Ferramentas Visuais para Análise e Seleção de Atributos

Erasmo Artur da Silva Júnior

Erasmo Artur da Silva Júnior

Ferramentas Visuais para Análise e Seleção de Atributos

Monografia apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, para o Exame de Qualificação, como parte dos requisitos para obtenção do título de Doutor em Ciências – Ciências de Computação e Matemática Computacional.

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Rosane Minghim

USP – São Carlos
de 2017

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

Júnior, Erasmo Artur da Silva
J634f Ferramentas Visuais para Análise e Seleção de Atributos / Erasmo Artur da Silva Júnior; orientadora Rosane Minghim. - São Carlos - SP, 2017.
59 p.

Monografia (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2017.

1. Seleção de atributos visual. 2. visualização do espaço de características. I. Minghim, Rosane, orient. II. Título.

Erasmo Artur da Silva Júnior

Visual Tools for Feature Analysis and Selection

Monograph submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP, as part of the qualifying exam requisites of the Doctorate Program in Computer Science and Computational Mathematics.

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Rosane Minghim

USP – São Carlos
2017

RESUMO

SILVA JR., E. A.. **Ferramentas Visuais para Análise e Seleção de Atributos**. 2017. 59 f. Monografia (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Dados multidimensionais têm sido foco de várias pesquisas em virtude do grande volume de informações coletado nos dias atuais. A seleção de atributos representa parte deste esforço, conduzindo métodos que reduzem a dimensionalidade dos dados sem perda de informações pertinentes. O objetivo central é escolher o subconjunto ótimo de atributos, mantendo a representatividade dos dados com uma quantidade mínima de dimensões. Vários métodos de seleção de atributos foram propostos, porém, poucos exploram esta etapa para, além de realizar a seleção, promover descoberta de conhecimento a partir da exploração dos relacionamentos entre atributos, instâncias e classes. Uma das formas de se explorar o espaço de características e, em tempo, empregar a expertise do usuário para exercer a seleção, é realizá-la de maneira interativa com suporte de técnicas de visualização. Este trabalho propõe uma abordagem para análise e seleção interativa de atributos baseada em correlação e apoiada na técnica de projeção multidimensional. O principal objetivo é revelar informações pertinentes no espaço de características e, ao mesmo tempo, promover suporte para que o usuário realize a seleção de atributos de qualidade similar ou superior aos algoritmos automáticos.

Palavras-chave: Seleção de atributos visual, visualização do espaço de características.

ABSTRACT

SILVA JR., E. A.. **Ferramentas Visuais para Análise e Seleção de Atributos**. 2017. 59 f. Monografia (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação (ICMC/USP), São Carlos – SP.

Multidimensional data have been the focus of several researches due to the large volume of information collected nowadays. Feature selection represents part of this effort, leading to methods that reduce the data dimensionality without loss of pertinent information. The central objective is to choose the optimal subset of features, maintaining the representativeness of the data with a minimum amount of dimensions. Several feature selection methods have been proposed, however few exploit this step in order to promote knowledge discovery by the exploration of relationships between features, instances and classes. One way to explore the feature space and in the same time to employ the user expertise in the feature selection step is to perform it in an interactive mode with visualization techniques support. This work proposes an interactive feature selection approach based on correlation and supported by the point-based multidimensional projection technique. The main objective of the approach is to reveal relevant information in the feature space and at the same time to promote support for the user to perform the feature selection with similar or superior quality relative to automatic algorithms.

Key-words: Visual feature selection, Feature analysis, Feature space visualization.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CIFE ..... *Conditional Infomax Feature Extraction*
CMIN .... *Conditional Mutual Information Maximization*
DA ..... Âncoras Dimensionais
FCBF .... *Fast Correlation Based Filter*
IMOS .... insuficiência de múltiplos órgãos e sistemas
IRA ..... insuficiencia respiratoria aguda
JMI ..... *Joint Mutual Information*
MIFS ..... *Mutual Information Feature Selection*
mRMR ... *Minimum Redundancy Maximum Relevance*
NISS ..... *New Injury Severity Score*
NTRIS ... *New Trauma and Injury Severity*
RadViz ... *Radial Coordinate Visualization*
RFE ..... *Recursive Feature Elimination*
SARA .... síndrome da angústia respiratória aguda
SBFS ..... *sequential backward floating selection*
SBS ..... *sequential backward selection*
SFFS ..... *sequential forward floating selection*
SFS ..... *sequential forward selection*
SVM ..... *Support Vector Machine*
TCE ..... trauma encéfalo craniano
VHDR ... *Visual Hierarchical Dimension Reduction*
Visual-FSSEM *Visual Feature Subset Selection using Expectation-Maximization Clustering*

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

Enquanto as capacidades de coleta e armazenamento de dados crescem rapidamente nos dias de hoje, a habilidade do homem em processar e analisar essas informações aumenta em ritmo mais lento. Essa assincronia abre uma lacuna que gera novos desafios dentro dos processos que realizam computação com esses dados (KEIM *et al.*, 2006). Esforços significativos em pesquisas vêm sendo realizado para o desenvolvimento de ferramentas capazes de analisar e, eventualmente, sumarizar eficientemente estes dados, tanto para exposição ao usuário quanto para emprego em processos subsequentes.

Frequentemente essas bases de dados se apresentam de forma não estruturada ou semi-estruturada, demandando métodos específicos para seu tratamento. Um modelo de estrutura bastante simples, e largamente empregada, é a representação de exemplos no formato atributo-valor, onde cada exemplo possui um conjunto de m atributos, e, considerando cada atributo como uma dimensão de um vetor, tem-se então uma coleção de exemplos representados por vetores multi-dimensionais. O problema deste tipo de representação é que sua interpretação nem sempre é trivial e a alta dimensionalidade normalmente reflete em aumento da complexidade dos métodos que realizam algum processamento nestes dados.

Quanto maior a dimensionalidade dos dados, maior a complexidade de sua análise, como também maiores são as chances de vários atributos representarem obstáculos à tentativa de extração de informação pertinente perante algum critério desejado. Estes contratempos podem ser provenientes tanto da redundância de atributos, quanto da baixa relevância que alguns destes podem representar. Dados de alta dimensionalidade aumentam significativamente os requisitos de armazenamento de memória e os custos computacionais para análise e computação dos dados (LI *et al.*, 2016).

Uma alternativa para contornar o problema da complexidade de compreensão de dados multidimensionais é proporcionar ferramentas de apoio visual à análise os dados. Valiosas informações frequentemente surgem diante da observação das inter-relações existentes entre

atributos (ou variáveis) (ZHANG *et al.*, 2015). Nesse contexto, a análise visual dos dados no espaço de atributos representa etapa relevante na tarefa de revelar ao usuário informações que surgem dos relacionamentos entre os atributos e demais entidades dos dados. Uma boa representação visual do espaço de atributos, além de transmitir conhecimento, permite que o usuário realize uma seleção de atributos de maneira satisfatória.

A seleção de atributos busca a escolha de um subconjunto de variáveis que melhor represente o conjunto de dados diante da classificação/clusterização estimada (DASH; LIU, 1997). Vários benefícios podem ser extraídos desta etapa, como a ordenação de atributos a partir de algum critério, definição de pesos concedendo diferentes níveis de importância aos atributos, redução da dimensionalidade original do conjunto de dados, consequente aumento do desempenho dos processos subsequentes, bem como incremento da acurácia da etapa de classificação/clusterização (GUYON; ELISSEEFF, 2003).

A seleção de atributos pode ser classificada de acordo com a implementação de suas estratégias para encontro do subconjunto ideal. Em (LIU; MOTODA, 1998), os algoritmos são agrupados de acordo com três parâmetros, sendo eles: estratégia de busca, direção de busca e medida de avaliação, cada uma delas possuindo três principais abordagens. Em (CHANDRASHEKAR; SAHIN, 2014) os métodos são divididos de acordo com seu princípio de funcionamento, podendo ser: filtro, *wrapper* ou *embedded*. O capítulo 2 descreve a taxonomia dos métodos com mais detalhes.

Diversos métodos automáticos para seleção de atributos foram propostos, como em (HOQUE; BHATTACHARYYA; KALITA, 2014), (YU; LIU, 2003), (GU; LI; HAN, 2012), (HE; CAI; NIYOGI, 2005) e (ROBNIK-ŠIKONJA; KONONENKO, 2003), porém, o isolamento do usuário nesta etapa pode implicar na perda da oportunidade de revelar informações relevantes inerentes ao espaço de características. O usuário pode ser visto como peça chave na melhoria da seleção de atributos. Em (RAGHAVAN; MADANI; JONES, 2005), os autores demonstram que ele consegue apontar a relevância dos atributos de forma bastante satisfatória. Para viabilizar a interatividade com usuário e ao mesmo tempo tirar proveito da capacidade deste de reconhecer padrões, ferramentas visuais representam boas alternativas para suporte na seleção de atributos.

A análise visual no espaço de atributos busca fornecer ao usuário informações de relevância dos atributos, o relacionamento entre eles, o relacionamento com as demais entidades dos dados e a incidência de redundância, tornando tão explicita quanto possível a perspectiva de seleção de subconjuntos ótimos. Uma boa representação visual, além de habilitar até mesmo usuários leigos diante do conjunto de dados a realizarem seleção de atributos relevante, também incrementa o nível de interpretação dos dados, uma vez que os relacionamentos entre instâncias, atributos e classes são visualmente evidenciados.

A inserção humana na análise e seleção de atributos permite adicionar ao processo flexibilidade, criatividade e conhecimento tácito não presentes nos métodos automáticos. Os usuários podem concentrar em suas capacidades cognitivas e perceptivas completas no processo

analítico, ao mesmo tempo em que lhes permite aplicar capacidades computacionais avançadas para aumentar o processo de descoberta (KEIM *et al.*, 2006).

Como suporte visual para este trabalho, a projeção empregada para visualização dos dados multidimensionais é a *Radial Coordinate Visualization* (RadViz) (HOFFMAN *et al.*, 1997)(HOFFMAN; GRINSTEIN; PINKNEY, 1999). Nela, as dimensões são pontos (chamados âncoras dimensionais ou DAs) em torno de uma circunferência, e cada ponto localizado no interior desta representa um registro, ou conjunto de registros, do conjunto de dados. A localização de cada ponto dentro do círculo é calculada como uma função de sua atração relativa, semelhante a uma mola, às âncoras dimensionais (SHARKO; GRINSTEIN; MARX, 2008). O diferencial desta projeção está na riqueza semântica na exposição da relação entre instâncias e atributos, fornecendo uma visão global sobre os dados multidimensionais.

Mudanças significativas são realizadas na projeção RadViz proposta por este trabalho. A princípio, os rótulos das instâncias são separados por classe a partir da abordagem inspirada no método de relevância binária (TSOUMAKAS; KATAKIS, 2007). Em seguida, é feita a montagem da matriz de correlação entre atributos e rótulos individuais, assim como entre atributos e o vetor de rótulos completo. Por fim, a matriz é transposta, tornando os rótulos como as âncoras dimensionais na projeção, e os atributos são representados pelos elementos projetados.

Com intuito de aumentar a capacidade cognitiva do usuário diante da abordagem, o tamanho dos elementos na projeção é proporcional à relevância dos mesmos diante do panorama geral. Opcionalmente, o usuário pode interagir destacando os rótulos desejados, alterando, assim, o tamanho dos elementos de acordo com a relevância referente ao rótulo destacado. É possível também manipular âncoras dimensionais na busca exploratória de novos padrões. A codificação de cores pode ser proveniente da combinação linear sobre a contribuição da relevância para cada rótulo ou, de forma mais simples, o elemento pode assumir a cor do rótulo de maior relevância absoluta.

Diversas limitações persistem e devem ser analisadas durante a pesquisa. Algumas herdadas da projeção RadViz. O posicionamento dos atributos no espaço visual nem sempre reflete na informação semântica fornecida, podendo induzir o usuário ao erro. Problemas de *overlapping* também são comuns, em situações mais críticas podem ocorrer com elementos totalmente distintos. A escalabilidade dos atributos projetados é limitada, principalmente diante das estratégias para melhoria de *gestalt* [1]. A quantidade de rótulos também influencia na qualidade da projeção, e o excesso destes empobrece semanticamente a visualização. Conjuntos de dados desbalanceados tendem a gerar *visual clutter* [2]. O modelo de interação com usuário também deve ser trabalhado.

[1] Referência a habilidade humana de perceber o todo a partir de suas partes de maneira intuitiva e mantendo um mapa cognitivo do todo

[2] Situação em que o excesso de elementos causam desordem na representação visual

1.1 Motivação

Esforço significativo vem sendo aplicado em pesquisas com foco na análise de dados multidimensionais juntamente com formas de realizar sua redução de dimensionalidade. A seleção de atributos emerge como ramificação relevante dentre as abordagens exploradas. Vários trabalhos foram propostos, a maioria funcionando como métodos do tipo *black box* [3], onde o processo tem sua metodologia ocultada, isolando por completo o usuário.

Abordagens automáticas para seleção de atributos são relativamente eficientes do ponto de vista computacional, no entanto representam um atalho ao processo, transpondo a etapa que pode revelar informações pertinentes ao usuário. Técnicas de visualização vêm sendo empregadas como forma de inserir o usuário nesta etapa. Fornecer suporte visual para seleção de atributos tornou-se importante diante das várias aplicações de natureza preditiva bastante empregada na análise de dados multidimensionais (KRAUSE; PERER; BERTINI, 2014).

Pode ser tomado como exemplo o conjunto de dados de registros em saúde do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (conjunto de dados referência em diversos momentos deste projeto), onde em uma possível restruturação do protocolo de atendimento realizada pelos programas de qualidade e auditoria médica, é relevante a participação do usuário na etapa de seleção de atributos, tanto por sua contribuição significante diante de sua capacidade analítica, quanto pela oportunidade de ampliar seu conhecimento sob o conjunto de dados. Métodos automáticos podem descartar atributos pouco relevantes sob o ponto de vista estatístico, como por exemplo os atributos *"tem_rx"*, *"rx_coluna_cervical"* e *"rx_face"* (atributos que indicam exames de raios x realizados pelos pacientes) e que na prática dificilmente podem ser filtrados numa tarefa de seleção por se tratarem de atividades essenciais dentro dos protocolos de atendimento aos pacientes traumatizados.

