM Trans. Graph.**, ACM, New York, NY, USA, v. 5, n. 2, p. 110–141, Apr. 1986. ISSN 0730-0301. Citation on page 38.

MANLY, B. F.; ALBERTO, J. A. N. **Multivariate statistical methods: a primer**. [S.l.]: CRC Press, 2016. Citation on page 55.

MAY, T.; BANNACH, A.; DAVEY, J.; RUPPERT, T.; KOHLHAMMER, J. Guiding feature subset selection with an interactive visualization. In: **2011 IEEE Conference on Visual Analytics Science and Technology (VAST)**. [S.l.: s.n.], 2011. p. 111–120. Citations on pages 25 and 30.

MCKENNA, S.; MEYER, M.; GREGG, C.; GERBER, S. s-corrplot: An interactive scatterplot for exploring correlation. **Journal of Computational and Graphical Statistics**, Taylor & Francis, v. 25, n. 2, p. 445–463, 2016. Citations on pages 23 and 30.

METZ, C. E. Basic principles of roc analysis. In: ELSEVIER. **Seminars in nuclear medicine**. [S.l.], 1978. v. 8, p. 283–298. Citation on page 56.

MÜHLBACHER, T.; PIRINGER, H. A partition-based framework for building and validating regression models. **IEEE Transactions on Visualization and Computer Graphics**, v. 19, n. 12, p. 1962–1971, Dec 2013. ISSN 1077-2626. Citations on pages 28 and 30.

NIE, F.; XIANG, S.; JIA, Y.; ZHANG, C.; YAN, S. Trace ratio criterion for feature selection. In: **Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2**. [S.l.]: AAAI Press, 2008. (AAAI'08), p. 671–676. ISBN 978-1-57735-368-3. Citation on page 50.

ONO, J. H. P.; SIKANSI, F.; CORRêA, D. C.; PAULOVICH, F. V.; PAIVA, A.; NONATO, L. G. Concentric radviz: Visual exploration of multi-task classification. In: **2015 28th SIBGRAPI Conference on Graphics, Patterns and Images**. [S.l.: s.n.], 2015. p. 165–172. ISSN 2377-5416. Citation on page 34.

PENG, H.; LONG, F.; DING, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 27, n. 8, p. 1226–1238, Aug 2005. ISSN 0162-8828. Citation on page 50.

RAUBER, P. E.; FALCAO, A. X.; TELEA, A. C. Projections as visual aids for classification system design. **Information Visualization**, 2017. Citations on pages 27 and 30.

ROBNIK-ŠIKONJA, M.; KONONENKO, I. Theoretical and empirical analysis of relieff and rrelieff. **Machine learning**, Springer, v. 53, n. 1-2, p. 23–69, 2003. Citation on page 50.

SANCHEZ, A.; SOGUERO-RUIZ, C.; MORA-JIMéNEZ, I.; RIVAS-FLORES, F.; LEHMANN, D.; RUBIO-SÁNCHEZ, M. Scaled radial axes for interactive visual feature selection: A case study for analyzing chronic conditions. **Expert Systems with Applications**, v. 100, p. 182 – 196, 2018. ISSN 0957-4174. Citations on pages 26 and 30.

SEO, J.; SHNEIDERMAN, B. A rank-by-feature framework for interactive exploration of multidimensional data. **Information Visualization**, Palgrave Macmillan, v. 4, n. 2, p. 96–113, Jul. 2005. ISSN 1473-8716. Citations on pages 23, 24, and 30.

SHARKO, J.; GRINSTEIN, G.; MARX, K. A. Vectorized radviz and its application to multiple cluster datasets. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 6, p. 1444–1427, Nov 2008. ISSN 1077-2626. Citation on page 34.

SHEVADE, S. K.; KEERTHI, S. S. A simple and efficient algorithm for gene selection using sparse logistic regression. **Bioinformatics**, v. 19, n. 17, p. 2246–2253, 11 2003. ISSN 1367-4803. Available: <https://doi.org/10.1093/bioinformatics/btg308>. Citation on page 57.

