Empregando técnicas de visualização de
informação para transformação interativa de
dados multidimensionais

Francisco Morgani Fatore

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Francisco Morgani Fatore

Empregando técnicas de visualização de informação para transformação interativa de dados multidimensionais

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Fernando Vieira Paulovich

USP – São Carlos
Junho de 2015

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

M254e   Morgani Fatore, Francisco
              Empregando técnicas de visualização de informação
              para transformação interativa de dados
              multidimensionais / Francisco Morgani Fatore;
              orientador Fernando Vieira Paulovich. -- São
              Carlos, 2015.
              74 p.

              Tese (Doutorado - Programa de Pós-Graduação em
              Ciências de Computação e Matemática Computacional) --
              Instituto de Ciências Matemáticas e de Computação,
              Universidade de São Paulo, 2015.

              1. visualização computacional. 2. Projeção multi-
              dimensional. 3. Biplot. 4. Aprendizado de máquina.
              I. Vieira Paulovich, Fernando, orient. II. Título.

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Francisco Morgani Fatore

Transforming multidimensional data using information visualization techniques

Master dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Master Program in Computer Science and Computational Mathematics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. Fernando Vieira Paulovich

USP – São Carlos
June 2015

A exploração de conjuntos de dados é um problema abordado com frequência em diversos domínios e tem como objetivo uma melhor compreensão de fenômenos simulados ou medidos. Tal atividade é precedida pelas etapas de coleta e armazenamento de dados que buscam registrar o máximo de detalhes sobre algum fenômeno observado. Porém, a exploração efetiva dos dados envolve uma série de desafios. Um deles é a dificuldade em identificar quais dados são realmente relevantes para as análises. Outro problema está relacionado com a falta de garantias de que os fatores fundamentais para a compreensão do problema tenham sido coletados. A transformação interativa de dados é uma abordagem que utiliza técnicas de visualização computacional para resolver ou minimizar esses problemas. No entanto, os trabalhos disponíveis na literatura possuem limitações, como interfaces demasiadamente complexas e mecanismos de interação pouco flexíveis. Assim, este projeto de mestrado teve como objetivo desenvolver novas técnicas visuais interativas para a transformação de dados multidimensionais. A metodologia desenvolvida se baseou no uso de *biplots* e na ação conjunta dos mecanismos de interação para superar as limitações das técnicas do estado da arte. Os resultados dos experimentos realizados sobre diversos conjuntos de dados dão indícios de que os métodos desenvolvidos possibilitam a obtenção de conjuntos de dados mais representativos. Mais especificamente, foram obtidos melhores resultados em tarefas de classificação de dados ao utilizar os métodos desenvolvidos.

Palavras-chave: Visualização computacional; Projeção multidimensional; *Biplot*; Aprendizado de Máquina.

Abstract

The exploration of datasets is a frequently task in several fields and aims at a better understanding of simulated or measured phenomena. Such activity is preceded by the steps of collecting and storing data, which seek to record as much detail possible about an observed phenomenon. The exploration task is challenging due to many aspects. One of them is the difficulty in identifying which collected data are actually relevant to the analysis. Another one is related to the lack of guarantees that the key factors for understanding the problem have been collected. The interactive transformation of data is a visual based approach that seeks to solve or mitigate these problems. However, the available methods in the literature have limitations in several aspects, such as complex user interfaces and inflexible interactive mechanisms. So, this master project had the goal to develop novel visual techniques for the transformation of datasets. The proposed methodology was based on the use of biplots and interaction mechanisms to overcome the limitations of the state of the art techniques. Empirical results show that by using the proposed approach, it is possible to make the data more representative. Therefore, exploratory activities, classifications, were performed more efficiently and thus provided better results.

Keywords: Computer visualization; Multidimensional projections; Biplot; Machine Learning.

Sumário

1	Introdução	1
1.1	Objetivos	3
1.2	Contribuições	3
1.3	Organização da dissertação	4
2	Trabalhos Relacionados	5
2.1	Redução de Dimensionalidade	5
2.1.1	Métodos Automáticos	6
2.1.1.1	Extração de Características	7
2.1.1.2	Seleção de Características	9
2.1.2	Métodos Iterativos	10
2.1.2.1	Matrizes de Correlação	11
2.1.2.2	Hierarquias de Dimensões	13
2.1.2.3	Mapeamento de Elementos no Plano	15
2.1.2.4	Visualização de Métodos Automáticos	18
2.2	Construção Interativa de Atributos	20
2.3	Considerações Finais	22
3	Conceitos Teóricos	25
3.1	Projeções Multidimensionais	25
3.2	Biplots	28
4	Transformação Interativa de Dados Multidimensionais	31
4.1	Considerações Iniciais	31
4.2	Mecanismo de Seleção de Itens	34
4.3	Mecanismo de Seleção de Atributos	35
4.4	Mecanismo de Combinação de Atributos	37
4.5	Mecanismo de Construção de Atributos	39

5	Experimentos e Resultados	45
5.1	Avaliação quantitativa	45
5.1.1	Mecanismo de Seleção	46
5.1.2	Mecanismo de combinação	50
5.1.3	Mecanismo de Construção	51
5.2	Avaliação qualitativa	57
5.3	Ferramenta	62
6	Conclusão	65

Lista de Figuras

2.1	Variância capturada pelos PCs	9
2.2	Matrizes de Correlação	11
2.3	Ferramenta proposta por (Guo, 2003)	12
2.4	VHDR: Visual Hierarchical Dimension Reduction	14
2.5	VaR: Value and Relation	15
2.6	Brushing Dimensions	16
2.7	DimStiller	18
2.8	iPCA	19
2.9	INFUSE	20
2.10	Ferramenta desenvolvida por Mamani et al. (2013).	22
3.1	Exemplo de projeção multidimensional.	27
3.2	Exemplo de <i>biplot</i>	29
4.1	Ciclo do processo de transformação interativa	32
4.2	<i>Biplot</i> construído a partir do conjunto de dados fictício sobre características de países utilizado como estudo de caso.	33
4.3	Exemplo de seleção sobre itens.	35
4.4	Mecanismo de seleção em lotes	36
4.5	Exemplo de combinação de atributos.	38
4.6	Combinação de atributos em lotes.	39
4.7	Exemplo de construção de atributos.	40
4.8	Visualização da incerteza do mecanismo de construção	42
5.1	Avaliação do mecanismo de seleção (1 de 2). Apresenta-se a acurácia em vermelho, o tempo relativo em azul e a dimensionalidade relativa em verde.	48
5.2	Avaliação do mecanismo de seleção (2 de 2). Apresenta-se a acurácia em vermelho, o tempo relativo em azul e a dimensionalidade relativa em verde.	49

5.3	Avaliação do mecanismo de combinação (1 de 2). Apresenta-se a acurácia em vermelho, o tempo relativo em azul e a dimensionalidade relativa em verde.	52
5.4	Avaliação do mecanismo de combinação (2 de 2). Apresenta-se a acurácia em vermelho, o tempo relativo em azul e a dimensionalidade relativa em verde.	53
5.5	Avaliação do mecanismo de criação (1 de 2). Apresenta-se a acurácia do classificador utilizando (vermelho) e sem utilizar (preto) o novo atributo. .	55
5.6	Avaliação do mecanismo de criação (2 de 2). Apresenta-se a acurácia do classificador utilizando (vermelho) e sem utilizar (preto) o novo atributo. .	56
5.7	<i>Biplot</i> inicial do conjunto de dados <i>DNA</i>	57
5.8	Seleção em lote com $\alpha = 0.1$ sobre o conjunto de dados <i>DNA</i>	58
5.9	Combinação em lote com $\beta = 0.3$ sobre o conjunto de dados <i>DNA</i>	59
5.10	Construção de 3 novas variáveis para o conjunto de dados <i>DNA</i>	60
5.11	Mapa de incerteza antes e após as transformações.	61
5.12	Ilustração da ferramenta desenvolvida.	62

Lista de Tabelas

2.1	Conjunto de dados fictício usado para ilustrações.	7
2.2	Resultado de PCA	8
2.3	Características de interesse dos principais trabalhos estudados.	24
4.1	Símbolos mais frequentemente utilizados e suas descrições.	34
5.1	Características dos conjuntos de dados utilizados na avaliação.	46

Lista de Siglas

BD	Brushing Dimensions
DOSFA	Dimension Ordering Spacing and Filtering Approach
iPCA	Interactive Principal Component Analysis
LDA	Linear Discriminant Analysis
MDS	Multidimensional Scaling
MVD	Mineração Visual de Dados
PC	Principal Component
PCA	Principal Component Analysis
SOM	Self Organizing Maps
SVD	Singular Value Decomposition
SVM	Support Vector Machine
VaR	Value and Relation
VHDR	Visual Hierarchical Dimensions Reduction

Introdução

A exploração de conjuntos de dados é um problema abordado com frequência em computação, tanto na área acadêmica quanto na indústria (Ngai et al., 2009; Harding et al., 2006). Tal exploração tem como objetivo uma melhor compreensão de fenômenos simulados ou medidos. Com base nos conhecimentos adquiridos durante a exploração, espera-se aprimorar o processo de tomadas de decisões, como previsão de condições climáticas, diagnósticos médicos, detecção de fraude, análise de mercado, etc.