Projetar dados no espaço de atributos e expor ao usuário o processo de seleção pode revelar informações úteis não presentes nas projeções convencionais. Este tipo de projeção evidencia os relacionamentos entre os atributos e demais entidades presentes nos dados, fornecendo nova perspectiva ao usuário para exploração e coleta de padrões. Como no caso dos registros em saúde, onde é possível observar características não triviais do conjunto de dados, em especial informações dos índices de trauma, calculados por equações nem sempre de fácil interpretação. Em sua exploração no espaço de atributos evidencia-se a importância de cada índice e suas peculiaridades. O *New Trauma and Injury Severity* (NTRIS), por exemplo, tem alta correlação com casos de óbito, sendo um bom índice para sua predição, contudo tem baixa correlação com os demais rótulos, onde o índice *New Injury Severity Score* (NISS) consegue maior correlação, sendo mais balanceado com o conjunto inteiro de rótulos. Estas informações são complexas de se extrair apenas observando as equações que envolvem estes índices.

[3] Abordagem em que o processo é visto em termos de sua entrada e saída, sem disponibilidade do detalhamento de seu funcionamento interno

A inserção do usuário na seleção de atributos permite que ele empregue o seu conhecimento especializado nesta etapa. O usuário pode discernir acerca da filtragem de atributos e ainda flexibilizar a seleção de forma alternativa, como priorizar atributos correlacionados a um determinado rótulo. No caso dos registros em saúde, os atributos correlacionados com o rótulo *"obito"*, definitivamente, devem ganhar prioridade.

1.2 Hipótese

Uma boa análise visual de atributos permite ao usuário reconhecer padrões e exceções que levam não somente ao maior conhecimento do conjunto de dados, como também a capacidade de realizar seleção de atributos para melhoria dos processos subsequentes que realizam computação com os dados.

1.3 Objetivos

O objetivo deste trabalho é desenvolver métodos e ferramentas que forneçam apoio visual à análise de atributos de forma a dar condições ao usuário de entender o relacionamento entre atributos e o seu impacto dentro do conjunto de dados, levando assim a uma seleção de atributos de qualidade semelhante, ou superior, aos métodos automáticos. As mesmas ferramentas devem permitir a exploração do espaço de atributos de forma a confirmar hipóteses ou identificar exceções às expectativas do usuário.

1.4 Contribuições

As contribuições iniciais esperadas para este trabalho são:

- abordagem visual interativa para análise de atributos que permita identificar relações não estatisticamente óbvias através da exposição dos relacionamentos entre atributos e classes;
- e uma abordagem de apoio a estimativa e mapeamento de predição de instâncias por atributo para cada classe.

É importante destacar que, até a data da escrita deste trabalho e considerando a busca exploratória na literatura até então, não há conhecimento de abordagens que proporcionem as contribuições acima citadas.

1.5 Metodologia

Esta seção apresenta o delineamento básico da pesquisa. A Figura 1 resume o *design* da proposta dividido em sete fragmentos, descritos a seguir.

1.5.1 Conjunto de dados

Inicialmente serão adotados conjuntos de dados multidimensionais contendo atributos numéricos e/ou categóricos dispostos no formato estruturado atributo-valor. Posteriormente, a tarefa de análise será estendida para os formatos de texto e séries temporais.

O conjunto de dados referência no desenvolvimento deste projeto são os registros em saúde coletadas no Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo (HC-FMRP-USP), realizada em um intervalo de oito anos (2006 a 2014). Possui 21.295 instâncias com registro de 145 atributos tanto numéricos quanto categóricos.

1.5.2 Análise de similaridade

Para análise de similaridade deste projeto é adotada a medida de correlação, método largamente utilizado no contexto de aprendizado de máquina, mineração de dados e estatística, para análise de relevância. O coeficiente de correlação identifica e quantifica o relacionamento linear entre duas variáveis numéricas. Adaptações para o cálculo de correlação envolvendo dados categóricos devem ser investigadas.

1.5.3 Matriz de similaridade

A matriz de correlação armazena os valores de correlação entre pares formados por atributo e rótulo. Assim, cada atributo possui um valor de correlação referente a todos os possíveis rótulos. Somente após a montagem da matriz que os dados ficam aptos a serem projetados para o usuário.

1.5.4 Projeção multidimensional

Para projetar a matriz de correlação, que possui dimensões $k \times m$ (k rótulos por m dimensões), é necessária uma técnica multidimensional que codifique esses dados de forma facilmente compreensível ao usuário, uma vez que ele participará do processo de seleção.

Neste caso particular, a projeção RadViz caracteriza uma boa alternativa por ser considerada de fácil interpretação e por permitir a representação dos dados em um formato flexível. Para sua implementação é empregada a linguagem de programação JavaScript somada à biblioteca de manipulação de documentos baseados em dados D3js.

1.5.5 Interatividade e seleção de atributos

Uma vez projetados, os dados podem ser analisados pelo usuário que, eventualmente, deve realizar a seleção de atributos. A interatividade deve então abordar duas perspectivas distintas: a exploração dos dados e a seleção de atributos.

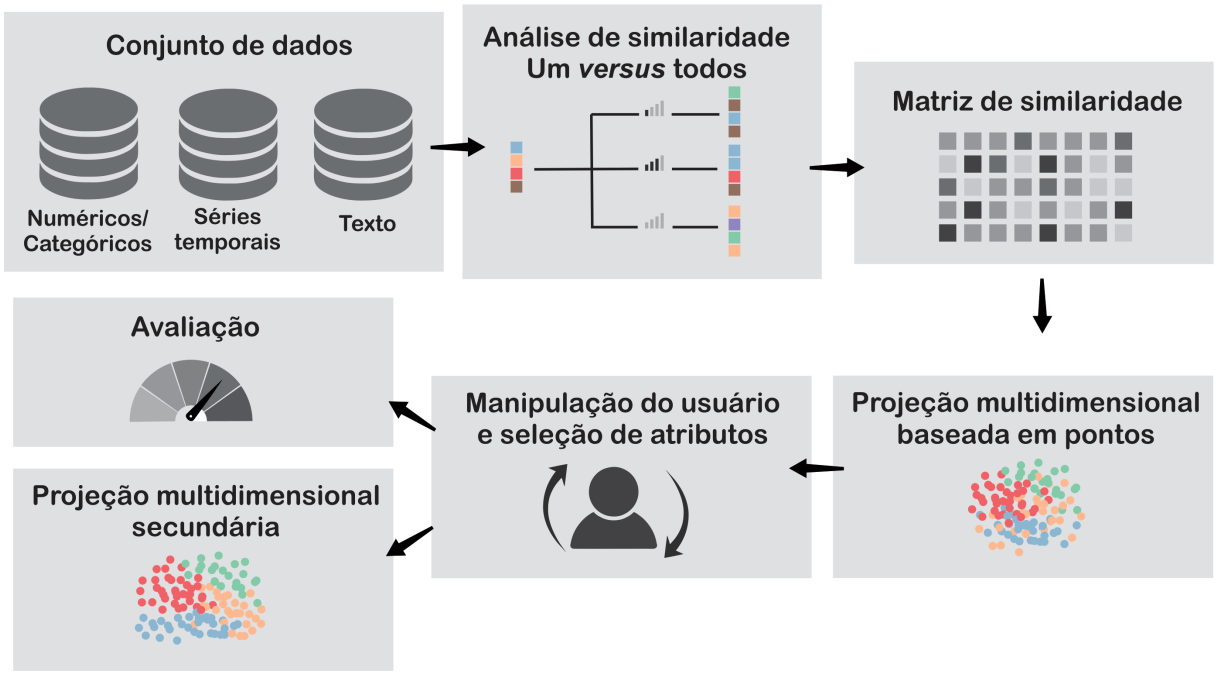

Figura 1 – Delineamento básico desta proposta dividido em suas etapas fundamentais.

Fonte: Elaborada pelo autor.

Na perspectiva de análise, é desejável que o usuário tenha liberdade de manipulação dos elementos visuais de forma a facilitar o reconhecimento de padrões na visualização. Para a seleção de atributos, o usuário poderá, a qualquer momento, definir os atributos selecionados através das operações de inclusão, exclusão ou alteração da seleção.

1.5.6 Avaliação

Inicialmente deve ser realizada uma avaliação qualitativa acerca da densidade semântica a partir de observações coletadas do modelo de projeção dos dados de correlação. Esta avaliação visa examinar a qualidade da análise que este projeto propõe.

Dentro do contexto de seleção de atributos, uma das formas mais comuns de avaliação é o emprego dos próprios classificadores para posterior cálculo de acurácia. É um método preciso, porém custoso. Nesta proposta, é empregada uma segunda técnica de visualização, para realizar a avaliação precoce da seleção de atributos. Projeções multidimensionais, tal como demais processos que lidam como este tipo de dados, são sensíveis à qualidade dos atributos, e uma vez que uma nova seleção é realizada, é possível fazer uma avaliação prévia a partir de alguma projeção multidimensional rápida baseada em pontos.

Para testar o desempenho da seleção de atributos, pretende-se confrontar a proposta com os principais métodos de seleção automáticos. A avaliação dos subconjuntos selecionados deve ser realizada através do cálculo de acurácia após submissão a algum modelo classificador, usualmente o *Support Vector Machine* (SVM).

1.6 Organização da monografia

Esta monografia está organizada da seguinte maneira: o capítulo 2 apresenta uma visão geral acerca da seleção de atributos e descreve os principais trabalhos relacionados presentes na literatura; no capitulo 3 a abordagem proposta é apresentada e os resultados preliminares que buscam validar a hipótese deste projeto de pesquisa. Por fim, o Capítulo 4 apresenta a descrição das atividades e o cronograma.

CAPÍTULO

2

CONCEITOS FUNDAMENTAIS E TRABALHOS RELACIONADOS

Este capítulo apresenta conceitos fundamentais acerca dos métodos de seleção de atributos na Seção 2.1, e os trabalhos relacionados à análise e seleção de atributos com suporte de técnicas de visualização em 2.2, que para melhor organização é dividido em duas subseções: análise visual dos dados multidimensionais no espaço de atributos, em 2.2.1, e apoio visual na seleção de atributos em 2.2.2.

2.1 Seleção de atributos

Dados multidimensionais são geralmente representados a partir de coleções de itens, que por sua vez possuem conjuntos de variáveis. Cada variável simboliza um atributo (ou uma dimensão) na perspectiva do conjunto de dados. No cenário atual, diante da capacidade de coleta e armazenamento dos dispositivos modernos, os dados multidimensionais com frequência dispõem de excesso de atributos. Muitos deles não apresentam significância diante dos critérios desejados para extração de conhecimento. Outros se apresentam de maneira redundante. Em ambas as situações, os atributos se manifestam em forma de ruído para os métodos que processam esses dados.

Para contornar o problema do excesso de dimensões representando ruído ao conjunto de dados, estratégias de redução de dimensionalidade foram propostas. Dentro delas está a seleção de atributos. O objetivo dos métodos de seleção de atributos é selecionar o subconjunto de variáveis que descreva efetivamente o conjunto de dados, através da redução de variáveis redundantes ou irrelevantes (GUYON; ELISSEEFF, 2003). A exclusão desses atributos implica em redução da capacidade de armazenamento e custo computacional enquanto evita-se perdas significantes de informação ou gere degradação na qualidade dos processos que lidam com tais dados (LI *et al.*, 2016).

Nesta seção são descritos os conceitos básicos de seleção de atributos para o correto entendimento do conteúdo exposto nos capítulos subsequentes. Também apresenta o formato genérico dos algoritmos e suas abordagens mais comuns. Por fim, descreve os principais métodos automáticos de seleção de atributos.

2.1.1 Atributos

Atributos representam as variáveis de entrada do conjunto de dados (GUYON; ELISSEEFF, 2003). No modelo vetorial, cada atributo caracteriza uma dimensão, desta forma vários atributos constituem dados multidimensionais. Os atributos podem ser: discretos – podendo assumir uma quantidade finita de valores; numéricos – podem assumir valores no domínio dos números reais; complexos – assumem dados complexos não tão comuns, como sons e imagens (podem ser convertidos em atributos discretos e contínuos a partir de técnicas de extração de atributos); e compostos – podendo representar combinações dos outros tipos de atributos.

2.1.2 Algoritmo genérico para seleção de atributos

Algumas similaridades são encontradas nos diversos métodos de algoritmos de seleção de atributos, com isso é possível descrever um modelo genérico para melhor ilustrar estas abordagens (Figura 2). O objetivo do algoritmo é encontrar o subconjunto de atributos que melhor representa o domínio de dados em conformidade com algum critério. De acordo com (DASH; LIU, 2003)(LEE, 2005), existem alguns passos básicos para um método de seleção de atributos típico:

- um procedimento de geração para gerar os próximos subgrupos candidatos;
- uma função de avaliação para avaliar o subconjunto corrente;
- um critério de parada para estabelecer um ponto de parada para o método;
- um procedimento de validação para validar a solução encontrada.

A etapa de geração usualmente segue uma de três possíveis estratégias, duas sequenciais, conhecidas como *sequential forward selection* (SFS) e *sequential backward selection* (SBS) e outra aleatória, conhecida como *randomized* (ANG *et al.*, 2016). Cada uma delas relacionada a uma direção de busca. Na etapa de avaliação, alguns tipos de medidas podem ser adotados, como consistência e acurácia. Uma descrição mais detalhada é realizada a seguir.