SURVEILLANCES, V. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19)—china, 2020. **China CDC Weekly**, v. 2, n. 8, p. 113–122, 2020. Citation on page 70.

TABACHNICK, B. G.; FIDELL, L. S. **Using Multivariate Statistics (5th Edition)**. Needham Heights, MA, USA: Allyn & Bacon, Inc., 2006. ISBN 0205459382. Citation on page 36.

TATU, A.; ALBUQUERQUE, G.; EISEMANN, M.; BAK, P.; THEISEL, H.; MAGNOR, M.; KEIM, D. Automated analytical methods to support visual exploration of high-dimensional data. **IEEE Transactions on Visualization and Computer Graphics**, v. 17, n. 5, p. 584–597, May 2011. ISSN 1077-2626. Citations on pages 23 and 30.

TATU, A.; MAASS, F.; FÄRBER, I.; BERTINI, E.; SCHRECK, T.; SEIDL, T.; KEIM, D. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In: IEEE. **Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on**. [S.l.], 2012. p. 63–72. Citations on pages 22 and 30.

TSOUMAKAS, G.; KATAKIS, I. Multi-label classification: An overview. **International Journal of Data Warehousing and Mining**, v. 3, n. 3, 2006. Citation on page 35.

TURKAY, C.; FILZMOSER, P.; HAUSER, H. Brushing dimensions - a dual visual analysis model for high-dimensional data. **IEEE Transactions on Visualization and Computer Graphics**, v. 17, n. 12, p. 2591–2599, Dec 2011. ISSN 2160-9306. Citations on pages 17, 26, and 30.

TURKAY, C.; LEX, A.; STREIT, M.; PFISTER, H.; HAUSER, H. Characterizing cancer subtypes using dual analysis in caleydo stratomex. **IEEE Computer Graphics and Applications**, v. 34, n. 2, p. 38–47, Mar 2014. ISSN 0272-1716. Citation on page 27.

WANG, J.; TANG, K.; FENG, K.; LV, W. High temperature and high humidity reduce the transmission of covid-19. **SSRN Electronic Journal**, Elsevier BV, 2020. ISSN 1556-5068. Available: <http://dx.doi.org/10.2139/ssrn.3551767>. Citation on page 70.

WANG, Y.; LI, J.; NIE, F.; THEISEL, H.; GONG, M.; LEHMANN, D. J. Linear discriminative star coordinates for exploring class and cluster separation of high dimensional data. **Comput. Graph. Forum**, The Eurographs Association & John Wiley & Sons, Ltd., Chichester, UK, v. 36, n. 3, p. 401–410, Jun. 2017. ISSN 0167-7055. Citations on pages 26 and 30.

WARE, M.; FRANK, E.; HOLMES, G.; HALL, M.; WITTEN, I. H. Interactive machine learning: letting users build classifiers. **International Journal of Human-Computer Studies**, v. 55, n. 3, p. 281 – 292, 2001. ISSN 1071-5819. Available: <http://www.sciencedirect.com/science/article/pii/S1071581901904999>. Citation on page 18.

WRIGHT, S. The interpretation of population structure by f-statistics with special regard to systems of mating. **Evolution**, v. 19, n. 3, p. 395–420, 1965. Citation on page 50.

YANG, H. H.; MOODY, J. Data visualization and feature selection: New algorithms for non-gaussian data. In: **Proceedings of the 12th International Conference on Neural Information Processing Systems**. Cambridge, MA, USA: MIT Press, 1999. (NIPS'99), p. 687–693. Citation on page 50.

YANG, J.; WARD, M. O.; RUNDENSTEINER, E. A.; HUANG, S. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In: **Proceedings of the Symposium on Data Visualisation 2003**. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2003. (VISSYM '03), p. 19–28. ISBN 1-58113-698-6. Citations on pages 21, 22, and 30.