As investigações sobre os dados são precedidas pelas etapas de coleta e armazenamento que podem ser realizadas por sensores, sistemas de monitoramento, simulações computacionais ou aplicações diversas. Nessas etapas, busca-se registrar o máximo de detalhes sobre o fenômeno observado. No caso de análises climáticas, por exemplo, poderiam ser observadas variáveis como temperatura, velocidade do vento, umidade do ar, etc. As observações poderiam ser realizadas em diferentes posições geográficas, onde para cada posição teria-se uma coleção de variáveis. Comumente, denomina-se cada uma dessas coleções como uma instância de dados. Como cada variável observada também pode ser chamada de dimensão, quando diversas instâncias de dados são coletadas obtém-se um conjunto de dados multidimensional.

Explorar conjuntos de dados multidimensionais envolve uma série de problemas. Um deles é a dificuldade em identificar quais dados coletados são relevantes para as análises. Isso faz com que, muitas vezes, se utilize nas investigações todas as variáveis observadas, sejam elas relevantes ou não. Essas condições impactam diretamente na complexidade das análises e também no custo computacional dos métodos que são aplicados sobre os

dados (Beyer et al., 1999). Outro problema está relacionado à falta de garantias de que os fatores fundamentais para a compreensão do problema tenham sido coletados. A maioria das aplicações está sujeita a essa situação, pois os sistemas de coleta de dados são suscetíveis a falhas e não reconhecem facilmente fatores subjetivos.

Um modo de amenizar esses problemas é transformar os dados para torná-los mais representativos para a execução das tarefas subjacentes. Uma transformação comumente aplicada é a redução de dimensionalidade. Esta tem por objetivo encontrar o menor espaço dimensional que é capaz de descrever os dados mantendo informações que são relevantes segundo algum critério. O processo de redução pode ser realizado tanto pela eliminação de dimensões irrelevantes ou redundantes quanto pela combinação entre dimensões.

Um dos problemas dos métodos de redução de dimensionalidade é que o conceito de relevância é subjetivo e pode variar de acordo com a aplicação. Além disso, os métodos tradicionais de redução apresentam uma natureza dita “caixa-preta”, pois o usuário inspeciona apenas os dados de entrada e saída, desconhecendo o processamento interno. Isso torna esses métodos pouco compreensíveis e impossibilita que o usuário contribua com a sua experiência. Para tratar essas limitações, novos métodos têm sido propostos. Eles permitem que o usuário guie o processo de redução por meio da interação com representações gráficas dos dados.

Técnicas baseadas em visualizações que permitem a interação do usuário têm sido aplicadas não somente em tarefas de redução de dimensionalidade, mas em diversas áreas de exploração de dados. Elas ganharam grande popularidade nos últimos anos (Milojević et al., 2012) e propiciaram a consolidação da área de visualização de informação (Keim, 2002). Grande parte do sucesso desta área pode ser atribuído ao uso efetivo da capacidade preemptiva da visão humana na exploração dos dados. Foi demonstrado que quando os dados são representados por gráficos, o ser humano é capaz de detectar e reconhecer padrões de forma mais fácil e rápida (Healey et al., 1995), mesmo em grandes conjuntos de dados (Fodor, 2002).

Entretanto, utilizar a capacidade preemptiva da visão humana não é a única vantagem das técnicas de visualização de informação. Ao permitirem que o usuário participe ativamente na geração dos resultados, essas técnicas viabilizam novas abordagens para explorar e transformar os dados. A construção interativa de dimensões é uma dessas novas abordagens e refere-se justamente a uma alternativa à redução de dimensionalidade para transformar os conjuntos de dados.

A necessidade de construir novas dimensões surge em situações em que a redução de dimensionalidade, seja ela automática ou interativa, não é suficiente para tornar os dados mais representativos. Isso ocorre quando fatores fundamentais para a compreensão do problema não foram capturados na etapa de coleta dos dados. Em situações como essa,

o problema só poderá ser descrito por completo ao se agregar o conhecimento do usuário nos dados.

Redução de dimensionalidade é um tema que tem sido estudado há mais de um século (Pearson, 1901) e se encontra em um estado avançado de desenvolvimento. Porém, os métodos mais populares apresentam a natureza “caixa-preta”, mencionada anteriormente, a qual inviabiliza que o usuário intervenha no processo de redução. Os métodos de redução interativa buscam dar maior participação ao usuário, mas apresentam limitações tanto em relação às visualizações em que se baseiam quanto aos mecanismos de interação propostos. Muitos desses métodos interativos partem de interfaces demasiadamente complexas e não fornecem a flexibilidade necessária para que o usuário modifique os dados adequadamente. Por sua vez, a pesquisa em construção interativa de dimensões ainda é muito recente (Mamani et al., 2013) e por isso os métodos existentes apresentam diversos aspectos que podem ser melhorados.

1.1 Objetivos

De um modo geral, o objetivo deste trabalho pode ser declarado da seguinte maneira:

“Este trabalho de mestrado tem como objetivo desenvolver mecanismos interativos para a transformação de dados multidimensionais, utilizando visualizações para permitir a intervenção do usuário em processos de redução de dimensionalidade e construção de atributos.”

Mais especificamente, busca-se que por meio de representações visuais e de mecanismos interativos, o usuário seja capaz de lidar com as duas situações descritas anteriormente. A primeira, em que é necessário reduzir a dimensionalidade dos dados para eliminar variáveis irrelevantes e redundantes. E a segunda, em que novas dimensões devem ser construídas com base no conhecimento do usuário para representar informações ausentes nos dados. Com isso, permite-se ao usuário tornar os conjuntos de dados mais representativos para tarefas subjacentes. A abordagem desenvolvida se baseou na simplicidade das visualizações e na integração dos mecanismos de interação para superar as limitações dos trabalhos do estado-da-arte.

1.2 Contribuições

A análise dos resultados obtidos pelos experimentos realizados indicam que os métodos desenvolvidos foram capazes de fornecer conjuntos de dados mais representativos:

- Mecanismo interativo de seleção de variáveis;
- Mecanismo interativo de combinação de variáveis;
- Mecanismo interativo de construção de variáveis;
- Metáfora visual para transformações de dados;

Individualmente, esses métodos apresentam valor, mas a maior contribuição deste trabalho vai além disso. O grande feito foi integrar esses métodos em um arcabouço unificado, permitindo que os usuários lidem prontamente com as principais atividades da transformação de dados, proporcionando uma análise mais efetiva em tarefas subjacentes.

1.3 Organização da dissertação

O restante desta dissertação está organizado da seguinte maneira. No Capítulo 2, discute-se sobre as características e limitações dos trabalhos que apresentam objetivos similares aos aqui estabelecidos. No Capítulo 3, apresenta-se os métodos e técnicas da literatura que serviram como base para o desenvolvimento deste trabalho de mestrado. Os métodos de transformação interativa desenvolvidos são apresentados no Capítulo 4. No Capítulo 5, apresenta-se o arcabouço e os resultados dos experimentos realizados. Por fim, no Capítulo 6, conclui-se esta dissertação com uma discussão das contribuições deste trabalho e dos trabalhos futuros que consideramos relevantes para a continuidade desta pesquisa.

Trabalhos Relacionados

Os trabalhos que mais se relacionam ao o aqui proposto são aqueles que buscam transformar conjuntos de dados para torná-los mais representativos para execução de uma determinada tarefa. Esses trabalhos dividem-se basicamente em dois grupos: métodos de redução de dimensionalidade e métodos de construção interativa de atributos.

A seguir, na Seção 2.1, apresenta-se uma discussão sobre os métodos de redução de dimensionalidade, com um enfoque especial para métodos interativos. Na Seção 2.2, apresenta-se um levantamento sobre pesquisas em construção interativa de atributos, um tema que não conta com uma literatura tão vasta quanto à dos métodos de redução, mas vem ganhando popularidade nos últimos anos.

2.1 Redução de Dimensionalidade

O problema de reduzir a dimensionalidade de conjuntos de dados pode ser descrito da seguinte forma: dado um conjunto de dados representado por uma matriz $\mathbf{X}$ composta por n vetores $\mathbf{x}_i$ ($i \in \{1, 2, \dots, n\}$) m -dimensionais, deseja-se encontrar uma transformação $t : \mathbf{X} \rightarrow \mathbf{Y}$, em que $\mathbf{Y}$ é uma matriz composta por n vetores $\mathbf{y}_i$ ($i \in \{1, 2, \dots, n\}$) de dimensionalidade p ($p < m$). Normalmente $p \ll m$ e, idealmente, p equivale à

*dimensionalidade intrínseca dos dados*¹, fazendo com que t mantenha em $\mathbf{Y}$ o máximo das propriedades de $\mathbf{X}$ quanto for possível.

Um dos principais objetivos da redução de dimensionalidade é amenizar os efeitos da *maldição da dimensionalidade*² e com isso fazer com que os métodos que operam sobre os dados tenham uma melhor eficiência e um menor custo computacional (Maaten et al., 2009). König (2000), por exemplo, apresenta melhorias na precisão de sistemas de classificação e no desempenho de sistemas de reconhecimento automático ao preceder os procedimentos com um processo de redução de dimensionalidade. Até mesmo outras melhorias não tão diretas podem ser alcançadas por meio do uso de técnicas de redução. Trata-se do caso do mesmo trabalho apresentado por König (2000), onde métodos de redução de dimensionalidade são utilizados para reduzir a complexidade de projetos de circuitos integrados, resultando em uma redução na área e no consumo de energia dos circuitos.

Uma outra utilidade dos métodos de redução de dimensionalidade é viabilizar a construção de representações visuais de dados multidimensionais, permitindo que sejam mapeados em um espaço bidimensional (tela computador). Representações visuais têm sido fundamentais para análises exploratórias de dados, principalmente em investigações iniciais, onde não se conhece as propriedades dos dados (Kaski et al., 2011).