2.1.3 Direção da busca

Relacionada ao ponto de partida do espaço de busca (completo, vazio ou aleatória), a direção de busca remete ao sentido em que o algoritmo investiga as soluções, podendo estrategi-

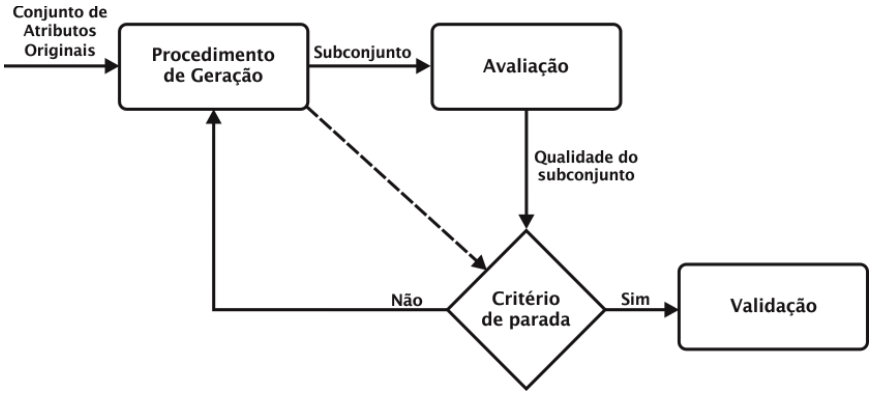

Figura 2 – Processo genérico de seleção de atributos com validação. Quatro tarefas principais compõem o algoritmo. Ao final, a solução passa por uma validação.

Fonte: Dash e Liu (2003), Lee (2005).

camente prosseguir adicionando atributos ao subconjunto candidato, removendo atributos, ou até mesmo ambos.

No modo *sequential forward selection* (SFS), o algoritmo inicia sua execução com o espaço de busca vazio, e a cada passo iterativo do processo, adiciona novo atributo ao subconjunto. A Figura 3 ilustra esse tipo de exploração. Sua vantagem é a simplicidade de implementação, porém sofre com a pouca flexibilidade do modelo. Uma vez que um atributo é selecionado, ou descartado, não há meios de reverter a ação. Para contornar este problema, uma variante chamada *sequential forward floating selection* (SFFS) é proposta, permitindo mecanismos de selecionar atributos já descartados ou remover atributos já selecionados.

No modo *sequential backward selection* (SBS), o algoritmo parte com espaço de busca completo, e iterativamente remove elementos. Análogo ao *SFS*, este modo tem simples implementação, mas não é flexível. A variante *sequential backward floating selection* (SBFS) foi proposta para contornar o problema de forma semelhante ao *SFFS*.

Por fim, no modo randômico, a estratégia de busca parte de um subconjunto gerado aleatoriamente, e iterativamente prossegue a partir da adoção de umas das estratégias de busca (SFS, SFFS, SBS ou SBFS), ou alternativamente pode adotar estratégia que não emprega movimentos regulares (LI *et al.*, 2016).

2.1.4 Estratégia da busca

Uma vez definida a direção de busca, é necessário determinar a estratégia de exploração do espaço de busca. Considerando os m atributos do conjunto de dados, a quantidade de possíveis subconjuntos é 2^m. O que torna proibitiva a realização de uma busca exaustiva para uma razoável quantidade de atributos. Algumas estratégias podem ser adotadas, sendo as principais: completa, heurística e aleatória.

A completa (também conhecida como busca exponencial) investiga, a princípio, todo o

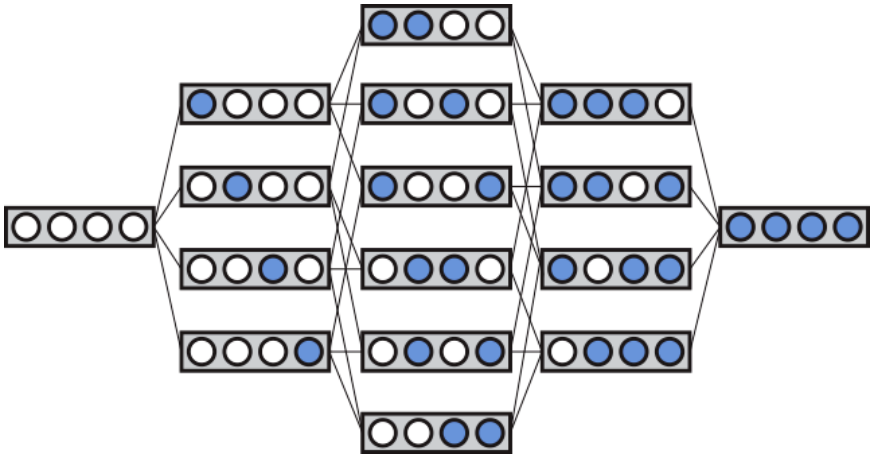

Figura 3 – Exploração do espaço de busca incremental tipo SFS. A cada passo da busca, um elemento é inserido, até algum critério de parada ser satisfeito, ou a exploração concluir.

Fonte: Langley (1994), Lee (2005).

espaço de busca com a finalidade de extrair a solução ótima, sendo assim, uma busca exaustiva. Considerando o espaço de busca como 2^m, a pesquisa se torna impraticável para um d grande. Portanto, diversos métodos são propostos para alcançarem uma solução ótima local (LI *et al.*, 2016), não sacrificando o desempenho desta forma.

Outro modelo de estratégia de busca emprega o conceito de heurística, onde em vez de realizar busca exaustiva e cega, o algoritmo segue, a cada passo iterativo, um modelo que representa boa probabilidade para encontrar a solução ótima (local ou, eventualmente, global). Diferentes funções heurísticas podem ser usadas para reduzir o espaço de busca sem comprometer as chances de encontrar o resultado ideal. Assim, embora a ordem do espaço de pesquisa seja o mesmo da busca completa, um número menor de subconjuntos é avaliado (LIU; YU, 2005).

Na estratégia de busca aleatória não há um critério na escolha do próximo subconjunto candidato. Apesar do espaço de busca ser $O(2^m)$, esta abordagem não chega a investigar todos os 2^m subconjuntos por, geralmente, empregar um valor máximo de iterações. Pode, ou não, empregar o subconjunto corrente como referência para o próximo.

2.1.5 Método de avaliação

A maneira que o algoritmo realiza a avaliação dos atributos (ou subconjuntos de atributos) define o método de avaliação. Estes podem ocorrer de forma incorporada ao algoritmo (*embedded*), anteriormente ao algoritmo (filtro), ou até mesmo empregando o algoritmo como avaliador dos subconjuntos gerados (*wrapper*).

O método de avaliação *embedded* (embutido) acontece durante o processamento do algoritmo de classificação, representando soluções específicas para um conjunto de algoritmos de aprendizado de máquina (GUYON; ELISSEEFF, 2003). Como a seleção de atributos acontece dentro da classificação, geralmente esta abordagem apresenta boa eficiência. Contudo, possui a

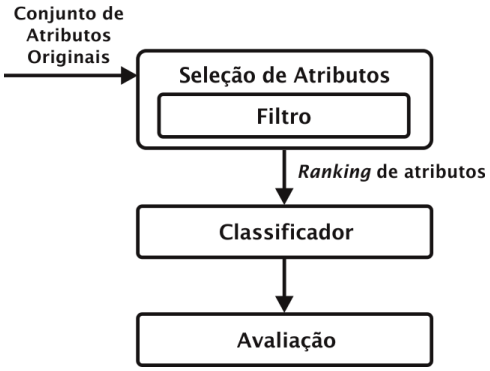

Figura 4 – Representação da abordagem Filtro, nela a etapa de seleção de atributos se comporta de forma independente do algoritmo de classificação.

Fonte: Elaborada pelo autor.

desvantagem de se caracterizar como solução restrita.

A avaliação tipo filtro atua de forma independente do algoritmo de classificação (Figura 4), empregando métricas geralmente estatísticas para definir a importância dos atributos. Apenas depois de seu processamento completo é que os algoritmos subsequentes obtém acesso ao subconjunto selecionado (JOVIĆ; BRKIĆ; BOGUNOVIĆ, 2015). Seu resultado é frequentemente um *ranking* dos atributos mais importantes, cabendo às etapas seguintes decidirem quais utilizar. A vantagem desta abordagem é não depender da complexidade do algoritmo de indução. Como desvantagem, o fato de não considerar o critério de avaliação real do classificador pode trazer inconsistências aos resultados, dado que nem sempre os atributos mais relevantes no ponto de vista do algoritmo de seleção de atributos representam a solução ótima para o classificador.

A abordagem *wrapper* (Figura 5) emprega o próprio classificador como caixa preta para avaliação dos subconjuntos gerados. Tem a vantagem de trabalhar com a acurácia real do classificador, no entanto, é necessário considerar os custos da avaliação. O método funciona com base em dois passos principais: busca por subconjuntos de atributos e avaliação do subconjunto selecionado (LI *et al.*, 2016).

2.1.6 Medida de avaliação

A função de avaliação mensura a qualidade de um subconjunto de atributos gerado e o compara com a melhor solução encontrada até então, que em caso afirmativo, passa a ser a melhor solução atual. Em (DASH; LIU, 2003) são descritos 5 tipos de parâmetros para avaliação, definidas resumidamente a seguir.

- Medida de distância: conhecida também como medida de separabilidade, divergência ou discriminação. Supondo um problema de duas classes, a importância de um atributo é maior que outro quando este induz uma maior diferença entre as probabilidades condicionais das duas classes. Caso não haja diferença, eles são indistinguíveis. Um exemplo de medida de

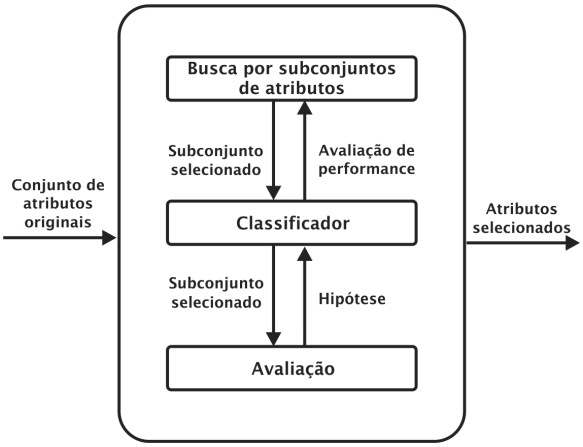

Figura 5 – Representação da abordagem *Wrapper*. O próprio classificador é utilizado como ferramenta de avaliação do subconjunto de atributos gerados. A execução finaliza quando o avaliação satisfaz algum critério.

Fonte: Adaptada de Li *et al.* (2016).

distância é a distância Euclidiana.

- Medida de informação: é dada pela diferença da incerteza antes e depois da inserção do atributo ao subconjunto selecionado. Um atributo é considerado mais relevante quando seu ganho de informação é superior.
- Medida de dependência: conhecida também como medida de correlação, qualifica a habilidade de um atributo predizer o valor de outro. O coeficiente da medida de dependência pode ser empregado para calcular a correlação de um atributo com uma classe. Se a correlação de um atributo com a classe for maior que outro, então ele será considerado mais importante.
- Medida de consistência: o conceito de consistência remete ao fato de que, se dois exemplos do conjunto de dados forem semelhantes (todos os atributos semelhantes), implica que esses exemplos devem pertencer à mesma classe. De forma similar, se os exemplos foram diferentes, espera-se que pertençam a classes diferentes. A medida de consistência não tenta maximizar a separabilidade de classes, a ideia é encontrar o menor conjunto de atributos que possa distinguir classes de forma semelhante ao conjunto original (DASH; LIU, 2003).
- Medida de acurácia: emprega a própria função de avaliação do algoritmo para calcular precisão do subconjunto selecionado. A abordagem que utiliza essa medida é conhecida como *Wrapper*.

2.1.7 Principais métodos de seleção de atributos

Esta subseção apresenta um resumo dos principais métodos de seleção de atributos. É importante destacar que, diante do interesse desta proposta, são considerados apenas abordagem do tipo filtro que objetivam a seleção para subsequente classificação.

- *ReliefF*: (ROBNIK-ŠIKONJA; KONONENKO, 2003): pontua atributos de acordo com a similaridade de instâncias sorteadas com seus k vizinhos. Destaca atributos com valores distantes entre vizinhos mais próximos de classes diferentes e penaliza atributos que possuem valores ditantes para vizinhos próximos de mesma classe.

- *Fisher Score* (DUDA; HART; STORK, 2012): baseada em similaridade, emprega o critério *fisher* para ranquear o conjunto de atributos de acordo com sua pontuação. A ideia central é selecionar os atributos que mantenham as instâncias de mesma classe o mais próximo possível enquanto que as instâncias de classes diferentes estejam o mais afastado possível (análogo à medida de silhueta).

- *Trace Ratio* (NIE *et al.*, 2008): baseada em similaridade, calcula a relevância do atributo a partir do cálculo do critério *trace ratio*. De forma semelhante ao *Fisher Score*, o critério busca maximizar a similaridade das instancias de mesma classe e minimizar em relação às instâncias de classes diferentes. Quanto maior o *score*, maior a relevância do atributo.

- *Information Gain* (LEWIS, 1992): baseado em medida de informação, o método mensura a relevância dos atributos de acordo com sua correlação com os rótulos dos dados. Atributos com alto valor de correlação com as classes são considerados importantes e selecionados.

- *Mutual Information Feature Selection* (MIFS) (BATTITI, 1994): o método considera que, além de alta correlação com os rótulos, é necessário que os atributos possuam baixa correlação entre si. Minimiza o problema de redundância presente no método *Information Gain*.

- *Minimum Redundancy Maximum Relevance* (mRMR) (PENG; LONG; DING, 2005): baseado em medida de informação, o método busca selecionar atributos relevantes evitando redundâncias. A máxima relevância busca pelos atributos de maior correlação com os rótulos, contudo estes possuem a tendência de se apresentar de forma redundante. Então a mínima redundância busca atributos distantes. O algoritmo busca, então, os atributos mais relevantes que sejam distantes uns dos outros.

- *Fast Correlation Based Filter* (FCBF) (YU; LIU, 2003): baseado em medida de informação, o método explora a correlação entre atributos e classes como também entre atributos de forma simultânea. Neste método não são calculados *scores* para cada atributo, sendo sua saída um subconjunto de atributos e não um *ranking*. O método aplica uma medida de incerteza simétrica para fazer a filtragem dos atributos redundantes mantendo a representatividade do subconjunto selecionado.

- *Joint Mutual Information* (JMI) (YANG; MOODY, 1999)(MEYER; BONTEMPI, 2006): enquanto MIFS e mRMR preocupam-se em reduzir a redundância de atributos selecionados, JMI propões incrementar a complementaridade de informações compartilhadas

pelos atributos. A ideia é escolher atributos candidatos (aleatoriamente) e testar sua complementaridade com o conjunto atual, em caso positivo esta deve ser incluído ao novo subconjunto.