YUAN, X.; REN, D.; WANG, Z.; GUO, C. Dimension projection matrix/tree: Interactive subspace visual exploration and analysis of high dimensional data. **IEEE Transactions on Visualization and Computer Graphics**, v. 19, n. 12, p. 2625–2633, Dec 2013. ISSN 1077-2626. Citations on pages 27 and 30.

ZHANG, C.; YANG, J.; ZHAN, F. B.; GONG, X.; BRENDER, J. D.; LANGLOIS, P. H.; BARLOWE, S.; ZHAO, Y. A visual analytics approach to high-dimensional logistic regression modeling and its application to an environmental health study. In: **2016 IEEE Pacific Visualization Symposium (PacificVis)**. [S.l.: s.n.], 2016. p. 136–143. Citations on pages 28 and 30.

ZHANG, Z.; MCDONNELL, K. T.; ZADOK, E.; MUELLER, K. Visual correlation analysis of numerical and categorical data on the correlation map. **IEEE Transactions on Visualization and Computer Graphics**, v. 21, n. 2, p. 289–303, Feb 2015. ISSN 1077-2626. Citations on pages 37 and 38.

ZHENG, Y.; WU, W.; CHEN, Y.; QU, H.; NI, L. M. Visual analytics in urban computing: An overview. **IEEE Transactions on Big Data**, v. 2, n. 3, p. 276–296, Sep. 2016. ISSN 2332-7790. Citation on page 18.

ZHOU, F.; HUANG, W.; LI, J.; HUANG, Y.; SHI, Y.; ZHAO, Y. Extending dimensions in radviz based on mean shift. In: **2015 IEEE Pacific Visualization Symposium (PacificVis)**. [S.l.: s.n.], 2015. p. 111–115. ISSN 2165-8765. Citation on page 34.

ZHOU, F.; LI, J.; HUANG, W.; ZHAO, Y.; YUAN, X.; LIANG, X.; SHI, Y. Dimension reconstruction for visual exploration of subspace clusters in high-dimensional data. In: **2016 IEEE Pacific Visualization Symposium (PacificVis)**. [S.l.: s.n.], 2016. p. 128–135. ISSN 2165-8773. Citations on pages 23 and 30.

ZHOU, X.-h.; MCCLISH, D. K. **Statistical methods in diagnostic medicine /**. 2nd ed.. ed. Hoboken, N.J. :: Wiley,, 2011. (Wiley series in probability and statistics). Citation on page 59.

APPENDIX

A

RELEVANT PSEUDOCODES

In the following, we present the pseudocodes of algorithms described in this project. Algorithm 1 describes the construction of the correlation matrix between attributes and presence vectors that represent the data labels. Algorithm 2 illustrates the greedy method used to order the attributes in the form of DAs in the item view. Lastly, Algorithm 3 presents the ordering method of labels represented by DAs in Attribute-RadViz.

Algorithm 1 – Attribute vs label correlation matrix building

```
1:  procedure BUILDMATRIX(D, L, t)        ▷ Data matrix, attribute list, target attribute index
2:      T                                  ▷ Target attribute
3:      U                                  ▷ Unique label values
4:      B                                  ▷ Binary presence vectors
5:      CoM                                ▷ The correlation matrix
6:      D ← trasnpose(D)                   ▷ Transposes to handle attributes
7:      T ← D.splice(t, 1)                 ▷ Splits the target attribute from the rest
8:      U ← uniques(T)                     ▷ Collecting unique values inside the target
9:      for i ← 0, U.length do
10:         for j ← 0, T.length do         ▷ Building the binary presence vectors
11:             if T_j = U_i then
12:                 B_ij ← 1
13:             else
14:                 B_ij ← 0
15:     B ← transpose(B)                   ▷ Also transposing
16:     for i ← 0, D.length do
17:         for j ← 0, B.length do         ▷ Building the correlation matrix
18:             CoM_ij ← getCorrelation(D_i, B_i)
19:     return CoM
```