A literatura em redução de dimensionalidade é extensa e os métodos desenvolvidos apresentam grande diversidade em relação a aspectos matemáticos e computacionais. Para uma melhor organização, esta seção foi dividida em duas subseções. Na Subseção 2.1.1 busca-se descrever sucintamente os métodos automáticos e apresentar suas limitações, evidenciando que a falta da participação do usuário no processo faz com que muitas vezes os resultados obtidos não sejam facilmente compreendidos. Já a Subseção 2.1.2, apresenta os métodos que permitem ao usuário participar no processo de redução de dimensionalidade por meio de interações com representações visuais.

2.1.1 Métodos Automáticos

A redução de dimensionalidade automática pode ser realizada seguindo duas abordagens (Pudil et al., 1998). A primeira transforma os atributos de entrada em um novo conjunto de dimensões que busca conservar propriedades ou relacionamentos do conjunto original. Por extrair um novo conjunto de atributos a partir dos originais, esta abordagem

---

¹A dimensionalidade intrínseca dos dados é o conjunto mínimo de variáveis necessárias para descrever as propriedades dos dados (Fukunaga, 1990).

²A maldição da dimensionalidade foi um termo introduzido por Bellman (1961) para se referir aos problemas que surgem ao lidar com conjuntos de dados com um elevado número de dimensões (centenas ou milhares). Sua maior implicação é a falta de separabilidade entre elementos que se encontram em um espaço de alta dimensão (Kouiroukidis et al., 2011).

recebe o nome de extração de características (*feature extraction*). Já a segunda abordagem busca selecionar quais dos atributos do conjunto de dados são realmente relevantes para as análises. Como os dados não são modificados, esta segunda abordagem é chamada de seleção de características (*feature selection*). Ambas abordagens serão discutidas a seguir.

2.1.1.1 Extração de Características

Como apresentado por Maaten et al. (2009), existe uma grande variedade de métodos de extração de características. Não é intuito desta subseção detalhar cada uma dessas técnicas e levantar suas limitações particulares, mas sim ilustrar a limitação comum que a maioria apresenta, isto é, retornar resultados pouco intuitivos para o usuário e impedi-lo de interagir com os dados. Para este fim, o conjunto de dados fictício apresentado na Tabela 2.1 será utilizado como exemplo.

Tabela 2.1: Conjunto de dados fictício. Os valores foram estabelecidos arbitrariamente e não apresentam necessariamente alguma relação com índices oficiais.

	<i>Padrão</i>	<i>Clima</i>	<i>Gastro- nomia</i>	<i>Segu- rança</i>	<i>Infra- estrutura</i>
Alemanha	8	3	2	8	9
Brasil	5	8	7	3	3
Croácia	5	6	6	6	6
Espanha	7	9	9	5	8
França	8	4	7	7	8
Itália	7	8	9	5	7
Marrocos	4	7	8	2	2
México	2	5	5	2	3
Nigéria	2	4	4	2	2
Peru	5	6	6	3	4
Rússia	6	2	2	3	6
Turquia	5	8	9	3	3

Um dos primeiros métodos desenvolvidos para a redução de dimensionalidade trata-se da análise de componentes principais (PCA) (Pearson, 1901), sendo que até hoje é um dos mais utilizados (Jolliffe, 2002). Neste método, as dimensões extraídas, ou componentes, são combinações lineares das dimensões originais, onde cada uma busca capturar características distintas das outras.

A Tabela 2.2 apresenta o resultado obtido pela técnica PCA para os dados da Tabela 2.1. Nota-se a dificuldade em interpretar o resultado obtido, pois aparentemente os componentes extraídos não apresentam nenhuma relação com os dados originais. Com base somente nesse resultado não é possível, por exemplo, compreender quais atributos foram combinados para a obtenção de tais componentes.

Tabela 2.2: Resultado obtido pela técnica PCA para os dados da Tabela 2.1. As novas dimensões são combinações lineares das dimensões originais.

	<i>Comp.1</i>	<i>Comp.2</i>	<i>Comp.3</i>	<i>Comp.4</i>	<i>Comp.5</i>
Alemanha	-8.3095	0.5446	1.9864	-0.3819	-0.8144
Brasil	3.4480	-0.9062	0.2499	0.2469	-0.6788
Croácia	-1.7908	-0.8416	0.0145	-1.5911	0.1811
Espanha	0.4513	-5.7962	1.2968	0.3819	0.9095
França	-6.6956	-2.0063	-2.7319	-0.2799	-0.0306
Itália	0.0796	-4.7684	0.0922	0.3013	0.3266
Marrocos	5.1923	-0.0863	-0.4045	0.2746	-0.5028
México	2.7158	3.3806	0.0858	-0.6728	1.1078
Nigéria	2.5911	4.8643	0.1585	-0.6316	0.2394
Peru	0.9130	0.6241	-0.4731	0.4676	0.0180
Rússia	-4.5417	4.3595	-0.0156	1.7600	0.4539
Turquia	4.5991	-2.2306	-0.3223	0.1513	-0.6417

Em PCA, os componentes são gerados em uma ordem decrescente de importância, de modo que o primeiro captura mais informação dos dados que o segundo e assim por diante. Deste modo, a redução de dimensionalidade em si ocorre ao se manter os k primeiros componentes gerados pelo método. Em tarefas de visualização, por exemplo, é comum escolher $k = 2$ para manter somente os dois primeiros componentes e então criar uma representação bidimensional dos elementos. Uma possibilidade menos arbitrária para definir k é analisar a parcela da variância dos dados que cada componente captura. Tal análise pode ser realizada com o auxílio de gráficos, como o apresentado na Figura 2.1. Com base neste gráfico, nota-se que os dois primeiros componentes capturam grande parte da variância (aproximadamente 92%). No entanto, mesmo com o auxílio de tais recursos o usuário acaba dependendo de medidas estatísticas para medir valores que muitas vezes podem ser subjetivos.

Quando existem relações não lineares entre os atributos, PCA não é capaz de capturá-las. Em situações como esta, métodos não lineares como *Multidimensional Scaling* (MDS) (Cox et al., 2008) e *Self Organizing Maps* (SOM) (Kohonen, 1990) podem ser utilizados para uma maior eficácia. Porém, independentemente de quais técnicas se sobressaem sobre as outras, uma limitação que os métodos de extração compartilham entre si é a dificuldade em se compreender o resultado obtido, ou seja, o espaço dimensional gerado tem pouco significado para o usuário.

Como pode ser observado na Tabela 2.2, não existe uma correspondência direta entre as dimensões extraídas e as originais. Logo, o usuário pode ter dificuldades em associar quais características dos dados foram responsáveis por aquele resultado. A seguir apresenta-se

Figura 2.1: Variância dos dados capturada por cada componente para os dados da Tabela 2.1. Os dois primeiros componentes capturam cerca de 92% da variância.

uma abordagem alternativa que leva em consideração tal problema e busca manter uma correspondência entre o espaço dimensional reduzido e o original.

2.1.1.2 Seleção de Características

O objetivo dos métodos de seleção de características é encontrar o subconjunto dos atributos de entrada mais adequado para a aplicação em estudo. Assim, busca-se identificar e eliminar atributos redundantes (Kohavi et al., 1997) ou que não apresentem relação com o fenômeno investigado (Nilsson et al., 2007). Por exemplo, em tarefas de classificação supervisionada, pode-se determinar a importância de um atributo ao se avaliar sua correlação em respeito ao atributo classe, de modo que atributos com altos valores de correlação são de maior importância. Os métodos de seleção dividem-se basicamente em *filtros*, *wrappers* e *métodos embutidos* (Guyon et al., 2003).

Filtros utilizam um atributo alvo como referência e determinam, a partir de alguma medida de correlação, quanto cada atributo se relaciona com esta referência. A filtragem

é realizada por meio da eliminação de atributos que apresentam relação menor do que um valor fixado. Uma das desvantagens de filtros é que pelo fato de considerarem somente relações par-a-par, não são capazes de detectar dependências indiretas entre os atributos.

O funcionamento de *wrappers* e métodos embutidos consiste em realizar uma busca sobre subconjuntos candidatos e tomar como resultado o subconjunto que resulta na melhor precisão de um algoritmo de predição. O caso completo trata-se da avaliação de 2^m subconjuntos, onde m corresponde ao número de atributos do conjunto de entrada. Tal situação equivale a um problema *np*-completo (Amaldi et al., 1998), consequentemente para grandes conjuntos de dados a solução ótima não pode ser obtida em tempo viável, exigindo assim a adoção de alguma heurística. São justamente essas heurísticas que definem os diferentes métodos que podem ser utilizados. De um modo geral, a distinção entre *wrappers* e métodos embutidos vem de que os primeiros enxergam o método de predição como uma “caixa-preta”, se interessando somente pelo resultado obtido e permitindo que diferentes preditores sejam aplicados sem a necessidade de modificar o método de seleção. Já os métodos embutidos são incorporados às etapas de treinamento dos preditores, sendo assim específicos para cada situação.

Em comparação aos métodos de extração de características, os métodos de seleção apresentam a vantagem de que o resultado obtido é mais intuitivo ao usuário, pois se trata de um subconjunto dos atributos de entrada. Assim, se o usuário tem certo conhecimento sobre o conjunto de entrada, então será capaz de compreender os resultados obtidos. No entanto, eles compartilham da mesma natureza caixa-preta dos métodos de extração. Isto é, impedem qualquer tipo de interação durante o processo de redução, impedindo que o usuário contribua com seu conhecimento sobre o domínio e compreenda quais características dos seus dados foram responsáveis por aquele resultado. A seguir apresenta-se um levantamento dos trabalhos que inserem o usuário no processo de redução para contornar essa limitação.