- *Conditional Infomax Feature Extraction* (CIFE) (LIN; TANG, 2006): maximiza a informação conjunta relevante em relação às classes ao reduzir explicitamente as redundâncias relevantes entre elas. Funciona diante da hipótese em que é mais vantajoso eliminar redundâncias entre classes que entre atributos.

- *Conditional Mutual Information Maximization* (CMIN) (FLEURET, 2004): iterativamente seleciona atributos que maximize sua informação mútua com os rótulos considerando o subconjunto de atributos atual. O critério (CMIN) não permite que um atributo similar a outro já selecionado seja escolhido, dado que este não carrega nenhuma informação extra que possa ser empregada na predição das classes.

- *Low Variance* (PEDREGOSA *et al.*, 2011): método baseado em medida de estatística, elimina atributos com valores de variância abaixo de determinado limiar. Baseia-se na ideia que atributos com baixa variância tem pouca capacidade de discriminar classes.

- *F-Score* (WRIGHT, 1965): método baseado em medida de estatística, busca selecionar atributos aptos a separar corretamente as instâncias das diferentes classes. Para tanto, emprega cálculo de variância das classes e variâncias entre classes. O critério testa se, ao longo do atributo, há variância que faça distinção das classes.

- *Chi-Square Score* (LIU; SETIONO, 1995) : emprega o teste de independência para avaliar o quanto um atributo é independente das classes. Atributos com o *score* calculado alto são considerados mais importantes.

2.2 Visualização no espaço de atributos

Técnicas de visualização são convencionalmente empregadas em sistemas que focam na análise de conjunto de dados a partir da projeção das instâncias e subsequente coleta de observações. Contudo, há um crescimento na quantidade de pesquisas interessadas em mais que apenas interpretar os dados, existe o interesse de entender os dados e suas capacidades preditivas (KRAUSE; PERER; BERTINI, 2014). A análise dos dados no espaço de atributos pode revelar informações importantes, portanto vem ganhando ferramentas para tal. Esta seção apresenta de forma sucinta os principais trabalhos propostos para apoio visual à análise e/ou seleção de atributos.

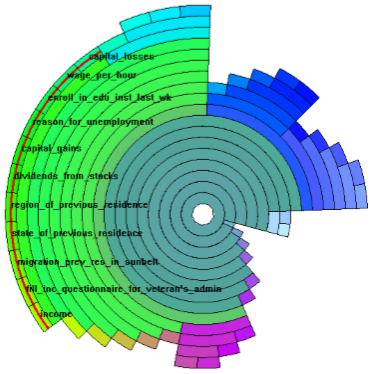

Figura 6 – Visualização dos dados utilizando o método *InterRing* (YANG *et al.*, 2003), nele o usuário pode realizar interações para mudar a estrutura de hierarquia das dimensões.

2.2.1 Análise visual de dados multidimensionais no espaço de atributos

Alguns métodos visuais para exploração de dados no espaço de atributos foram propostos. Estas abordagens se diferenciam das projeções tradicionais por exporem os dados sob a perspectiva dos atributos (em vez de instâncias). Nesse tipo de projeção geralmente é evidenciado o relacionamento entre os atributos, adicionalmente pode-se exibir suas relações também com os rótulos dos dados.

Em (YANG *et al.*, 2003), os autores propõem uma abordagem para manuseio de dados de alta dimensionalidade denominado *Visual Hierarchical Dimension Reduction* (VHDR). A ideia central é visualizar dados multidimensionais sem perder significado intuitivo para o usuário. Na abordagem VHDR, as dimensões são organizadas em uma estrutura hierárquica. A projeção dos dados é realizada a partir do método de visualização radial hierárquica chamado *InterRing* (Figura 6). De forma geral, o método segue cinco passos: geração da hierarquia de dimensão, navegação e modificação da hierarquia de dimensão, seleção de agrupamento de dimensão, geração de representação dimensional e visualização e projeção de dados. Em complemento, usuários podem manusear as representações de dimensões, dando nomes significantes, alterando seus relacionamentos, ou até mesmo unificando agrupamentos semanticamente próximos. Apesar de oferecer boas representações do espaço original reduzido a partir do agrupamento de atributos, o método não busca objetivamente a seleção de atributos para redução de dimensionalidade. O objetivo principal é tornar a visualização mais intuitiva sumarizando os atributos na representação dimensional corrente.

Em (SEO; SHNEIDERMAN, 2005), os autores descrevem um *framework* conceitual (Figura 7) onde, de acordo com um *ranking* gerado a partir de critério escolhido pelo usuário, os atributos são exibidos graficamente por diferentes visões. O critério de ranqueamento é escolhido pelo usuário para então os atributos serem impressos em gráficos, expondo o relacionamento par a par, intensidade do valor do critério (cores), resumo da distribuição da dimensão, entre outras informações. A abordagem revela dados pertinentes com ponto de vista nos atributos, porém

Figura 7 – Tela do *Hierarchical Clustering Explorer*, ferramenta proposta em (SEO; SHNEIDERMAN, 2005). O usuário escolhe um critério de ranqueamento e todas as projeções da ferramenta passam a ser ordenadas pelo critério

não visa à seleção, tornando desconhecida sua eficiência para tal. Informações de relevância de atributos também poderiam ser exploradas, tendo em vista que a análise realizada no método aborda a relação no espaço de características.

Turkay et al. (TURKAY; FILZMOSER; HAUSER, 2011) descrevem um modelo que permite a análise visual interativa a partir da combinação da visão do conjunto de dados associada à visão de propriedades estatísticas dos atributos. A interatividade acontece apoiada na manipulação estilo *linking and brushing*[1] de uma das visualizações, que subsequentemente atualiza a outra com abordagem *focus+context*[2]. A visualização dos itens está vinculada a visualização de dimensões e vice versa. Desta forma, o usuário é capaz de perceber, em conjunto, a estrutura do espaço de dimensões, bem como a distribuição de itens de dados com relação às dimensões. Apesar de indiretamente realizar seleção de atributos com apoio na visualização dupla (de itens e atributos), o método não visa a eliminação de atributos redundantes e/ou irrelevantes.

No trabalho proposto em (MAMANI *et al.*, 2013), é descrito um modelo que molda o espaço de características a partir da percepção do usuário realizada com base na transformação interativa de projeções de amostras e mapeamentos locais. A Figura 8 exibe as etapas básicas da abordagem. Inicialmente uma amostra do conjunto de dados é selecionada e projetada (por algum método de projeção multidimensional que respeite, na medida do possível, as distâncias originais). Em seguida, o usuário pode manipular a projeção, ajustando-a de acordo com seu

[1] Combinação de diferentes projeções onde a manipulação interativa de uma reflete em mudanças na outra.
[2] Destaca visualmente áreas de interesse ao tempo em que preserva a visualização global com menos detalhes

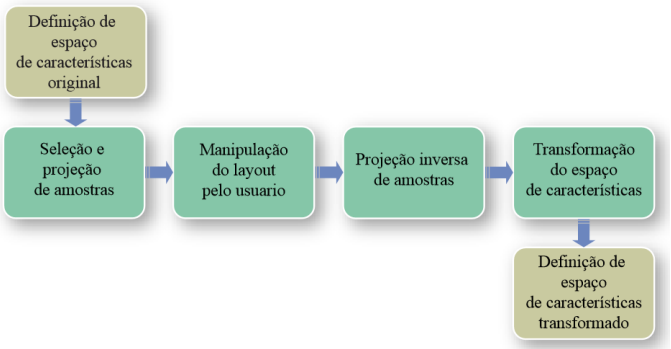

Figura 8 – Sequência de passos realizada pela abordagem apresentada em (MAMANI *et al.*, 2013), onde o espaço de atributos é modificado a partir de informações coletadas pela interação do usuário.

ponto de vista. A informação inserida pelo usuário é empregada na aproximação denominada pelos autores como *Neighbor-Sample Inverse Force Scheme* que atualiza a matriz de distâncias para subsequente transformação dos dados. Para avaliar os resultados, as medidas de silhueta e acurácia de agrupamento são adotadas. A interferência do usuário nas relações de similaridade através das projeções melhoram a coesão e a separação entre grupos dos conjuntos de dados multidimensionais. A limitação principal do método está relacionada aos conjuntos de dados que possuem grande quantidade de grupos com poucas amostras representativas. A partir do momento em que amostras pouco representativas são manipuladas, a abordagem gera distorções que não refletem no ponto de vista real do usuário.

Um modelo de visualização baseado na construção de um mapa de correlações dos atributos é proposto em (ZHANG *et al.*, 2015). O mapa é construído em duas dimensões, a partir da matriz de correlações, expondo os relacionamentos e intensidade dos atributos. Usuários podem manipular parâmetros para melhorar a interpretação do mapa de correlação. Os autores também ressaltam a possibilidade de tratar dados numéricos e categóricos de forma unificada, aplicar *zoom* multi-escala, possibilidade de alterar o limiar de correlação e a capacidade de visualização de subespaços de variáveis correlacionadas. O trabalho não visa a seleção de atributos, mesmo podendo servir de referência. O objetivo é proporcionar uma análise a partir do ponto de vista de correlações dos atributos.

2.2.2 Apoio visual na seleção de atributos

Métodos automáticos para seleção de atributos vêm atingindo seus objetivos com relativo sucesso, porém isola o analista desta etapa que geralmente revela informações significativas. Com base nisto, ferramentas de apoio visual vêm ganhando espaço no contexto de seleção de atributos. Nas abordagens descritas nesta seção, além da alternativa de exploração do espaço de atributos, é possível a realização de seleção de atributos.

Em (DY; BRODLEY, 2000b) é apresentado o *Visual Feature Subset Selection using Expectation-Maximization Clustering* (Visual-FSSEM) que emprega técnicas de visualização e

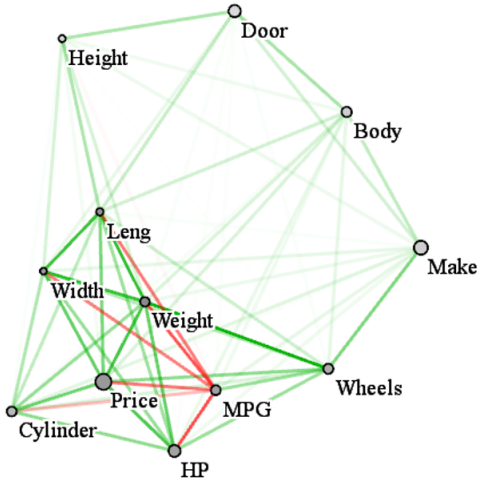

Figura 9 – Resultado a visualização do mapa de correlação apresentado em (ZHANG *et al.*, 2015). A abordagem expõe o relacionamento entre os atributos, evidenciando a correlação positiva ou negativa (verde ou vermelho).

agrupamento juntamente com interações com o usuário para guiar a escolha do subconjunto de atributos e maximizar o conhecimento sobre o conjunto de dados. Os autores adotam o método *linear discriminant analysis* (LDA) como modelo de visualização. A base do método é o FSSEM original (DY; BRODLEY, 2000a), que faz uso de algoritmos de agrupamento para proceder com a seleção de atributos. Interativamente, o usuário pode escolher o melhor passo (SFS ou SBS) da iteração exposto em forma de gráficos de dispersão pelo método. O método não é eficiente para grandes quantidades de dimensões por ter que exibir as possibilidades para o consecutivo discernimento do usuário. Em situações de alta dimensionalidade, as possibilidades para algoritmos sequenciais (seja SFS ou SBS) são consideravelmente variadas, e inviáveis de ilustrar individualmente.

É proposto, em (GUO *et al.*, 2003), um método de seleção de atributos para agrupamentos baseada nas relações de entropia condicional para cada par de dimensões, montando, assim, uma matriz de valores de entropia e correlação, que revelam o potencial destes em conter níveis significantes de agrupamentos. Uma visualização da matriz obtida é realizada para que o usuário possa proceder com a seleção das dimensões desejadas (Figura 10). Alternativamente, um método automático é também proposto. Para melhorar a visualização da matriz, os autores ordenam as dimensões de acordo com suas correlações, posicionando dimensões correlacionadas o mais próximo possível, para tanto, empregam árvore geradora mínima. O método possui boa escalabilidade, podendo descrever um valor significante de dimensões. Contudo, representa pouca melhoria em termos de interpretabilidade do conjunto de dados, também aproveita pouco a expertise do usuário, uma vez que a simples definição de um limiar para os valores de entropia selecionados pode automatizar o método.

Uma técnica de visualização, denominada *SmartStripes*, que permite aos usuários participarem do processo de seleção de atributos é proposta em (MAY *et al.*, 2011). Ela possibilita a investigação de dependências (e interdependências) entre subconjuntos diversos de atributos e

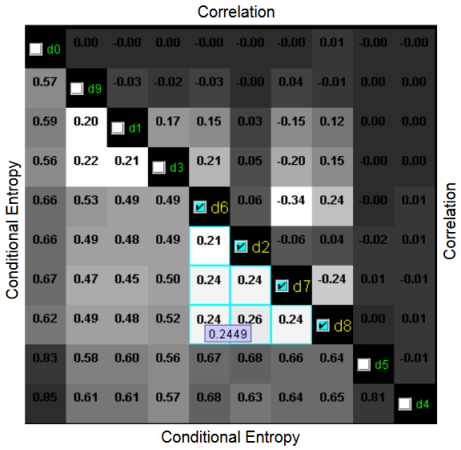

Figura 10 – Visualização da matriz de entropia condicional (GUO *et al.*, 2003). A coloração de cada célula faz referência ao calor de entropia (esquerda/inferior) ou correlação (direita/superior).

de itens do conjunto de dados. A ideia central é encontrar atributos com máxima correlação em partições de instâncias do conjunto de dados. Para cada partição poderá existir um atributo que forneça separabilidade às classes no âmbito local. Para evitar que o usuário faça a busca em todo espaço de atributos, o método prevê ranqueamento de atributos com métodos automáticos para reduzir escopo com bons candidatos. Em alguns contextos a visualização da relação atributo *versus* classe se apresenta de forma severamente dispersa, dificultando a tarefa do usuário de encontrar correlações. Por fim, o método não é devidamente avaliado, uma vez que os autores fazem comentários acerca dos resultados sem a exposição clara destes.