Algorithm 2 – Greedy DA ordering for the item view

1: **procedure** SORTATTRIBUTEDAS(D, L) ▷ Selected attributes data, attribute list
2: L' ▷ New ordered attribute list
3: CoM ▷ Correlation matrix of the selected attributes
4: **for** $i \leftarrow 0, L.length$ **do** ▷ Building the correlation matrix
5: **for** $j \leftarrow 0, i$ **do**
6: $CoM_{ij} \leftarrow PersonCorrelation(D_i, D_j)$
7: $L'_0 \leftarrow L_{CoM.indexOf(max(CoM))}$ ▷ Highest correlation as first element
8: **for** $i \leftarrow 1, L.length$ **do** ▷ For each remaining label
9: $mMax, nMax, maxValue \leftarrow 0$
10: **for** $n \leftarrow 0, L.length$ **do** ▷ Search the next highest correlation
11: **for** $m \leftarrow 0, n$ **do** ▷ among the included ones
12: $a \leftarrow L_n \in L'$? ▷ L_n is already placed?
13: $b \leftarrow L_m \in L'$? ▷ L_m is already placed?
14: **if** $(maxValue < abs(CoM_{nm}))\&(\text{a}\oplus\text{b})$ **then**
15: $maxValue = abs(CoM_{nm})$
16: $mMax = m$
17: $nMax = n$
18: **if** $CoM_{mMax,nMax} > 0$ **then** ▷ Correlation direction test
19: $L' \leftarrow placeClose(L_{(mMax)}, L_{(nMax)})$ ▷ Place them close
20: **else**
21: $L' \leftarrow placeFar(L_{(mMax)}, L_{(nMAx)})$ ▷ Place them far
22: **return** L'

Algorithm 3 – Greedy DA ordering for the attribute view

1: **procedure** SORTLABELDAS(D, L, t) ▷ Data matrix, label list, and target attribute index
2: C ▷ The data matrix split by the categories
3: I ▷ Representative item list
4: CoM ▷ Correlation matrix of the representative items
5: L' ▷ New ordered label list
6: **for** $i \leftarrow 0, D.length$ **do** ▷ Spliting data by categories
7: $C_{L.indexOf(D_{it})}.add(D_i)$ ▷ Splitting data by category
8: **for** $i \leftarrow 0, L.length$ **do** ▷ Getting the medoid of each category
9: $I_i \leftarrow getMedoid(C_i)$
10: **for** $i \leftarrow 0, L.length$ **do** ▷ Building the correlation matrix
11: **for** $j \leftarrow 0, i$ **do**
12: $CoM_{ij} \leftarrow PersonCorrelation(I_i, I_j)$
13: $L'_0 \leftarrow L_{CoM.indexOf(max(CoM))}$ ▷ Highest correlation as first element
14: **for** $i \leftarrow 1, L.length$ **do** ▷ For each remaining label
15: $mMax, nMax, maxValue \leftarrow 0$
16: **for** $n \leftarrow 0, L.length$ **do** ▷ Search the next highest correlation
17: **for** $m \leftarrow 0, n$ **do** ▷ among the included ones
18: $a \leftarrow L_n \in L'$? ▷ L_n is already placed?
19: $b \leftarrow L_m \in L'$? ▷ L_m is already placed?
20: **if** $(maxValue < abs(CoM_{nm}))\&(\mathrm{a} \oplus \mathrm{b})$ **then**
21: $maxValue = abs(CoM_{nm})$
22: $mMax = m$
23: $nMax = n$
24: **if** $CoM_{mMax,nMax} > 0$ **then** ▷ Correlation direction test
25: $L' \leftarrow placeClose(L_{(mMax)}, L_{(nMax)})$ ▷ Place them close
26: **else**
27: $L' \leftarrow placeFar(L_{(mMax)}, L_{(nMAx)})$ ▷ Place them far
28: **return** L'


ICMC
USP
SÃO CARLOS