2.1.2 Métodos Iterativos

Métodos visuais que permitem a interação do usuário ganharam popularidade nos últimos anos (Milojević et al., 2012). Grande parte deste sucesso deve-se ao uso efetivo da capacidade preemptiva da visão humana. Foi demonstrado que quando os dados são representados por alguma forma gráfica, o ser humano é capaz de detectar e reconhecer padrões facilmente e com rapidez (Healey et al., 1995), mesmo em grandes conjuntos de dados (Fodor, 2002). Mas utilizar a capacidade preemptiva da visão humana não é a única vantagem dos métodos visuais. Permitir que o usuário participe ativamente nos processos e na geração dos resultados é um dos grandes benefícios desses métodos, pois proporciona uma maior confiança e controle sobre o processo.

A seguir serão apresentados os trabalhos desta vertente que buscam executar redução de dimensionalidade de forma interativa. São trabalhos que não somente fazem uso da capacidade perceptiva humana, mas que também permitem que o usuário participe ativamente na geração dos resultados com o seu conhecimento sobre o domínio.

2.1.2.1 Matrizes de Correlação

Uma das maneiras mais utilizadas para inspecionar relações entre dimensões são as matrizes de correlação (Friendly, 2002). A Figura 2.2 apresenta um exemplo deste tipo de representação para um conjunto de dados de *baseball* (Friendly, 2002). As cores azul e vermelho indicam uma correlação positiva e negativa entre os pares de dimensões, respectivamente. A intensidade da cor indica a magnitude da correlação, sendo que branco representa dimensões não correlacionadas. Com base em uma investigação visual sobre essa figura é possível levantar algumas hipóteses sobre os dados. Observa-se, por exemplo, uma relação direta entre os anos de carreira do jogador (*Years*) e o seu salário (*logSal*), ao mesmo tempo a experiência tem uma relação inversa com o número de erros.

Figura 2.2: Exemplo de matriz de correlação. A cor azul indica correlação positiva entre as variáveis, enquanto a vermelha correlação negativa. A intensidade da cor é proporcional à magnitude da correlação.

Esse tipo de representação é útil para ter uma visão geral das relações entre pares de dimensões. No entanto, para análises mais detalhadas, ou que exijam uma comparação entre mais do que simplesmente pares de elementos, não é uma representação adequada.

Devido à sua simplicidade, matrizes de correlação têm sido adotadas por diversos métodos visuais que viabilizam a investigação de atributos de conjuntos de dados. A ferramenta desenvolvida por Guo (2003), por exemplo, utiliza matrizes de correlação para apresentar as relações entre os atributos e utiliza um método de agrupamento para ordenar as colunas da matriz de modo a destacar grupos de dimensões similares.

Figura 2.3: Visão geral da ferramenta desenvolvida por Guo (2003). Imagem extraída de Guo (2003)).

A Figura 2.3 apresenta uma visão geral da ferramenta proposta por Guo (2003). Na parte superior da figura encontra-se a matriz de correlação, onde o usuário seleciona

subconjuntos de atributos de interesse por meio da seleção dos elementos correspondentes na diagonal da matriz. Nota-se que, devido ao método de ordenação das colunas, os grupos de dimensões similares podem ser identificados mais facilmente. As outras visualizações são coordenadas com as seleções do usuário e então análises mais detalhadas podem ser realizadas para uma melhor compreensão das estruturas presentes naquele subconjunto de atributos.

O objetivo das matrizes de correlação não é propriamente reduzir a dimensionalidade do conjunto de dados, mas sim ajudar o usuário a encontrar subconjuntos de atributos com características de interesse. Outros trabalhos (Friendly, 2002; MacEachren et al., 2003; Seo et al., 2004; May et al., 2011a; Johansson et al., 2009; Ingram et al., 2010; May et al., 2011b) também adotam matrizes de correlação para atingir esse mesmo objetivo. Eles diferem na maneira como é construída a matriz de correlação e de quais recursos são disponibilizados para o usuário interagir sobre os subconjuntos de dimensões. Um problema geral desses trabalhos é que certas análises podem exigir demasiado esforço do usuário devido à necessidade de se explorar individualmente cada dimensão ou avaliar par-a-par as relações entre atributos. Com a ocorrência de dependências não lineares este problema torna-se ainda maior e o usuário pode se perder em suas análises e não extrair novos conhecimentos dos resultados.

A seguir apresenta-se alternativas às matrizes de correlação para se apresentar medidas de correlação. Essas abordagens fornecem mecanismos mais diretos para se reduzir a dimensionalidade dos conjuntos de dados.

2.1.2.2 Hierarquias de Dimensões

Em busca de construir espaços de baixa dimensionalidade mais intuitivamente do que pelo uso de métodos automáticos, Yang et al. (2003) desenvolveram o método de redução de dimensionalidade chamado VHDR (*Visual Hierarchical Dimensions Reduction*). O funcionamento deste método é ilustrado pela Figura 2.4a. Inicialmente (1), constrói-se uma organização hierárquica dos atributos com base na similaridade entre as dimensões. Em seguida (2), o usuário define os níveis da hierarquia que devem ser considerados pela última etapa do processo. Finalmente (3), o usuário, por meio de um método automático ou de seu conhecimento sobre os dados, escolhe dimensões representativas para os níveis definidos, reduzindo assim a dimensionalidade dos dados.

O processo de construção da hierarquia é muito semelhante aos algoritmos de agrupamento hierárquico. A distinção é que agrupa-se atributos semelhantes, ao invés de itens. Deste modo, qualquer método de agrupamento pode ser aplicado, exigindo-se apenas que o agrupamento resulte em uma estrutura hierárquica, no caso uma árvore, onde cada dimensão seja representada por um nó folha da árvore.

Figura 2.4: Em (a) ilustra-se o funcionamento do VHDR. Em (b) exemplifica-se a representação gráfica adotada pelos autores do VHDR. Ambas imagens extraídas de (Yang et al., 2003).

A representação gráfica utilizada no VHDR é a *InterRing* (Yang et al., 2002) e pode ser observada na Figura 2.4b. O nó raiz da árvore é representado pelo círculo mais interno e os nós folhas pelos elementos posicionados na borda. As cores são utilizadas para destacar grupos de dimensões com características em comum.

Os autores do VHDR desenvolveram uma extensão chamada DOSFA (*Dimension Ordering Spacing and Filtering Approach*) (Wang et al., 2003) que apresenta outras abordagens para investigar os atributos de um conjunto de dados. Mais especificamente, eles propõem ferramentas para ordenação, espaçamento e filtragem de atributos. As duas primeiras, ordenação e espaçamento, não estão diretamente relacionadas com redução de dimensionalidade. Já a filtragem de atributos é análoga aos métodos de seleção de características. Este mecanismo consiste em remover dimensões pouco representativas ou redundantes, de modo que se certas dimensões apresentam alta similaridade entre si, então apenas uma delas é mantida, ou se certas dimensões apresentam pouca relevância, então são descartadas. A grande complexidade do processo de filtragem está no modo como se define a redundância e a importância entre as dimensões. Um método semelhante para filtragem de atributos irrelevantes foi proposto por Artero et al. (2006).

Tanto o VHDR quanto o DOSFA, tratam todo o conjunto de dados de maneira uniforme, ou seja, não consideram que possa existir subconjuntos nos dados com diferentes características que devem ser analisados separadamente (May et al., 2011b). Uma maneira

de contornar este problema seria apresentar os itens simultaneamente com a representação das dimensões, assim o usuário poderia detectar grupos não somente nas dimensões mas também nos itens. Os trabalhos discutidos a seguir utilizam tal representação para reduzir a dimensionalidade dos conjuntos de dados. Esses trabalhos servem como inspiração para a construção das visualizações que utilizaremos na proposta deste projeto.

2.1.2.3 Mapeamento de Elementos no Plano

Abordando justamente o problema de se apresentar itens simultaneamente com as dimensões de um conjunto de dados, Yang et al. (2004) desenvolveram a ferramenta VaR (*Value and Relation*). A abordagem une os conceitos de MDS e glifos para representar as dependências entre as dimensões de uma base de dados.

Figura 2.5: Em (a) exemplifica-se a ferramenta VaR. Em (b) apresenta-se um exemplo da representação alternativa proposta como extensão da ferramenta VaR. Imagens extraídas de (Yang et al., 2004) e (Yang et al., 2007), respectivamente.

Na representação visual da ferramenta VaR, cada dimensão é representada por um glifo e seus posicionamentos refletem a similaridade entre as dimensões, de modo que glifos que se encontram próximos indicam atributos que apresentam alguma relação entre si. Como mostra a Figura 2.5a, de acordo com o posicionamento dos glifos no plano o usuário pode compreender como as dimensões se relacionam entre si. O usuário é capaz de construir espaços dimensionais reduzidos que conservam certas características dos dados por meio de seleções manuais sobre os dados ou pelo uso de um método automático. Este método automático parte de uma dimensão de referência e de um limiar definido pelo usuário e retorna as dimensões mais semelhantes à esta referência.

O procedimento para o mapeamento das dimensões tem início com a construção de uma matriz de distâncias que é responsável por capturar os relacionamentos entre pares de dimensões do conjunto de dados. Sobre esta matriz de distâncias aplica-se uma técnica de MDS para mapear cada dimensão em uma posição do espaço bidimensional. Finalmente, cria-se um glifo orientado a pixels para cada dimensão que é utilizado para representar as dimensões no plano.