Em (BOTELHO, 2011) é apresentada uma abordagem que emprega projeção multidimensional para auxílio na seleção de atributos. A matriz original contendo o conjunto de dados é transposta com finalidade de projetarem-se atributos em vez dos elementos. Em seguida três processos de seleção são efetuados (Figura 11). O primeiro projeta os dados transpostos, e com auxílio do usuário, agrupamentos são identificados e são escolhidos atributos representantes destes. No segundo processo, em vez de selecionar os atributos manualmente, é aplicado o algoritmo *k-means*. No terceiro processo, o algoritmo *k-means* é aplicado diretamente ao conjunto transposto, gerando *clusters* de características. Para avaliação dos resultados a autora emprega a medida de silhueta no espaço original e em seguida no espaço projetado. Dois estudos de casos são apresentados para análise. Com a aplicação da abordagem, a autora obteve resultados interessantes, visto que a projeção manteve qualidade diante da redução significativa da quantidade de atributos. A abordagem trata do apoio visual à seleção de atributos para subsequente clusterização, o que torna sua aplicação conveniente para dados não supervisionados. Com isso, o trabalho tem boa eficiência na eliminação de atributos redundantes, porém não é eficiente na remoção de atributos não relevantes.

Na abordagem descrita em (CRUZ, 2012), uma árvore de similaridade *Neighbor-Joining* é empregada apoiando o usuário na busca pelo subconjunto ótimo de atributos. A abordagem é descrita em 8 passos: extração de características de um conjunto de imagens, transposição da

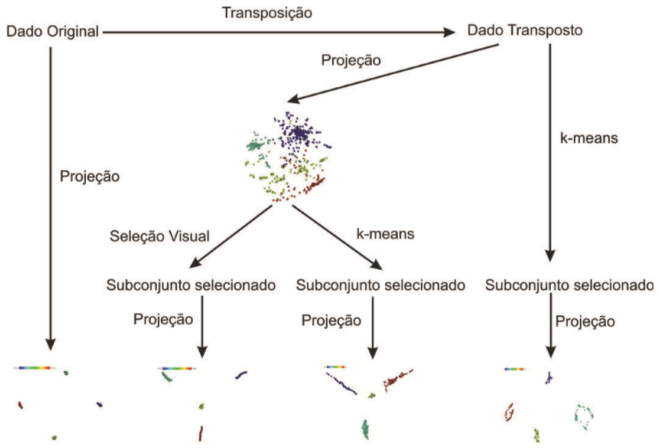

Figura 11 – *Pipeline* da abordagem apresentada em (BOTELHO, 2011), onde um método de projeção multidimensional é empregado para apoiar a tarefa de seleção de atributos.

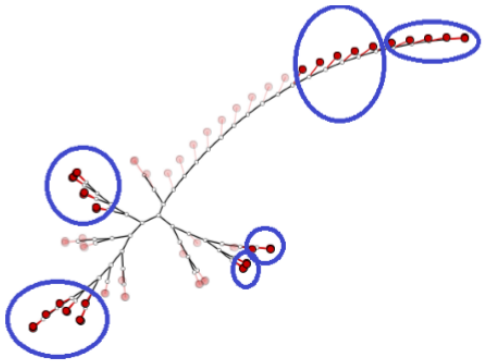

Figura 12 – Os atributos são representados pela técnica *Neighbor Joining* permitindo que o usuário realize a seleção de um subconjunto de atributos (CRUZ, 2012).

matriz de instâncias,pré-processamento, visualização das características, análise e seleção das características, obtenção das características selecionadas (Figura 12), avaliação, retroalimentação. Assim como o trabalho descrito em (BOTELHO, 2011), o principal critério considerado para realização da avaliação é a medida de silhueta. Também emprega medidas baseadas na vizinhança entre as projeções antes e depois da seleção. A abordagem revela bons resultados, obtidos principalmente pela eficácia da árvore *Neighbor Joining* no agrupamento de atributos similares para posterior seleção do usuário. Suas limitações também se assemelham ao trabalho (BOTELHO, 2011), onde há pouca eficiência na redução de atributos irrelevantes ao conjunto de dados.

Buscando auxiliar o usuário a entender como os atributos estão sendo ranqueados (ou selecionados) diante dos algoritmos de seleção de atributos convencionais, (KRAUSE; PERER; BERTINI, 2014) propõem uma ferramenta de análise visual, chamada INFUSE (*INteractive FeatUre SElection*). A tela principal da ferramenta apresenta quatro componentes: visualização de atributos, visualização de lista, visualização do classificador e construtor de modelo interativo. O primeiro exibe os atributos a partir de glifos divididos em fatias para representação dos valores

obtidos pelos algoritmos de seleção de atributos. O segundo exibe a lista completa e ordenada de todos os atributos. O terceiro componente mostra a análise dos atributos a partir dos cinco classificadores. O ultimo componente é voltada a customização de modelos para subsequente avaliação. A ferramenta apresenta uma visualização bastante completa, no entanto o excesso de informação pode sobrecarrega-la, por vezes, tornando a seleção de atributos customizada algo não trivial.

Um método para seleção de atributos de imagens interativa e iterativa auxiliada por técnicas de redução de dimensionalidade (entre outras ferramentas complementares) é descrito em (RAUBER *et al.*, 2015). A ferramenta conta com cinco visualizações que expõem o panorama da seleção e de interação, sendo elas: *observation view*, *feature view*, *group view*, *projection view* e *feature scoring view*. A sequência de passos para a seleção é iniciada pela filtragem inicial de atributos, realizada a partir da abordagem *Recursive Feature Elimination* (RFE) (GUYON *et al.*, 2002). A etapa é exposta ao usuário através da tela de projeção. Em seguida é aplicada outra técnica para filtragem de atributos conhecida como *Randomized Decision Trees*. O resultado é novamente exposto ao usuário, que por sua vez procede com um refinamento manual para finalizar a seleção.

2.3 Considerações finais

As abordagens apresentadas neste capítulo visam fornecer análise no espaço de características, e em alguns casos, acompanhados com métodos de redução de dimensionalidade por seleção de atributos. Todos são interativos e contam com apoio de técnicas de visualização, o que os diferenciam das abordagens tradicionais de seleção de atributos.

Um resumo das características de cada método exposto neste capítulo é apresentado na Tabela 1. É possível notar alguma deficiência dos métodos em usar a expertise do usuário. Também há carência na apresentação das relações entre classes tal como relações entre classes e atributos. A proposta deste projeto pretende agregar todas as características elencadas na tabela.

Para realizar a avaliação das abordagens, os trabalhos (YU; LIU, 2003),(SEO; SHNEIDERMAN, 2005),(BOTELHO, 2011),(TURKAY; FILZMOSER; HAUSER, 2011),(KRAUSE; PERER; BERTINI, 2014) e (ZHANG *et al.*, 2015) empregam estudos de caso ou demonstrações, enquanto em (MAMANI *et al.*, 2013) é apresentado evolução da eficiência a cada passo da técnica, em (RAUBER *et al.*, 2015) é feita uma pequena análise de todas os atributos *versus* os selecionados pela técnica. Por fim, o trabalho apresentado em (GUO *et al.*, 2003) é o único que faz confrontamento com outras técnicas existentes.

Ferramentas visuais para análise e seleção de atributos são excelentes alternativas quando se deseja, além da realização da seleção, transmitir conhecimento contido no espaço de características. Simultaneamente em que o usuário aumenta sua familiaridade com os dados, sua contribuição na seleção se torna mais precisa e significativa.

Tabela 1 – Comparativo das características encontradas nas abordagens apresentadas neste trabalho. As propriedades analisadas foram: interatividade com usuário, capacidade de escalabilidade, foco na melhoria em técnicas de agrupamento, foco na melhoria em técnicas de classificação, emprega a expertise do usuário, realiza seleção de atributos, promove descoberta do conhecimento, expõe o relacionamento entre atributos, expões o relacionamento entre classes e expões o relacionamento entre atributos e classes.

Técnica	Interativo	Escalabilidade	Agrupamento	Classificação	Expertise do Usuário	Seleção de Atributos	Descoberta de Conhecimento	Atributo vs Atributo	Classe vs Classe	Atributo vs Classe
(DY; BRODLEY, 2000b)	✔	✗	✔	✗	✗	✔	✗	✔	✗	✗
(YANG *et al.*, 2003)	✔	✔	✔	✗	✗	✗	✔	✔	✗	✗
(GUO *et al.*, 2003)	✔	✔	✔	✗	✗	✔	✗	✔	✗	✗
(SEO; SHNEIDERMAN, 2005)	✔	✔	✔	✗	✗	✔	✔	✔	✗	✗
(MAY *et al.*, 2011)	✔	✔	✗	✔	✗	✔	✗	✔	✗	✗
(TURKAY; FILZMOSER; HAUSER, 2011)	✔	✔	✗	✗	✗	✗	✔	✔	✗	✗
(BOTELHO, 2011)	✔	✔	✔	✗	✗	✔	✗	✔	✗	✗
(CRUZ, 2012)	✔	✔	✔	✗	✗	✔	✗	✔	✗	✗
(MAMANI *et al.*, 2013)	✔	✔	✔	✔	✔	✗	✗	✔	✗	✗
(KRAUSE; PERER; BERTINI, 2014)	✔	✔	✗	✔	✗	✔	✔	✔	✗	✗
(ZHANG *et al.*, 2015)	✔	✔	✗	✔	✗	✗	✔	✗	✔	✗
(RAUBER *et al.*, 2015)	✔	✔	✗	✔	✔	✔	✔	✔	✗	✗

CAPÍTULO

3

PROJETO DE PESQUISA

Este capítulo descreve a proposta de pesquisa adotada neste trabalho. Após breves considerações iniciais, em 3.1, uma estrutura de abordagem visual para análise e seleção de atributos é descrita em 3.2, tal como suas observações iniciais. Os próximos passos para o desenvolvimento são expostos em 3.3. A conclusão e discussão sobre a metodologia são realizadas em 3.4.

3.1 Considerações iniciais

Este trabalho tem a proposta de fornecer apoio visual à tarefa de análise no espaço de atributos, evidenciando os relacionamentos destes com demais entidades do conjunto de dados. A abordagem dá condições ao usuário de realizar seleção de atributos com qualidade semelhante, ou eventualmente melhor, aos métodos automáticos. Ao mesmo tempo, flexibiliza a seleção através da intervenção humana, que insere seu conhecimento implícito ao processo.

3.2 Resultados iniciais: análise de atributos baseada em correlação com projeção RadViz

A ideia básica da abordagem visual é codificar as informações contidas na matriz de correlação (extraídas pelo cálculo do coeficiente de Pearson) a partir de uma projeção multidimensional baseada em pontos. Uma representação semanticamente intensa implica em maior capacidade de instigar a habilidade analítica do usuário, permitindo tirar proveito do potencial cognitivo natural humano. Esta seção descreve o processo, desde o tratamento inicial dos dados, até as primeiras observações extraídas pela abordagem.

Antes de disponibilizar a visualização ao usuário, o modelo passa por algumas etapas básicas. A abordagem em desenvolvimento é resumida na Figura 13, sendo detalhada nas próximas subseções, que descrevem a forma da realização do cálculo da matriz de correlação

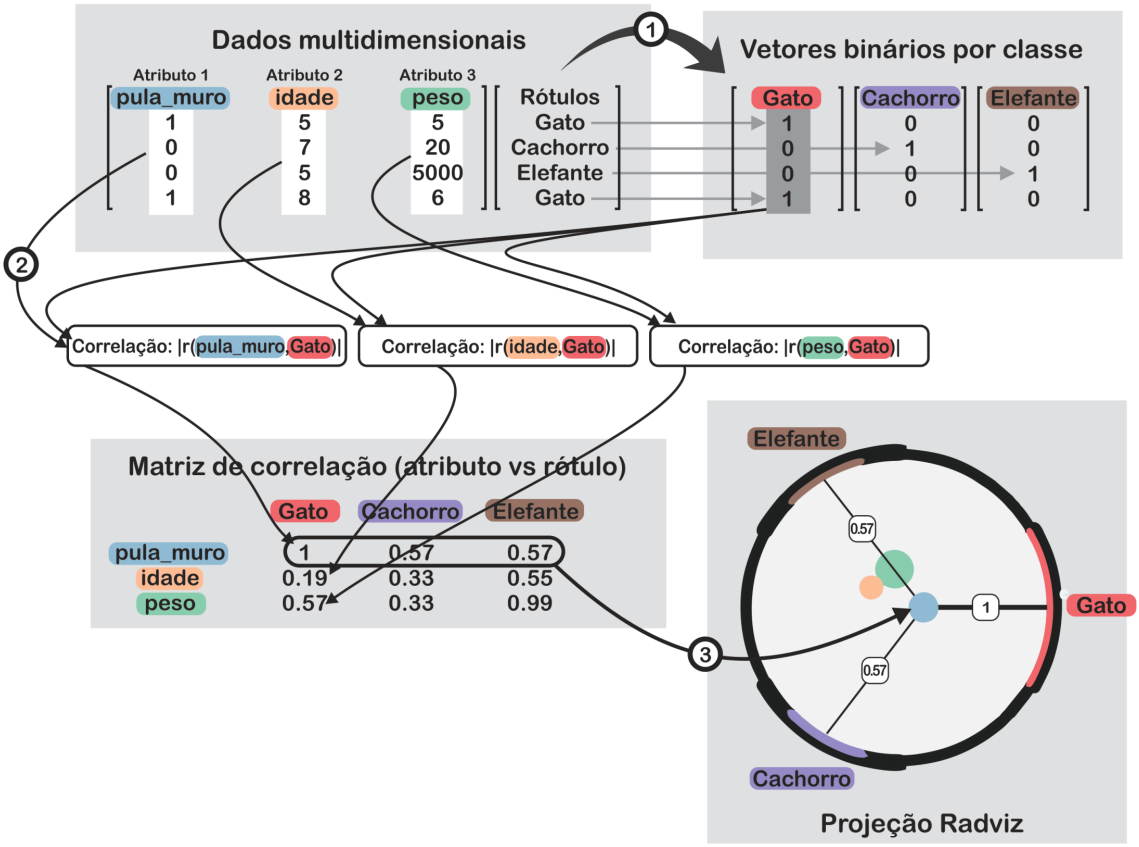

Figura 13 – Resumo gráfico da abordagem proposta. (1) Decomposição dos rótulos em vetores binários separando as classes. (2) Cálculo de correlação entre cada classe com demais atributos do conjunto de dados. (3) Projeção da matriz de correlação resultante a partir da técnica de RadViz.