Observando a Figura 2.5a nota-se que o uso de glifos faz com que ocorram sobreposições, pois cada glifo requer um espaço relativamente grande para que seja analisado adequadamente. As sobreposições dificultam as análises de regiões de interesse e podem fazer com que o usuário alcance conclusões inválidas, devido a oclusão de algum elemento importante.

Para tratar o problema de sobreposição de elementos, Yang et al. (2007) desenvolveram a extensão ilustrada na Figura 2.5b, onde apresentaram alternativas para o mapeamento dos glifos no plano. Porém, a abordagem adotada não considera disparidades nas distâncias entre elementos vizinhos, ou seja, independentemente da similaridade entre dois glifos adjacentes, eles sempre serão representados pela mesma distância no plano. Assim, o resultado obtido pela versão original transmite melhor os relacionamentos entre as dimensões do que a abordagem proposta na extensão.

Apesar de a ferramenta VaR apresentar informações sobre itens e dimensões simultaneamente, não é permitido ao usuário interagir com os itens. Consequentemente, esta abordagem sofre das mesmas limitações das ferramentas apresentadas anteriormente, ou seja, não é capaz de lidar com características locais em subconjuntos dos dados. Um outro aspecto importante que os próprios autores mencionam em relação ao uso de glifos é que os usuários têm dificuldade em comparar glifos que se encontram afastados.

Figura 2.6: Em (a) ilustra-se o conceito principal do trabalho de Turkay et al. (2011), *Brushing Dimensions*. Em (b) à esquerda (1) apresenta-se a representação visual dos itens e na direita (2) a das dimensões. Imagens extraídas de (Turkay et al., 2011)

O trabalho proposto por Turkay et al. (2011), *Brushing Dimensions* (BD), cobre essa limitação da ferramenta VaR, pois permite aos usuários interagir tanto com as dimensões dos conjunto de dados quanto com os itens. Como pode ser observado na Figura 2.6 o usuário pode realizar seleções em ambas direções. Semelhantemente à ferramenta VaR, as representações visuais do BD são baseadas em mapeamentos de elementos no plano. As representações dos itens são construídas com base em métodos automáticos, como PCA, e as das dimensões são criadas por *scatterplots* usando medidas estatísticas, como média e variância. Este modo de posicionamento das dimensões é uma das limitações da ferramenta, pois ao desconsiderar medidas par-a-par, como correlação, a visualização não apresentará dependências entre os atributos. O principal mecanismo de interação da ferramenta BD é a seleção que se reflete em outras visões e permite que se visualize, por exemplo, variações na importância de um atributo em diferentes subconjuntos dos dados. Uma das limitações de ambos os métodos, VaR e BD, é não permitir que o usuário construa novas dimensões com base nas originais ou com base em seu conhecimento.

Uma questão inerente de se mapear elementos de um espaço de alta dimensionalidade em um plano, sejam os elementos itens ou dimensões, é que não há garantias de que o mapeamento seja válido. Em casos onde a dimensionalidade intrínseca dos dados for maior do que a do espaço alvo, então poderá haver sobreposição de elementos sem necessariamente significar que os elementos sobrepostos sejam realmente semelhantes. Ambos VaR e BD não atentam para esta questão, mas Ingram et al. (2010) desenvolveram a ferramenta *DimStiller* buscando construir mapeamentos de dados multidimensionais levando em consideração este problema.

A Figura 2.7 ilustra a ferramenta *DimStiller*. Pelo gráfico de barras (janela canto inferior esquerdo) o usuário reconhece a dimensionalidade intrínseca dos dados, proporcionando um melhor entendimento dos resultados que será possível obter por meio da redução de dimensionalidade. O mapeamento resultante da redução é apresentado em um gráfico dos dois componentes principais (janela canto inferior direito). De acordo com esta visualização, não existem estruturas de interesse nos dados. No entanto, ao observar mapeamentos com outros componentes da redução (janela canto superior direito), o usuário pode identificar padrões nos dados.

Outro aspecto importante da redução de dimensionalidade, que muitas vezes não é levado em consideração, é que dependendo do método adotado, diferentes características dos dados podem ser mantidas e outras perdidas. Este problema é abordado no trabalho de Johansson et al. (2009), onde por meio de gráficos de perda de informação para diferentes medidas, o usuário pode entender quais características dos seus dados são mantidas e perdidas ao longo do processo de redução.

O trabalho de Johansson et al. (2009) e (Ingram et al., 2010) fornecem meios para o usuário avaliar a incerteza dos resultados. Apresentam aos usuários os fatores que podem

Figura 2.7: Abordagem proposta pela ferramenta *DimStiller* para criar mapeamentos de dados multidimensionais interativamente. Imagem extraída de (Ingram et al., 2010)

resultar em interpretações ambíguas dos resultados e na perda de possíveis informações de interesse. Tal característica não é presente em muitas das ferramentas de visualização atuais, mas tem se tornado cada vez mais uma exigência (Dill et al., 2012).

As ferramentas de redução de dimensionalidade não são restritas a totalmente automáticas ou integralmente interativas. Abordagens mistas podem ser adotadas, como é o caso dos trabalhos discutidos a seguir.

2.1.2.4 Visualização de Métodos Automáticos

Existem métodos que não fazem uso de representações visuais para realizar a redução de dimensionalidade em si, mas sim para tornar os métodos automáticos mais compreensíveis. Eles buscam incluir a participação do usuário nesse processo para tornar esses métodos “caixas-pretas” mais intuitivos.

No contexto de extração de características, a ferramenta iPCA (Jeong et al., 2009) provê meios para o usuário manipular a importância de cada atributo na técnica PCA e, assim, é capaz de ajudar no entendimento das transformações realizadas sobre os dados. O funcionamento dessa ferramenta é ilustrado na Figura 2.8. Em (A) apresenta-se o mapeamento dos elementos no plano utilizando dois componentes principais selecionados

Figura 2.8: Ilustração da ferramenta iPCA. Imagem extraída de (Jeong et al., 2009)

pelo usuário. (B) e (C) referem-se a visualizações dos dados originais e transformados, respectivamente. Em (D) é apresentada uma matriz de correlação. O principal mecanismo de interação é indicado por (E), o qual permite ao usuário definir a contribuição de cada atributo no resultado final.

Similarmente, Williams et al. (2004) permitem que o usuário guie o processo de redução de dimensionalidade a partir de métodos MDS ao escolher regiões de interesse para se concentrar os esforços computacionais. Neste mesmo sentido, Schreck et al. (2008) desenvolveram uma ferramenta que permite ao usuário monitorar visualmente os recursos computacionais utilizados pelo método SOM e definir interativamente os parâmetros para sua execução. Recentemente, Pagliosa et al. (2015) desenvolveram o chamado *Projection Inspector*, que consiste em um método interativo que permite combinar diversas técnicas de redução de dimensionalidade para a obtenção de uma representação visual dos dados.

Visualizações também podem ser utilizadas para tornar métodos de seleção de características mais intuitivos. O trabalho de Dy et al. (2000), por exemplo, possibilita que o usuário adicione ou remova atributos interativamente ao longo do processo de redução. Brandoli et al. (2010) propôs uma outra abordagem para o mesmo fim, na qual utiliza visualizações para auxiliar o usuário a definir os parâmetros dos métodos de seleção de características.

Encontra-se na literatura alguns trabalhos que buscam tornar o processo de redução de dimensionalidade mais intuitivo no contexto de classificação de dados Zhang et al. (2006), Choo et al. (2010) e Paiva et al. (2012). Nesses trabalhos, o usuário interage simultaneamente com representações dos resultados do processo de redução de dimensionalidade e também do modelo de classificação. Com base nessa interação conjunta, espera-se obter classificadores mais representativos para o problema.

Recentemente, Krause et al. (2014) desenvolveram uma ferramenta para a seleção interativa de atributos chamada INFUSE. Nesta ferramenta, permite-se que o usuário analise o potencial preditivo de cada variável para a construção de modelos lineares. Como mostra a Figura 2.9, na esquerda há uma visão geral de todas as variáveis de acordo com suas respectivas capacidades preditivas. No canto superior direito, apresenta-se uma lista das variáveis na qual o usuário pode remover variáveis. No canto direito inferior, apresenta-se a qualidade dos modelos construídos.

Figura 2.9: Ilustração da ferramenta INFUSE. Imagem extraída de (Krause et al., 2014)

Um problema das ferramentas que criam visualizações de métodos automáticos é a necessidade do usuário ter um certo conhecimento sobre o método utilizado para a construção da visualização. Por exemplo, o usuário pode não fazer um uso efetivo da ferramenta *iPCA* se não compreender o significado de um componente principal. Para pesquisadores da área pode ser até pressuposto que o usuário tenha este tipo de conhecimento, no entanto, se o objetivo for criar uma ferramenta para um uso mais abrangente, tal suposição pode restringir seu uso.

2.2 Construção Interativa de Atributos

Os métodos de redução de dimensionalidade realizam transformações sobre os dados, porém, mesmo as ferramentas interativas de redução não permitem que o usuário modifique os atributos livremente. A construção interativa de atributos significa mais do que

permitir que o usuário guie as transformações sobre os dados, trata-se de permitir que o usuário agregue seu conhecimento sobre os dados de forma mais incisiva.

Esse tipo de abordagem ainda não conta com uma literatura tão vasta quanto a dos métodos de redução de dimensionalidade. Sendo que a contribuição mais relevante é a ferramenta proposta por Mamani et al. (2013) que possibilita ao usuário modificar os atributos de um conjunto de dados com base na manipulação sobre amostras dos itens. Os autores fazem uso de mapeamento de elementos no plano para permitir interações intuitivas sobre os dados.

A Figura 2.10 apresenta o funcionamento dessa técnica. Inicialmente, dentre os n vetores pertencentes ao espaço m -dimensional, $\mathcal{X} = \{x_1, \dots, x_n\}$, escolhe-se uma amostra $\mathcal{X}_s \subset \mathcal{X}$, $\mathcal{X}_s = \{x_{s1}, \dots, x_{sk}\}$ com k elementos. De modo a obter uma representação visual da amostra, mapeia-se os elementos para um espaço bidimensional, obtendo-se $\mathcal{Y}_s = \{y_{s1}, \dots, y_{sk}\}$. Então, possibilita-se que o usuário manipule o posicionamento dessas amostras, buscando agrupar elementos que considera similares. Em seguida, utiliza-se o novo posicionamento das amostras, $\tilde{\mathcal{Y}}_s$, para redefinir o espaço inicial:

$$\tilde{x}_{sj} = \tilde{x}_{sj} + \frac{\vec{v}}{\|\vec{v}\|} (d_2(\tilde{y}_{si}, \tilde{y}_{sj}) - (d_m(\tilde{x}_{si}, \tilde{x}_{sj}))) \quad (2.1)$$

sendo v o vetor de x_{si} para x_{sj} e d_2 e d_m a distância dos vetores nos espaços de dimensionalidade 2 e m , respectivamente. Assim, as manipulações realizadas sobre a amostra são refletidas para o conjunto de dados original, fazendo com que este reflita o conhecimento do usuário em relação ao critério de similaridade entre os elementos. Este processo pode ser repetido até que se atinja o resultado esperado. Observa-se que para o exemplo a transformação do espaço foi bem sucedida, pois foram reveladas estruturas que não eram identificáveis no mapeamento original.

O trabalho de Liu et al. (2012), *Dis-function*, é similar ao de Mamani et al. (2013). No entanto, a obtenção de um novo espaço multidimensional não vem da direta transformação do espaço original, mas da elaboração interativa de uma métrica de distância. A interação do usuário também se baseia em uma representação bidimensional dos dados, mas ao invés de arrastar os elementos, o usuário estabelece relações de conectividade entre elementos que julga similares. Com base nessas interações refina-se a matriz A de $D_A = \sqrt{(x_i - x_j)^T A (x_i - x_j)}$ obtendo-se uma nova métrica de distância. Esta nova métrica, então pode ser aplicada sobre os dados originais para a obtenção de novas dimensões.

Os trabalhos Mamani et al. (2013) e Liu et al. (2012) não constróem propriamente novas dimensões, mas modificam as existentes para agregar o conhecimento do usuário. Ao se transformar os dados originais o usuário pode ter dificuldade em interpretar o resultado obtido, além de se ter pouco controle o que está de fato ocorrendo com os

Figura 2.10: Ferramenta desenvolvida por Mamani et al. (2013).

fatos, semelhantemente ao que acontece com os métodos automáticos de extração de características. Na abordagem aqui proposta pretendemos mapear o conhecimento do usuário em novas dimensões e conservar as dimensões originais.

2.3 Considerações Finais

Neste capítulo foram apresentados os trabalhos que modificam os conjuntos de dados de modo a torná-los mais representativos para a compreensão do fenômeno em estudo. Discutiu-se que a natureza “caixa-preta” dos métodos automáticos faz com que muitas vezes os resultados não sejam facilmente compreendidos pelos usuários. Foi discutido também que apesar dos métodos interativos se mostrarem uma interessante alternativa aos métodos automáticos, eles ainda apresentam limitações. As principais características dos métodos interativos foram elencadas na Tabela 2.3.

Nota-se por essa tabela que nenhum dos trabalhos consegue unir em único ambiente os três principais mecanismos de interação para a transformação dos dados, seleção, combinação e construção de variáveis. Essa é uma das maiores limitações do estado da arte, pois um único mecanismo não é capaz de operar otimamente para todas possíveis aplicações.

Observa-se também que nem todas as ferramentas conseguem apresentar itens e dimensões simultaneamente. Dentre as que conseguem, uma parcela ainda menor permite ao usuário interagir sobre ambas representações. Esse tipo de interação é importante para permitir que o usuário realize avaliações locais nos dados. Este é um recurso fundamental, pois dificilmente o conjunto de dados apresentará um comportamento uniforme globalmente, sendo mais provável que existam subconjuntos com diferentes características que devem ser avaliadas localmente.

Apesar da avaliação de incerteza ser um aspecto importante de ferramentas visuais, nota-se que poucos trabalhos atentam para essa questão. Além disso, algumas ferramentas se baseiam em interfaces demasiadamente complexas, as quais exigem do usuário um certo período de treinamento para um uso efetivo. Tendo em vista que o objetivo das ferramentas visuais é tornar as análises mais intuitivas, qualquer tipo de obstáculo, como a necessidade de treinamento do usuário, pode ser desfavorável ao se comparar com os métodos automáticos.

Tabela 2.3: Características de interesse dos principais trabalhos estudados.

Ferramenta	Mecanismos de Interação			Representação das Dimensões	Representação dos Itens	Interação sobre Itens	Avalia Incerteza	Complexidade de uso
	Seleção	Extração	Construção					
Guo (2003)	Não	Não	Não	Sim	Não	Não	Não	Baixa
VHDR (2003)	Sim	Não	Não	Sim	Não	Não	Não	Baixa
INFUSE (2014)	Sim	Não	Não	Sim	Não	Não	Não	Baixa
VaR (2007)	Sim	Não	Não	Sim	Sim	Não	Não	Baixa
BD (2009)	Sim	Não	Não	Sim	Sim	Sim	Não	Alta
DimStiller (2010)	Sim	Sim	Não	Sim	Sim	Não	Sim	Alta
Johansson et al. (2009)	Sim	Sim	Não	Sim	Sim	Não	Sim	Baixa
iPCA (2009)	Não	Sim	Não	Sim	Sim	Sim	Não	Alta
Mamani et al. (2013)	Não	Não	Sim	Não	Sim	Sim	Não	Baixa

Conceitos Teóricos

NESTE capítulo, apresenta-se os trabalhos que serviram como base para o desenvolvimento deste trabalho de mestrado. Eles podem ser divididos em dois grupos: métodos de projeção multidimensionais e *biplots*. De modo geral, os métodos de projeção multidimensional foram utilizados para apresentar visualmente a relação de similaridade entre os elementos e para permitir análises locais dos conjuntos de dados. Os *biplots* serviram como base para o desenvolvimento dos mecanismos interativos de transformação. A seguir, discute-se com mais detalhes sobre esses três conceitos.

3.1 Projeções Multidimensionais

As técnicas de projeção multidimensional permitem mapear elementos pertencentes a um espaço de alta dimensão em espaços p -dimensionais, com $p = \{1, 2, 3\}$, buscando preservar as relações de distâncias e de similaridade entre os dados (Paulovich et al., 2008). Mais formalmente, uma técnica de projeção multidimensional pode ser definida como (Tejada et al., 2003a):

Seja X um conjunto de objetos R^m com $\delta : R^m \times R^m \rightarrow R$ um critério de proximidade entre dois objetos em R^m , e Y um conjunto de objetos em R^p para $p = \{1, 2, 3\}$ e $d : R^p \times R^p \rightarrow R$ um critério de proximidade em R^p . Uma técnica de projeção multidimensional pode ser descrita como uma função

$f : X \rightarrow Y$ que visa tornar $|\delta(x_i, x_j) - d(f(x_i), f(x_j))|$ o mais próximo possível de zero, $\forall x_i, x_j \in X$.

Idealmente, a proximidade dos pontos indica a semelhança entre os objetos que representam. Pontos próximos indicam instâncias semelhantes de acordo com a medida de distância δ . Intuitivamente, pontos distantes representam objetos com pouca relação, também de acordo com δ . Assim, um ponto importante para a construção de uma projeção é a forma como as distâncias entre os objetos multidimensionais (δ) é definida.

Uma das primeiras técnicas de projeção multidimensional é a *classical multidimensional scaling*, ou simplesmente *multidimensional scaling* (MDS) (Torgerson, 1965). MDS se baseia na decomposição da matriz de distâncias entre os pares de elementos em auto-vetores, de tal modo que os que apresentam os maiores autovalores compõem o espaço transformado. Outra técnica tradicional é a *Sammon's Mapping* (Sammon, 1969). Para esta, define-se uma função de custo com base nas distâncias entre o espaço original e o transformado, a qual é minimizada por meio de um método de gradiente descendente.

Essas duas técnicas são muito precisas em termos de preservação global das distâncias, mas apresentam elevada complexidade computacional. Buscando contornar essa limitação, mantendo uma boa preservação de distância, *Landmarks MDS* (LMDS) (Silva et al., 2004), *Pivot MDS* (Brandes et al., 2007), *Part-Linear Multidimensional Projection* (PLMP) (Paulovich et al., 2010), e a abordagem de Pekalska (Pekalska et al., 1999), aplicam estratégias similares, onde uma pequena amostra dos dados é inicialmente projetada e então o restante é interpolado para a obtenção do *layout* final.

Existem técnicas que priorizam a eficiência computacional. *Fastmap* (Faloutsos et al., 1995) é uma técnica $O(n)$ que exige apenas o cálculo de distâncias dos elementos a um pequeno conjunto de pivôs. *Random Projection* (Achlioptas, 2003) é outro exemplo de uma abordagem $O(n)$. Nesta técnica, uma transformação linear aleatória é criada e aplicada à todas as instâncias dos dados para definir o espaço transformado. Apesar da eficiência computacional dessas técnicas, a preservação de distâncias resultante é prejudicada.