Fonte: Elaborada pelo autor.

atributo *versus* classe em 3.2.1, e metodologia de projeção da matriz em 3.2.2, a interatividade com usuário na subseção 3.2.3, visualização dupla em 3.2.4, e, por fim, algumas observações iniciais podem ser examinadas em 3.2.5.

3.2.1 Matriz de correlação (atributos versus classe)

A análise por correlação quantifica as associações entre variáveis podendo expor o quanto cada par destas são relacionadas e o quão forte é esse relacionamento (ZHANG *et al.*, 2015). Os métodos convencionais baseados em correlação geralmente fazem o cálculo a cada par de atributos, montando uma matriz de correlação. O modelo proposto por este trabalho faz uma abordagem diferente, calculando a correlação entre os atributos e uma versão binária de cada possível rótulo dos dados. Para tanto, é necessária a decomposição do vetor que representa os k rótulos em k vetores virtuais com valores binários simbolizando cada rótulo (passo (1) da Figura 13).

Para a montagem da matriz base para subsequente projeção, é realizado o cálculo de correlação de Pearson (COHEN *et al.*, 2013) entre cada atributo contra os vetores virtuais

binários representantes de cada classe (passo (2) da Figura 13). A correlação de Pearson está entre as mais populares empregadas para quantificar o relacionamento entre duas variáveis. A equação 3.1 descreve o cálculo do coeficiente de Pearson, onde x e y são vetores de mesmo tamanho, $\bar{x}$ e $\bar{y}$ são as médias aritméticas. Os valores variam de +1 a -1, onde o sinal implica na direção da relação e a magnitude está relacionada com a intensidade da correlação.

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

A matriz resultante tem dimensões $k \times m$, onde cada célula armazena o valor do coeficiente de correlação entre o atributo y_i e a classe representada pelo vetor y_{lj}. Por fim, a matriz é transposta com finalidade de projetar atributos como elementos da visualização.

3.2.2 Projeção RadViz

RadViz é uma técnica de visualização que representa as dimensões do conjunto de dados através de pontos conhecidos como Âncoras Dimensionais (DA), distribuídos em volta da circunferência da área de plotagem de forma inicialmente equidistante. Os elementos (instâncias) são projetados de acordo com a influência de aproximação por cada DA, em formato análogo a um sistema de molas. Os valores de atração para cada DA são geralmente normalizados evitando discrepâncias no posicionamento dos elementos. A Figura 14 apresenta a projeção.

As equações 3.2 e 3.3 são aplicadas para cada ponto a ser projetado, onde x_i e y_i são as coordenadas transformadas para a instância i, θ_j é a posição angular da dimensão j representada na circunferência da área de projeção, a_{ij} é o valor da instância i na dimensão j, m e n são as quantidades de dimensões e instâncias, respectivamente.

$$x_i = \frac{\sum_1^d a_{i,j} \cos\theta}{\sum_1^d a_{i,j}} \tag{3.2}$$

$$y_i = \frac{\sum_1^d a_{i,j} \sin\theta}{\sum_1^d a_{i,j}} \tag{3.3}$$

RadViz é uma projeção simples, intuitiva e semanticamente intensa. Contudo, esta projeção enfrenta vários problemas, dentre os principais, *overlapping* e *visual clutter*. Abordagens para aumento da robustez desta técnica devem ser investigadas e subsequentemente melhorias devem ser propostas.

3.2.3 Interação com usuário

O foco da abordagem em desenvolvimento é proporcionar, além da análise, seleção de atributos com qualidade similar aos algoritmos automáticos e adicionalmente promover uma

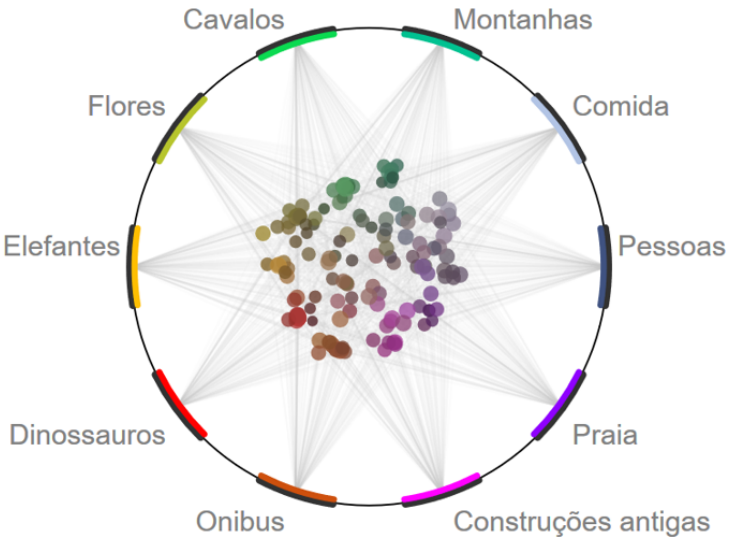

Figura 14 – Técnica de projeção RadViz. Os elementos posicionados sob a circunferência representando as âncoras dimensionais são os possíveis rótulos do conjunto de dados. Os elementos projetos internamente são atributos, onde o tamanho codifica o nível de correlação com o as classes e a cor é resultante da contribuição de correlação de cada rótulo.

Fonte: Elaborada pelo autor.

visualização que forneça ao usuário meios de revelar informações pertinentes. Com base nisso, é desejável que o usuário enriqueça seu entendimento sobre o conjunto de dados e empregue seu conhecimento para a realização da seleção.

Interativamente deve ser possível: observar a relevância dos atributos por rótulo; realizar a seleção de atributos, marcando visualmente os selecionados com fim de manter um mapa cognitivo da distribuição da seleção; destacar áreas de interesse do usuário; reordenar as âncoras dimensionais; definir parâmetros de amostras do conjunto de dados.

3.2.4 Dual RadViz

Complementando a análise dos dados no espaço de atributos, o usuário pode optar em visualiza-los no espaço de instâncias a partir de uma segunda projeção RadViz, agora em sua forma convencional. É possível expor a representação corrente da seleção de atributos e manuseá-la buscando a melhor organização visual. Uma abordagem semelhante é desenvolvida em (TURKAY; FILZMOSER; HAUSER, 2011). A figura 15 resume o modelo da visualização dupla.

Visando a melhoria da seleção de atributos e incremento da interpretação dos dados, algumas interações relacionadas às duas visualizações são possíveis. No espaço de atributos, ao selecionar um item, o mesmo é inserido na segunda visualização como uma âncora dimensional, e seus valores passam a exercer influência na projeção. De forma inversa, ao selecionar itens do espaço de instâncias, os rótulos dos mesmos se destacam na primeira visualização, apontando onde o usuário deve focar para um desejado ajuste.

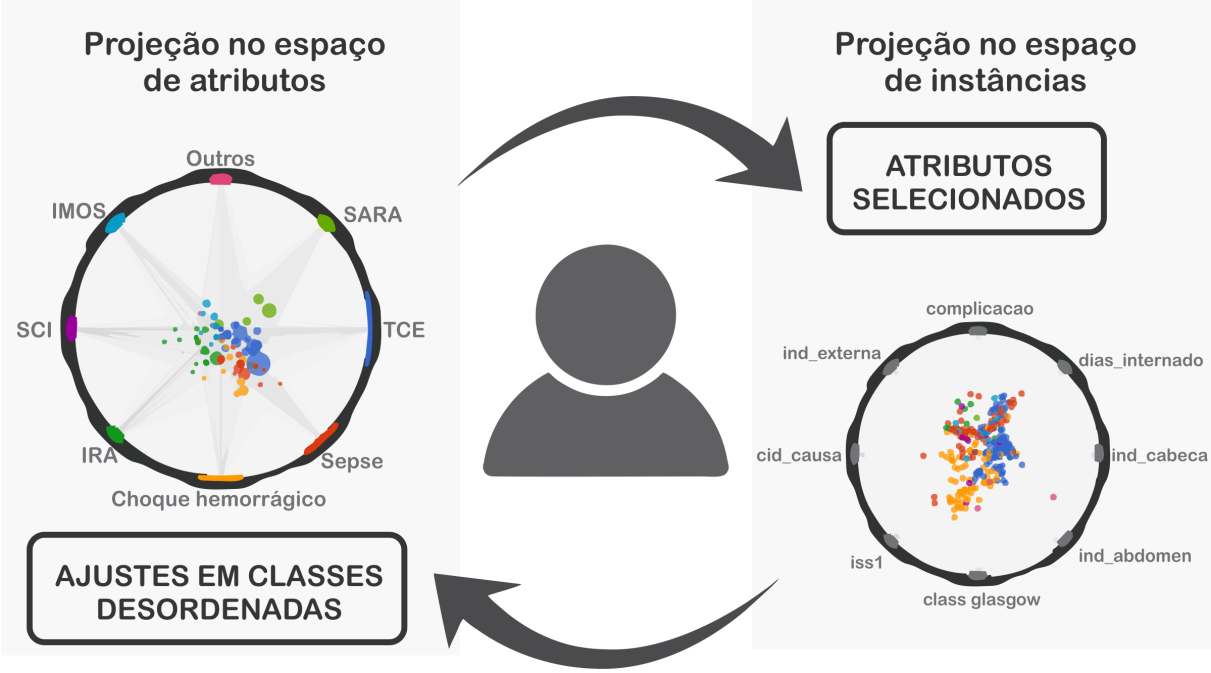

Figura 15 – Modelo de interação do *dual RadViz*. Ao selecionar atributos na primeira representação, os mesmos são expostos e modificam a segunda. Eventual desordem visual pode ser detectada no espaço de instâncias, possibilitando ao usuário selecionar áreas onde é necessário ajustes (nova seleção/alteração de atributos).

Fonte: Elaborada pelo autor.

3.2.5 Observações iniciais

Nesta subseção são apresentados resultados preliminares obtidos no atual estágio de desenvolvimento da proposta. Um estudo de caso empregando o conjunto de dados de registros em saúde (apresentado em 1.5.1) é realizado com a exposição de algumas observações obtidas.

Estudo de caso

Conforme mencionado anteriormente, o conjunto de dados deste estudo de caso é proveniente da coleta de informações de trauma e registros em saúde do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo (HC-FMRP-USP), realizada em um intervalo de oito anos (2006 a 2014). Na base estão presentes informações do paciente, dados acerca do evento do trauma, dados clínicos e índices calculados de trauma. Mais detalhes acerca dos índices de trauma podem ser encontrados em (JÚNIOR *et al.*, 1999) e (FUGIMOTO *et al.*, 2009).

Dois cenários são explorados para coleta de observações. O primeiro verifica a importância dos atributos em relação à condição final do paciente, podendo ser: “boa recuperação”, “limitações moderadas”, “limitações graves”, “óbito”, “estado vegetativo permanente”, “alta à pedido”, “evasão” e “transferência”. No segundo cenário, é analisada a importância dos atributos diante das possíveis causas de óbito: “choque hemorrágico”, “trauma encéfalo craniano (TCE)”, “sepse”, “arritmia”, “síndrome da angústia respiratória aguda (SARA)”, “insuficiencia respiratoria

aguda (IRA)",“ insuficiência de múltiplos órgãos e sistemas (IMOS)” e “outros”.

Condição final do paciente

Neste cenário o atributo selecionado como vetor de rótulos é intitulado de “condicao_-alta_i”. A Figura 16a exibe o estado inicial da visualização, expondo os atributos importantes em relação ao conjunto completo de rótulos, com destaque para os atributos “cirurgia_i”(indica se o paciente passou por procedimento cirúrgico), “ind_torax_1”, “ind_face_1”, “ind_extremidade_-1”, “ind_abdomem_1” e “ind_cabeca_1”, sendo atributos que armazenam níveis de gravidade das lesões pelas regiões do organismo e servem de base para outros índices de trauma. Estes atributos tem correlação de maior intensidade com os rótulos “Boa Recuperação” e “Limitações Moderadas” e moderada em “Limitações Graves” (Figura 16c).

Outra observação importante é perceptível na análise do rótulo “Estado Vegetativo Permanente”, onde nenhum atributo revelou alguma correlação significante. É possível perceber também que nenhum índice de trauma aponta para chances do paciente entrar nesta condição. A Figura 16b ilustra este fenômeno. A tentativa de predição para essa confição final, considerando este escopo de atributos, é impraticável.

Neste contexto é dada atenção diferenciada aos casos de óbito, onde é desejável o máximo de precisão nos índices que buscam sua predição. Como é possível observar na Figura 16d, os atributos “ntris_ue”, “rts_ue”, “pas”, “class_glasgow”, “niss_ue” e “iss_ue” tiveram alta correlação. O índice “ntris_ue” obteve o maior coeficiente seguido de perto pelo “tris_ue” e “iss_ue’, informações semelhantes às conclusões das pesquisas em (DOMINGUES *et al.*, 2011) e (DOMINGUES *et al.*, 2015).

Causa do óbito do paciente

Em circunstância de óbito do paciente, é registrada na base de dados a sua causa, compondo o atributo que serve de rótulo para este cenário. A motivação desta análise é, havendo a predição de alta probabilidade de óbito do paciente, qual seria a causa e quais medidas podem ser tomadas para evita-la? A Figura 17a expõe o estado inicial da visualização, onde é perceptível que o atributo mais importante no ponto de vista de todas as classes é “ind_cabeca”.

A Figura 17b apresenta os atributos pertinentes ao caso de “TCE”. O “ind_cabeca”, que já estava destacado diante do conjunto inteiro, passa a ter mais importância nesse contexto. Este atributo contém os registros do índice de avaliação de gravidade da região crânio-encefálica (incluindo face). Outros atributos pertinentes são: “iss_ue” e “class_glasgow”.