A *Force Scheme* (Tejada et al., 2003a) é uma técnica de projeção que se tornou muito popular por propor um balanceamento entre precisão e desempenho computacional. Essa técnica estabelece um sistema de forças, onde inicialmente posicionam-se os objetos de forma aleatória, ou por meio de alguma heurística, e forças de atração e repulsão entre os objetos levam o sistema a um estado de equilíbrio.

Após o posicionamento inicial dos pontos, a técnica realiza iterações para aproximar as distâncias d entre os objetos projetados e as distâncias δ entre os objetos no espaço original. Na primeira iteração, a técnica considera como conjunto de entrada, Y , um

posicionamento aleatório dos pontos. Para cada ponto projetado $y_i \in Y$, calcula-se um vetor $v_{i,j} = (y_j - y_i), \forall y_j \neq y_i$ e, então, move-se então y_i na direção de v . Ao término da iteração, cada objeto sofreu um deslocamento na direção de cada outro objeto, aproximando a distância entre os elementos no espaço projetado e a distância entre os elementos no espaço original. Iterações são repetidas sucessivamente até um número previamente estabelecido.

A Figura 3.1 apresenta um exemplo de uma projeção multidimensional de uma análise química realizada sobre a qualidade de vinhos produzidos em três regiões da Itália (conjunto *Wine* do repositório UCI (Newman et al., 1998)). Os círculos são coloridos de acordo com a região em que são produzidos. Pela projeção, é possível observar que, de um modo geral, as diferentes regiões produzem vinhos com características distintas. Ao mesmo tempo, nota-se que os vinhos representados pelos círculos laranjas se encontram mais espalhados, o que pode ser um indicativo de que tal produção pode seguir um processo menos controlado.

Figura 3.1: Exemplo de projeção multidimensional do conjunto *Wine*. A proximidade entre os pontos indica a semelhança entre os elementos que representam e a cor indica a classe dos elementos.

Uma projeção multidimensional permite que diversas investigações sejam realizadas sobre um conjunto de dados. No entanto, algo que pode se mostrar pouco intuitivo aos usuários de tal representação é falta de algum recurso que ajude a entender o posicionamento dos pontos. Por exemplo, no caso de um gráfico de dispersão (*scatterplot*), sempre há os eixos x e y para orientar as análises. Porém, o conceito de eixos não possui

significado quando lidamos com projeções de dados dimensionalidade maior do que 2. Contornar tal limitação das projeções é justamente um dos objetivos dos *Biplots*, que são discutidos a seguir.

3.2 Biplots

Um *biplot* (Gabriel, 1971) pode ser entendido como uma extensão das projeções multidimensionais, a qual permite avaliar também as relações entre as dos dados. A Figura 3.2 apresenta um exemplo de *biplot* para o conjunto de dados de vinhos. Novamente, a distância entre os pontos é relacionada com a similaridade entre os vinhos que representam, de modo que pontos que se encontram próximos indicam vinhos com características em comum.

Entretanto, agora há também a representação das dimensões dos dados por meio da metáfora visual de setas, o que auxilia a compreensão das posições dos pontos. Por exemplo, pela orientação das setas é possível dizer que os vinhos das classes azul e verde apresentam um teor alcoólico mais elevado. Por outro lado, os vinhos da classe laranja vão contra a variável *color*, o que indica que apresentam uma coloração pouco intensa, sendo possivelmente vinhos brancos.

Os *biplots* possibilitam também analisar a importância de cada variável para o posicionamento dos pontos no plano. Para o exemplo, o comprimento das setas é proporcional a uma aproximação da variância das dimensões que representam. A variável *cinzas* (*ash*), por exemplo, apresenta um comprimento reduzido em comparação ao das outras, logo não é um fator de grande significância para o posicionamento dos pontos. Um possível explicação para isso é que tal substância pode ser um resíduo do processo de produção de vinhos comum para as três regiões produtoras.

A construção de um *biplot* parte do princípio de que qualquer matriz S de tamanho $n \times m$ e posto r pode ser representada por:

$$S = XY^T \quad (3.1)$$

onde X é uma matriz $n \times r$ e Y uma matriz $m \times r$, ambas de posto r (Gabriel, 1971). Assim, os valores da matriz S são obtidos pelo produto escalar entre os vetores correspondentes de X e Y . Em casos onde o posto da matriz é dois, é possível desenhar os pontos de X e Y no plano. Os pontos referentes a X são os *pontos do biplot*, enquanto os referentes a Y são as *eixos do biplot*.

Na prática, o posto de uma matriz equivale ao menor valor entre n e m (Greenacre, 2010). Assim, ao lidar com grandes conjuntos de dados multidimensionais esse valor será

Figura 3.2: Exemplo de *biplot* para o conjunto de dados de vinhos.

maior que dois e consequentemente não será possível mapear os elementos das matrizes X e Y no plano. Para contornar tal situação é comum aproximar a matriz de dados original a uma matriz de posto igual a dois e utilizar essa aproximação para criar a representação visual.

Uma das maneiras mais adotadas para encontrar essa aproximação é por meio da decomposição em valores singulares, ou simplesmente SVD (*Singular value decomposition*) (Kalman, 1996). O uso do método SVD é adequado para a construção de *biplots*, pois além de resolver o problema da aproximação, seu resultado possui um formato muito similar ao exigido pela formulação de *biplots*, apresentada na Equação 3.1.

Basicamente, usando o método SVD qualquer matriz Y de tamanho $n \times m$ e posto r pode ser expressa como o produto de três matrizes:

$$Y = U D_{\alpha} V^T \quad (3.2)$$

onde U é uma matriz $n \times r$, V é uma matriz $m \times r$ e D_{α} é uma matriz diagonal $r \times r$ com autovalores $\alpha_1, \alpha_2, \dots, \alpha_r$ em uma ordem decrescente.

Para obter o formato estabelecido na Equação 3.1 basta distribuir a matriz D às outras matrizes. Dependendo do modo que essa distribuição é realizada diferentes resultados visuais são obtidos. Ao se atribuir D a U destaca-se as relações entre as instâncias de dados. Quando isso é feito em relação a V destaca-se as relações entre os atributos. E

quando se atribui parcialmente D a ambas matrizes U e V obtém-se um *biplot* simétrico que não prioriza características específicas dos dados. Independentemente do posto da matriz ser igual a dois, utiliza-se apenas os dois primeiros vetores de U e V para a criação da representação visual. Assim, a qualidade do resultado dependerá do erro da aproximação e da dimensionalidade intrínseca dos dados.

O método baseado em SVD permite a construção de *biplots* para conjuntos de dados multidimensionais. Porém, o layout obtido é restrito à matriz U , ou seja, esse método não permite que diferentes técnicas de projeção dimensional sejam utilizadas para definir o posicionamento dos elementos no plano. Há um método alternativo aos *biplots* tradicionais que viabilizam essa flexibilidade, são os chamados *biplots* de regressão (Greenacre, 2010). Dado um conjunto de dados X com n elementos m -dimensionais e sua projeção Y com n elementos bidimensionais, as coordenadas dos eixos do *biplot* de regressão, $C = \{c_1, \dots, c_m\}$, podem ser encontradas a partir dos coeficientes, $B = \{b_1, \dots, b_m\}$, da regressão linear:

$$X_i = Y B_i \quad (3.3)$$

É necessário um passo adicional para normalizar os comprimentos das setas do *biplot* de regressão, dado por:

$$\hat{C}_i = \frac{C_i \sigma(Y)}{\sigma(X_i)} \quad (3.4)$$

A precisão do *biplot* de regressão dependerá do erro de aproximação da regressão linear e também da precisão da técnica de projeção adotada. Para os resultados apresentados nos próximos capítulos, define-se a criação dos *biplots* por meio do método de regressão com base na técnica de projeção *Force Scheme*. Dentre as diversas técnicas que foram estudadas ao longo deste trabalho de mestrado, essa foi a que apresentou melhor relação entre desempenho computacional e precisão dos resultados.

No capítulo a seguir, apresenta-se os mecanismos interativos de transformação que foram desenvolvidos com base nos conceitos aqui discutidos.

Transformação Interativa de Dados Multidimensionais

4.1 Considerações Iniciais

N O Capítulo 2 foram apresentados os métodos que buscam modificar os conjuntos de dados para torná-los mais representativos para o problema em estudo. Discutiu-se que os métodos automáticos impedem que os usuários orientem essas modificações e ao mesmo tempo imponham seus conhecimentos sobre os resultados. Apresentou-se as ferramentas visuais que surgem como uma interessante alternativa aos métodos automáticos, pois permitem a interação dos usuários, mas que ainda apresentam certas limitações em relação às interfaces utilizadas e aos mecanismos de interação propostos.

O uso de ferramentas visuais que operam sobre grandes volumes de dados não é exclusivo aos trabalhos relacionados ao aqui proposto. Na verdade, toda a área de Mineração Visual de Dados (Wong, 1999) (MVD), *Visual Data Mining*, tem como objetivo justamente envolver os usuários em tarefas que até então eram executadas de maneira totalmente automática. A principal motivação desta área parte do princípio de que quando o usuário consegue compreender o resultado apresentado por uma representação visual, ele confia neste resultado e consegue obter melhor proveito das análises (Wong, 1999).

Uma característica fundamental para ferramentas MVD é manter a simplicidade em todos aspectos do sistema (Wong, 1999). No entanto, muitas das ferramentas discutidas

anteriormente se baseiam em interfaces demasiadamente complexas, as quais exigem do usuário um certo período de treinamento para um uso efetivo. Tendo em vista que o objetivo das ferramentas visuais é tornar as análises mais intuitivas, qualquer tipo de obstáculo, como a necessidade de um exaustivo treinamento do usuário, pode ser desfavorável ao se comparar com os métodos automáticos.