Na Figura 17c é possível observar a importância dos atributos para os casos de “Sepse”. Dois atributos possuem maior destaque: “dias_internacao” (contagem dos dias que o paciente permaneceu internado) e “complicacao_i”.

Os atributos de maior correlação para o caso de “Choque Hemorrágico” são apresentados

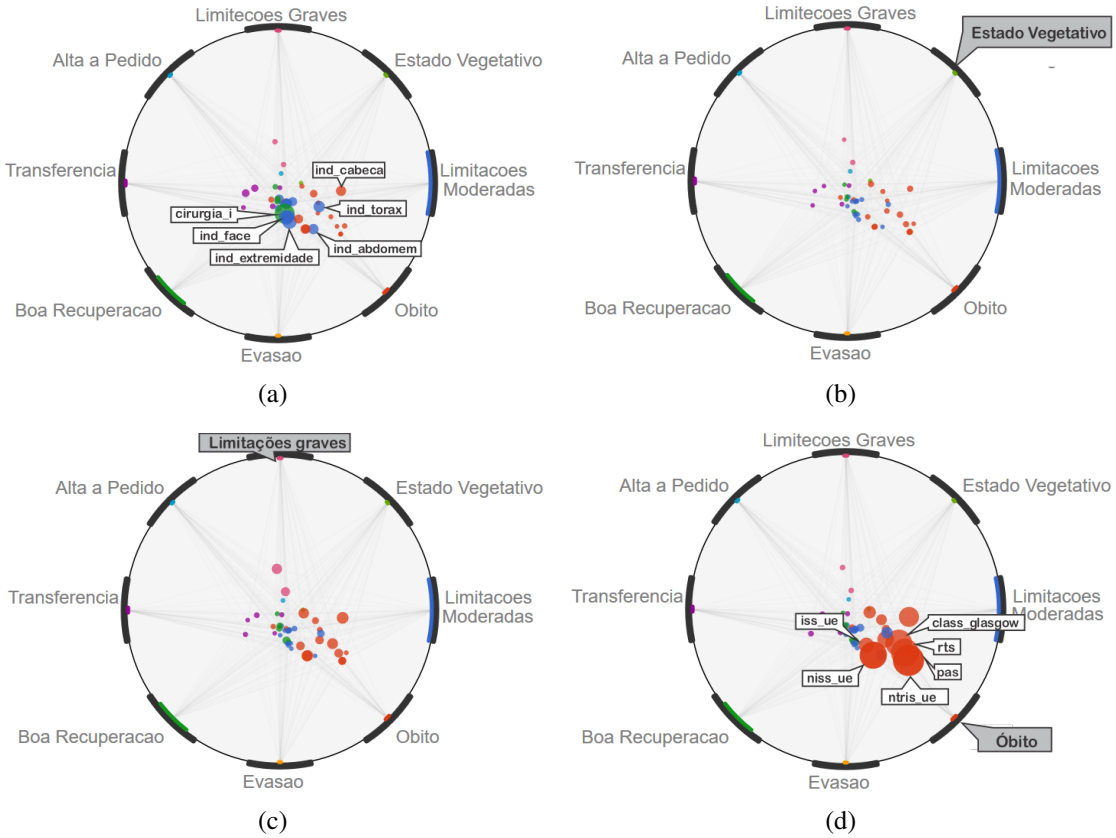

Figura 16 – Aplicação da proposta no conjunto de dados de registros em saúde. Em (a) é possível observar o estado inicial e os atributos pertinentes a todos os rótulos. Em (b) é selecionado (*mouse over*) apenas o rótulo "Estado Vegetativo" mostrando que não há bons atributos para sua predição. Em (c) é selecionado o rótulo "Limitações Graves" e, diferente de "Estado Vegetativo", alguns atributos possuem relevância. Na última figura (d), o rótulo "Óbito" é selecionado e expõe a alta relevância dos índices de trauma.

Fonte: Elaborada pelo autor.

na Figura 17d. Nela fica explícita a importância dos atributos "ind_abdomem", "ind_torax" e "pas", que registram o índice de avaliação de gravidade do abdomem, torax e a pressão arterial sistólica, respectivamente.

3.3 Desenvolvimento da pesquisa

Algumas lacunas seguem abertas e devem ser foco de investigação. Visando cumprir o objetivo desta proposta, esta seção descreve novos direcionamentos no desenvolvimento da abordagem e condução da pesquisa.

3.3.1 Método de avaliação interativa

Avaliar a qualidade da seleção de atributos de maneira interativa não é tarefa simples diante do possível alto custo computacional demandado. Frequentemente é necessário computar milhares, ou até mesmo milhões, de instâncias em uma janela de tempo curta que garanta a

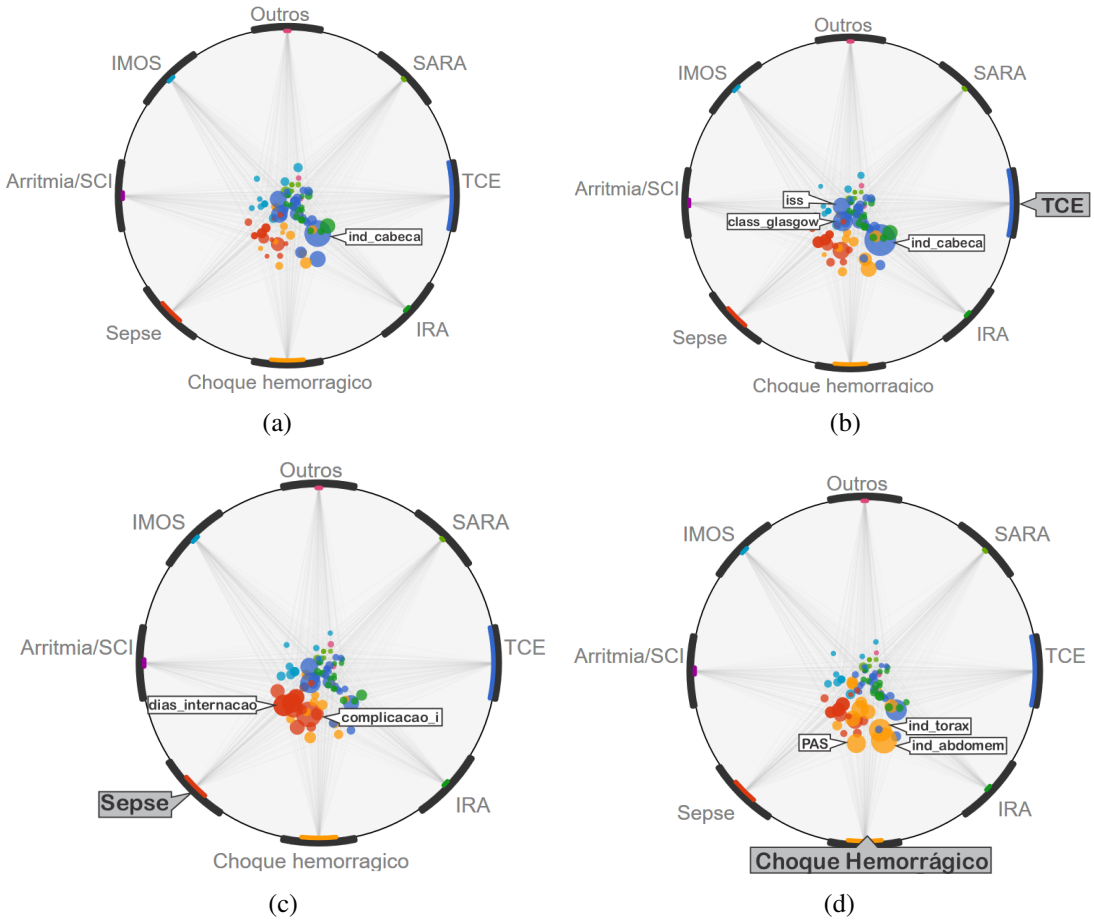

Figura 17 – Aplicação da abordagem proposta com o conjunto de dados de registros em saúde considerando como vetor de rótulos o atributo "obito_razao". Em (a) é exposto o estado inicial da visualização. Em (b), (c) e (d) os rótulos "TCE", "Sepse" e "Choque Hemorrágico", respectivamente, são selecionados.

Fonte: Elaborada pelo autor.

interatividade, podendo inviabilizar o processo.

Como apresentado no Capítulo 2, os métodos de classificação podem servir para realizar a avaliação (geralmente através da medida de acurácia), mas se tornam proibitivos diante da demanda computacional de o fazer de forma interativa.

Atualmente, neste projeto, a avaliação (ou estimativa de qualidade) da seleção é realizada a partir de uma segunda visualização (também empregando a projeção RadViz). Contudo, deverá ser investigada a aplicação de uma técnica de projeção baseada em pontos rápida e que reflita a qualidade da seleção de atributos do usuário.

3.3.2 Melhoria na estimativa de correlação na manipulação de dados heterogêneos

Apesar da abordagem apresentada em (ZHANG *et al.*, 2015) maximizar os valores dos coeficientes de correlação em casos do emprego de variáveis numéricas e categóricas, alguns

casos patológicos não são tratados. Situações com repetição excessiva de valores nulos, ou o inverso, onde todos os valores de um dos vetores são diferentes, causam incoerências no cálculo.

Formas de realização do cálculo de correlação para tratamento de dados heterogêneos devem ser investigados. A precisão dos coeficientes de correlação refletem diretamente nas informações repassadas ao usuário, influenciando na qualidade da seleção do subconjunto de atributos.

3.3.3 Mapeamento de instâncias predizíveis por atributo

Bons algoritmos de seleção de atributos são capazes de mensurar a relevância e filtrar eventuais atributos redundantes. Um problema comum é que atributos aparentemente redundantes podem ser, na verdade, complementares. Dois atributos que possuem alta correlação com determinada classe podem, eventualmente, serem filtrados na seleção, mesmo havendo a possibilidade destes representarem informações para predição de instâncias diferentes dentro desta mesma classe.

Um modelo para mapeamento de instâncias predizíveis por atributo (dentro do escopo dos dados supervisionados) será investigado. Deste modo é possível expor uma estimativa do subconjunto ideal de atributos por classe, evitando o descarte de atributos não redundantes e ao mesmo tempo que determina a quantidade ideal de atributos a serem selecionados.

3.3.4 Extensão para emprego em outros tipos de dados

Atualmente a abordagem trabalha com dados multidimensionais contendo atributos numéricos e/ou categóricos. Será estudada sua aplicação em séries temporais como também texto. Diante da natureza diferenciada destes tipos de dados, uma série de ajustes ao método deverão ser investigados

3.4 Considerações finais

O suporte visual à etapa de seleção de atributos, além de fornecer base para a escolha de atributos relevantes, pode transmitir informações e garantir maior interpretabilidade aos dados. A abordagem proposta preenche relativamente bem estes requisitos, porém, permanece com hiatos e falhas que devem ser investigados e aperfeiçoados.

Embora outros métodos possam eventualmente ser empregados, a projeção RadViz é a técnica de visualização de dados multidimensionais adotada pela abordagem. Seu estado da arte deve ser explorado e aplicado neste trabalho. Complementarmente, contribuições para esta projeção deverão ser propostas visando o seu enriquecimento.

Apesar da abordagem possuir opção de realização da seleção de atributos, ela se mostra uma ferramenta poderosa na análise dos dados no espaço de atributos. O ponto forte da proposta

está na capacidade de expor o relacionamento entre todos os elementos envolvidos no conjunto de dados (atributos, instâncias e classes).

CAPÍTULO

4

ATIVIDADES E CRONOGRAMA

Este capítulo apresenta uma síntese das atividades previstas para o cumprimento do programa juntamente com o detalhamento do seu cronograma de acompanhamento. As atividades são elencadas a seguir e a Tabela 2 expõe seu cronograma.

4.1 Passos metodológicos

Para atingir os objetivos do projeto, este pesquisa deve ser conduzida pelo procedimento metodológico, que regem as atividades, descrito a seguir.

- Fundamentação teórica: explorar a literatura acerca dos eixos que norteiam este trabalho. A seleção de atributos, investigando os modelos tradicionais, como também os, mais recentes, modelos interativos. As projeções multidimensionais, em especial a projeção Radviz, elemento chave desta pesquisa. E métodos interativos que buscam benefício máximo do contato com usuário;

- Revisão de literatura: realizar um mapeamento na literatura existente acerca das abordagens de seleção de atributos interativas com suporte visual, levantando suas características sinalizando as vantagens e desvantagens. O propósito é conhecer o estado da arte da área, reconhecer lacunas existentes, para assim propor novas funcionalidades que representem avanço científico;

- Desenvolvimento teórico: investigar e formalizar nova abordagem para seleção interativa de atributos apoiada pela projeção Radviz. Registrar a metodologia da proposta em documento científico.

- Desenvolvimento de algoritmos: fundamentado no desenvolvimento teórico, projetar novos algoritmos para seleção interativa de atributos e propor melhorias na abordagem tradicional Radviz.

- Delineamento experimental e validação: a avaliação da abordagem deve ser realizada em duas perspectivas diferentes envolvendo análises qualitativas e quantitativas. O primeiro ponto de vista refere-se à análise semântica, buscando examinar a qualidade visual e a riqueza de significado da metodologia de projeção. O segundo ponto de vista busca analisar quantitativamente a capacidade e precisão de realização de seleção de atributos. Para ambas perspectivas, o confrontamento com o estado da arte a partir de, preferencialmente, testes de *benchmark* será praticado.

- Aplicações: após desenvolvimento em contexto genérico, conjuntos de dados específicos serão empregados à abordagem a fim de aplicá-la em contextos do mundo real. Atenção especial é pretendida aos registros de saúde, que em testes preliminares se mostraram satisfatoriamente adequadas.

- Ajustes e publicações: durante o desenvolvimento, eventualmente surgirá a necessidade de realização de ajustes. Os mesmos serão realizados e validados visando alcance de nível de maturidade e, assim, viabilizando o foco em publicações da abordagem e suas aplicações.

4.2 Atividades e cronograma

1. Integralização de créditos obrigatórios exigidos pelo programa de Pós-Graduação, nível de Doutorado, tendo estas sido cumpridas parcialmente em Teresina, Piauí;

2. Revisão e acompanhamento bibliográfico acerca dos eixos temáticos deste trabalho, incluindo projeções multidimensionais, técnicas de seleção de atributos e modelos estatísticos com ênfase em análise de correlação;

3. Investigação e formalização de abordagem para apoio visual à seleção de atributo;

4. Desenvolvimento e implementação das abordagens formalizadas;

5. Realização de exame de proficiência em língua inglesa;

6. Inscrição e submissão de documento de monografia referente ao exame de qualificação do programa;

7. Submissão de artigo ao IEEE Conference on Visual Analytics Science and Technology (IEEE VAST 2017) expondo resultados preliminares da abordagem;

8. Realização do exame de qualificação diante de membros da comissão examinadora;

9. Avaliação experimental com ajustes na abordagem e coleta de resultados diante da aplicação com conjuntos de dados reais em especial aos registros em saúde provenientes da Faculdade de Medicina de Ribeirão Preto;

10. Estágio sanduíche mediante a visita a um centro de pesquisa no exterior;

11. Publicação dos resultados e pontos relevantes do modelo com suas aplicações através de relatórios técnicos e artigos direcionados às revistas e eventos da área;

12. Redação, defesa e entrega da tese.

Tabela 2 – Atividades planejadas previstas para este programa de Pós-graduação. Lacunas preenchidas em preto representam atividades já realizadas. Em cinza, atividades futuras.

		Atividades											
Trimestre		1	2	3	4	5	6	7	8	9	10	11	12
2015	3º	■											
	4º	■											
2016	1º	■	■										
	2º	■	■										
	3º	■	■	■									
	4º	■	■	■	■	■				■			
2017	1º		■	■	■		■	■	■	■			
	2º		■	■	■					■		■	
	3º		■	■	■					■	■	■	
	4º		■	■	■					■	■	■	
2018	1º		■		■					■	■	■	
	2º		■							■	■	■	
	3º		■							■		■	■
	4º		■							■		■	■
2019	1º		■										■
	2º												■

O desenvolvimento deste projeto de doutorado, sob orientação da Prof.ª Dr. Rosane Minghim, conta com apoio do Prof. Dr. Alexandru Telea do Instituto Johann Bernoulli, Universidade de Groningen. O professor teve contato inicial com o projeto em visita recente ao Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, e desde então colabora no seu desenvolvimento. O projeto conta também com a participação do grupo de pesquisa em trauma liderado pelo Prof. Dr. Gerson Alves Pereira Júnior da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo.

REFERÊNCIAS

ANG, J. C.; MIRZAL, A.; HARON, H.; HAMED, H. N. A. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. **IEEE/ACM transactions on computational biology and bioinformatics**, IEEE, v. 13, n. 5, p. 971–989, 2016. Citado na página 26.

BATTITI, R. Using mutual information for selecting features in supervised neural net learning. **IEEE Transactions on neural networks**, IEEE, v. 5, n. 4, p. 537–550, 1994. Citado na página 31.

BOTELHO, G. M. **Seleção de características apoiada por mineração visual de dados**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, University of São Paulo, São Carlos, 2011. Citado 5 vezes nas páginas 9, 37, 38, 39 e 40.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. **Computers & Electrical Engineering**, Elsevier, v. 40, n. 1, p. 16–28, 2014. Citado na página 18.

COHEN, J.; COHEN, P.; WEST, S. G.; AIKEN, L. S. **Applied multiple regression/correlation analysis for the behavioral sciences**. [S.l.]: Routledge, 2013. Citado na página 42.

CRUZ, L. E. F. **Uma abordagem baseada em técnicas de visualização de informações para avaliação de características de imagens e aplicações**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, University of São Paulo, São Carlos, 2012. Citado 4 vezes nas páginas 9, 37, 38 e 40.

DASH, M.; LIU, H. Feature selection for classification. **Intelligent Data Analysis**, v. 1, p. 131–156, 1997. Citado na página 18.

____. Consistency-based search in feature selection. **Artificial intelligence**, Elsevier, v. 151, n. 1-2, p. 155–176, 2003. Citado 4 vezes nas páginas 26, 27, 29 e 30.

DOMINGUES, C. d. A.; NOGUEIRA, L. d. S.; SETTERVALL, C. H. C.; SOUSA, R. M. C. d. Desempenho dos ajustes do Trauma and Injury Severity Score (TRISS): revisão integrativa. **Revista da Escola de Enfermagem da USP**, scielo, v. 49, p. 138 – 146, 12 2015. ISSN 0080-6234. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0080-62342015000700138&nrm=iso>. Citado na página 46.

DOMINGUES, C. d. A.; 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. Citado na página 46.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern classification**. [S.l.]: John Wiley & Sons, 2012. Citado na página 31.

DY, J. G.; BRODLEY, C. E. Feature subset selection and order identification for unsupervised learning. In: CITESEER. **ICML**. [S.l.], 2000. p. 247–254. Citado na página 36.

____. 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. Disponível em: <http://doi.acm.org/10.1145/347090.347168>. Citado 2 vezes nas páginas 35 e 40.

FLEURET, F. Fast binary feature selection with conditional mutual information. **Journal of Machine Learning Research**, v. 5, n. Nov, p. 1531–1555, 2004. Citado na página 32.

FUGIMOTO, P. M.; SALES, L. D. F.; JÚNIOR, G. A. P.; PASSOS, A. D. C.; ALVES, D.; BARANAUSKAS, J. A. Análise comparativa entre árvores de decisão e triss na predição de sobrevida de pacientes traumatizados. In: **IV Congresso da Academia Trinacional de Ciências**. [S.l.: s.n.], 2009. p. 10–20. Citado na página 45.

GU, Q.; LI, Z.; HAN, J. Generalized fisher score for feature selection. **arXiv preprint arXiv:1202.3725**, 2012. Citado na página 18.

GUO, D.; GAHEGAN, M.; PEUQUET, D.; MACEACHREN, A. Breaking down dimensionality: Effective and efficient feature selection for high-dimensional clustering. In: **Workshop on Clustering High-Dimensional Data**. [S.l.: s.n.], 2003. Citado 5 vezes nas páginas 9, 36, 37, 39 e 40.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **J. Mach. Learn. Res.**, JMLR.org, v. 3, p. 1157–1182, mar. 2003. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944919.944968>. Citado 4 vezes nas páginas 18, 25, 26 e 28.

GUYON, I.; WESTON, J.; BARNHILL, S.; VAPNIK, V. Gene selection for cancer classification using support vector machines. **Machine learning**, Springer, v. 46, n. 1-3, p. 389–422, 2002. Citado na página 39.

HE, X.; CAI, D.; NIYOGI, P. Laplacian score for feature selection. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2005. p. 507–514. Citado na página 18.

HOFFMAN, P.; GRINSTEIN, G.; MARX, K.; GROSSE, I.; STANLEY, E. Dna visual and analytic data mining. In: IEEE. **Visualization'97., Proceedings**. [S.l.], 1997. p. 437–441. Citado na página 19.

HOFFMAN, P.; GRINSTEIN, G.; PINKNEY, D. Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations. In: ACM. **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**. [S.l.], 1999. p. 9–16. Citado na página 19.

HOQUE, N.; BHATTACHARYYA, D.; KALITA, J. K. Mifs-nd: a mutual information-based feature selection method. **Expert Systems with Applications**, Elsevier, v. 41, n. 14, p. 6371–6385, 2014. Citado na página 18.

JOVIĆ, A.; BRKIĆ, K.; BOGUNOVIĆ, N. A review of feature selection methods with applications. In: IEEE. **Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on**. [S.l.], 2015. p. 1200–1205. Citado na página 29.

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. Citado na página 45.

KEIM, D. A.; MANSMANN, F.; SCHNEIDEWIND, J.; ZIEGLER, H. Challenges in visual data analysis. In: IEEE. **Information Visualization, 2006. IV 2006. Tenth International Conference on**. [S.l.], 2006. p. 9–16. Citado 2 vezes nas páginas 17 e 19.

KRAUSE, J.; PERER, A.; BERTINI, E. Infuse: interactive feature selection for predictive modeling of high dimensional data. **IEEE transactions on visualization and computer graphics**, IEEE, v. 20, n. 12, p. 1614–1623, 2014. Citado 5 vezes nas páginas 20, 32, 38, 39 e 40.

LANGLEY, P. Selection of relevant features in machine learning. In: **In Proceedings of the AAAI Fall symposium on relevance**. [S.l.]: AAAI Press, 1994. p. 140–144. Citado na página 28.

LEE, H. D. **Seleção de atributos importantes para a extração de conhecimento de bases de dados**. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, University of São Paulo, São Carlos, 2005. Citado 3 vezes nas páginas 26, 27 e 28.

LEWIS, D. D. Feature selection and feature extraction for text categorization. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Proceedings of the workshop on Speech and Natural Language**. [S.l.], 1992. p. 212–217. Citado na página 31.

LI, J.; CHENG, K.; WANG, S.; MORSTATTER, F.; TREVINO, R. P.; TANG, J.; LIU, H. Feature selection: A data perspective. **arXiv preprint arXiv:1601.07996**, 2016. Citado 6 vezes nas páginas 17, 25, 27, 28, 29 e 30.

LIN, D.; TANG, X. Conditional infomax learning: an integrated framework for feature extraction and fusion. In: SPRINGER. **European Conference on Computer Vision**. [S.l.], 2006. p. 68–82. Citado na página 32.

LIU, H.; MOTODA, H. **Feature Extraction, Construction and Selection: A Data Mining Perspective**. Norwell, MA, USA: Kluwer Academic Publishers, 1998. ISBN 0792381963. Citado na página 18.

LIU, H.; SETIONO, R. Chi2: Feature selection and discretization of numeric attributes. In: IEEE. **Tools with artificial intelligence, 1995. proceedings., seventh international conference on**. [S.l.], 1995. p. 388–391. Citado na página 32.

LIU, H.; YU, L. Toward integrating feature selection algorithms for classification and clustering. **IEEE Trans. on Knowl. and Data Eng.**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 17, n. 4, p. 491–502, abr. 2005. ISSN 1041-4347. Disponível em: <http://dx.doi.org/10.1109/TKDE.2005.66>. Citado na página 28.

MAMANI, G. M. H.; FATORE, F. M.; NONATO, L. G.; PAULOVICH, F. V. User-driven feature space transformation. In: **Proceedings of the 15th Eurographics Conference on Visualization**. Chichester, UK: The Eurographs Association & John Wiley & Sons, Ltd., 2013. (EuroVis '13), p. 291–299. Disponível em: <http://dx.doi.org/10.1111/cgf.12116>. Citado 5 vezes nas páginas 9, 34, 35, 39 e 40.

MAY, T.; BANNACH, A.; DAVEY, J.; RUPPERT, T.; KOHLHAMMER, J. Guiding feature subset selection with an interactive visualization. In: IEEE. **Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on**. [S.l.], 2011. p. 111–120. Citado 2 vezes nas páginas 36 e 40.

MEYER, P. E.; BONTEMPI, G. On the use of variable complementarity for feature selection in cancer classification. In: SPRINGER. **Workshops on Applications of Evolutionary Computation**. [S.l.], 2006. p. 91–102. Citado na página 31.

NIE, F.; XIANG, S.; JIA, Y.; ZHANG, C.; YAN, S. Trace ratio criterion for feature selection. In: **AAAI**. [S.l.: s.n.], 2008. v. 2, p. 671–676. Citado na página 31.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V. *et al.* Scikit-learn: Machine learning in python. **Journal of Machine Learning Research**, v. 12, n. Oct, p. 2825–2830, 2011. Citado na página 32.

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**, IEEE, v. 27, n. 8, p. 1226–1238, 2005. Citado na página 31.

RAGHAVAN, H.; MADANI, O.; JONES, R. Interactive feature selection. In: **Proceedings of the 19th International Joint Conference on Artificial Intelligence**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. (IJCAI'05), p. 841–846. Disponível em: <http://dl.acm.org/citation.cfm?id=1642293.1642428>. Citado na página 18.

RAUBER, P. E.; SILVA, R. da; FERINGA, S.; CELEBI, M. E.; FALCÃO, A. X.; TELEA, A. C. Interactive image feature selection aided by dimensionality reduction. In: **EuroVis Workshop on Visual Analytics (EuroVA). The Eurographics Association**. [S.l.: s.n.], 2015. Citado 2 vezes nas páginas 39 e 40.

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. Citado 2 vezes nas páginas 18 e 31.

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. Disponível em: <http://dx.doi.org.ez67.periodicos.capes.gov.br/10.1057/palgrave.ivs.9500091>. Citado 5 vezes nas páginas 9, 33, 34, 39 e 40.

SHARKO, J.; GRINSTEIN, G.; MARX, K. A. Vectorized radviz and its application to multiple cluster datasets. **IEEE transactions on Visualization and Computer Graphics**, IEEE, v. 14, n. 6, p. 1444–1427, 2008. Citado na página 19.

TSOUMAKAS, G.; KATAKIS, I. Multi-label classification: An overview. **Int J Data Warehousing and Mining**, v. 2007, p. 1–13, 2007. Citado na página 19.

TURKAY, C.; FILZMOSER, P.; HAUSER, H. Brushing dimensions—a dual visual analysis model for high-dimensional data. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 17, n. 12, p. 2591–2599, dez. 2011. ISSN 1077-2626. Disponível em: <http://dx.doi.org/10.1109/TVCG.2011.178>. Citado 4 vezes nas páginas 34, 39, 40 e 44.

WRIGHT, S. The interpretation of population structure by f-statistics with special regard to systems of mating. **Evolution**, JSTOR, p. 395–420, 1965. Citado na página 32.

YANG, H. H.; MOODY, J. E. Data visualization and feature selection: New algorithms for nongaussian data. In: **NIPS**. [S.l.: s.n.], 1999. v. 12. Citado na página 31.

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. Disponível em: <http://dl.acm.org/citation.cfm?id=769922.769924>. Citado 3 vezes nas páginas 9, 33 e 40.

YU, L.; LIU, H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: **ICML**. [S.l.: s.n.], 2003. v. 3, p. 856–863. Citado 3 vezes nas páginas 18, 31 e 39.

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**, IEEE, v. 21, n. 2, p. 289–303, 2015. Citado 8 vezes nas páginas 9, 18, 35, 36, 39, 40, 42 e 48.