Um outro aspecto que deve ser levado em consideração para o desenvolvimento dessas ferramentas é permitir seu uso em diversos domínios (Wong, 1999). Para isso, diferentes mecanismos de interação devem ser oferecidos, já que nenhum será capaz de operar otimamente para todas as aplicações. No entanto, unir em um único ambiente os principais mecanismos necessários para a modificação efetiva dos dados não é tarefa trivial e nenhum dos trabalhos estudados provê tal funcionalidade.

Uma questão que deve ser considerada em ferramentas de exploração de dados, sejam elas visuais ou não, é possibilitar investigações em subconjuntos dos dados. Isto é importante pois dificilmente o conjunto de dados apresentará um comportamento global, sendo mais provável que existam subconjuntos com diferentes características que devem ser avaliadas localmente (May et al., 2011b). Porém, poucos dos trabalhos estudados atentam para esta questão.

Figura 4.1: Ilustração do ciclo do processo de transformação interativa. Os dados coletados são mapeados em visualizações, nas quais o usuário pode interagir por meio de três mecanismos, seleção, combinação e construção, para transformar os dados e aprimorar as análises subjacentes na busca por conhecimento.

Levando em consideração os aspectos mencionados nos parágrafos acima: simplicidade da ferramenta, diversidade dos mecanismos de interação e avaliação global e local dos dados, este trabalho de mestrado se baseia no uso de *Biplots* (Gabriel, 1971) para superar as limitações dos atuais métodos de transformação interativa de dados. O estabelecimento

de *biplots* como base para este trabalho é adequado, pois oferecem uma representação simultânea entre itens e dimensões de forma simples. Poucas técnicas visuais são capazes de apresentar simultaneamente informações sobre itens e dimensões em uma única representação. Mesmo entre a minoria que apresenta essa característica, nenhuma é capaz de estabelecer uma coerência entre as duas representações e essa é uma propriedade de *biplots* que os tornam ferramentas únicas.

A Figura 4.1 ilustra o processo de transformação interativa de dados proposto. Os dados coletados sobre o problema em estudo são mapeados em visualizações, por meio dos mecanismos interativos o usuário pode transformar esses dados para aprimorar as análises subjacentes na busca por conhecimento. A abordagem desenvolvida se baseia no uso de *biplots* para as visualizações e na ação conjunta de três principais mecanismos para a transformação interativa dos dados. Os dois primeiros, seleção e combinação, possibilitam a redução da dimensionalidade dos dados em busca de eliminar variáveis irrelevantes e redundantes. O terceiro mecanismo, construção, permite que o usuário crie novas dimensões com base em seu conhecimento para representar informações ausentes nos dados.

Figura 4.2: *Biplot* construído a partir do conjunto de dados fictício sobre características de países utilizado como estudo de caso.

A seguir, apresenta-se os conceitos dos mecanismos de interação desenvolvidos. Para ilustrar e motivar os mecanismos, foi realizado um estudo de caso sobre o conjunto de

dados apresentado na Tabela 2.1, que contém informações fictícias sobre alguns países. O objetivo deste estudo é demonstrar os diferentes cenários nos quais os métodos desenvolvidos podem ser utilizados. O *biplot* correspondente a esse conjunto de dados é apresentado na Figura 4.2. Para facilitar o entendimento, os símbolos mais frequentemente utilizados e suas descrições podem ser encontrados na Tabela 4.1.

Tabela 4.1: Símbolos mais frequentemente utilizados e suas descrições.

Símbolo	Descrição
n	número de instâncias no espaço original e projetado.
m	número de dimensões do espaço original e número de eixos do <i>biplot</i> .
p	dimensionalidade do espaço projetado e dos eixos do <i>biplot</i> .
X	conjunto de dados no espaço original m -dimensional.
x_i	i -ésimo instância do conjunto de dados original. Sendo $x_i = (x_{i1}, \dots, x_{im})$ a representação de suas coordenadas.
$\delta(x_i, x_j)$	dissimilaridade entre as instâncias i e j no espaço original.
A	variáveis do conjunto de dados. Sendo $a_i = (a_{i1}, \dots, a_{in})$ a representação de cada variável.
Y	conjunto de pontos no espaço projetado p -dimensional.
y_i	i -ésimo ponto projetado. Sendo $y_i = (y_{i1}, \dots, y_{ip})$ a representação de suas coordenadas.
$d(x_i, x_j)$	dissimilaridade entre os pontos i e j no espaço projetado.
B	conjunto de eixos do <i>biplot</i> no espaço p -dimensional.
$\vec{b}_i$	i -ésimo eixo do <i>biplot</i> . Sendo $\vec{b}_i = (b_{i1}, \dots, b_{ip})$ sua representação vetorial.
$ \vec{b}_i $	comprimento do eixo i .
$\theta(\vec{b}_i, \vec{b}_j)$	ângulo entre os eixos i e j do <i>biplot</i> .

4.2 Mecanismo de Seleção de Itens

O mecanismo de seleção pode agir tanto sobre os itens quanto sobre as dimensões de um conjunto de dados. Do ponto de vista de itens, seu propósito é viabilizar análises locais por meio da criação de subconjuntos dos dados e para a remoção de *outliers*¹.

Por exemplo, no estudo de caso dos países, o usuário poderia estar interessado em investigar um subconjunto de países com mais detalhes. Para isso, ele selecionaria os círculos correspondentes aos países de interesse, e após a execução desta filtragem, poderia investigar o subconjunto com mais detalhes, como mostra, respectivamente as Figuras 4.3 (a) e (b). A cada seleção do usuário, projeta-se novamente os pontos e reconstrói-se o *biplot*, de modo que um novo layout 2D é estabelecido.

¹Um *outlier* é uma instância de dados que tem características incomuns em relação à maioria das instâncias do conjunto de dados (Tan et al., 2005).

Figura 4.3: Exemplo de seleção sobre itens. Em (a), o círculo azul corresponde a seleção realizada pelo usuário que é refletida em (b), onde apresenta-se o resultado da reconstrução do *biplot*, fazendo com que as relações entre os pontos e entre os eixos sejam detalhadas e melhor representadas.

O uso deste método para a remoção de *outliers* assume que a técnica de projeção utilizada é capaz de aproximar as relações de dissimilaridade δ e d , ou seja, que pontos isolados na projeção correspondem à instância de dados com valores espúrios, caso contrário, esse mecanismo irá falhar. A seguir, descreve-se o uso do mecanismo de seleção sobre dimensões.

4.3 Mecanismo de Seleção de Atributos

O mecanismo de seleção de atributos tem como objetivo auxiliar o usuário a remover dimensões irrelevantes para a compreensão do fenômeno observado. Uma aplicação deste problema pode ser encontrada no contexto de análise de expressões gênicas, onde deseja-se identificar quais genes apresentam maior relação com o diagnóstico de alguma doença. Esse mecanismo parte de uma importante propriedade dos *biplots*, a qual estabelece uma relação entre o comprimento dos eixos do *biplot* com a importância das variáveis que representam. Por exemplo, ao projetar os pontos utilizando PCA, o comprimento do eixo é proporcional à variância das dimensão que representa (Greenacre, 2010). Assim, eixos mais curtos tendem a representar dimensões menos relevantes que podem ser excluídas.

Caso o conjunto de dados sob análise apresente elevada dimensionalidade, é interessante viabilizar que o usuário tenha um método para interações em lote. Assim, ao invés

de seleccionar cada um dos eixos que indicam as dimensões que deseja remover, o usuário pode definir um comprimento mínimo para os eixos. A Figura 4.4 apresenta um exemplo de um conjunto de dados com um número maior de dimensões, onde o uso da interação em lote pode facilitar o processo de transformação dos dados. Nesse processo, toma-se o maior eixo como referência, e a partir de um limiar escolhido pelo usuário, descarta-se as dimensões representadas por eixos cuja proporção do comprimento em relação ao maior eixo não atinja um valor estabelecido. Mais formalmente, dado um limiar α definido pelo usuário, o conjunto de dimensões a serem removidas S_s resultante da seleção em lote é dado por:

$$a_i \in S_s \Leftrightarrow \frac{\|b_i\|}{\max_{b \in B} \|b\|} < \alpha, \forall a_i \in A \quad (4.1)$$

(a) Exemplos de faixas de cortes para diferentes valores do limiar α . (b) Resultado do mecanismo de seleção em lotes considerando $\alpha = 0.25$.

Figura 4.4: Ilustração da utilidade do mecanismo de seleção de dimensões em lotes. Com base em um limiar α definido pelo usuário, seleciona-se automaticamente o conjunto de dimensões a serem removidas.

O mecanismo de seleção de dimensões auxilia o usuário a lidar com o problema conhecido como todos relevantes (*all-relevants*) (Nilsson et al., 2007), que equivale a encontrar todos os atributos que são de algum modo relevantes para a compreensão do fenômeno observado. Porém, este mecanismo não é adequado para um segundo problema, chamado mínimo ótimo (*minimal optimal*) (Kohavi et al., 1997), que consiste em construir um subconjunto dos atributos de entrada evitando ao máximo a redundância entre eles. Por isso, desenvolvemos o mecanismo de combinação de atributos que é descrito a seguir.

4.4 Mecanismo de Combinação de Atributos

Como mencionado anteriormente, o objetivo do mecanismo de combinação de atributos é auxiliar o usuário a remover a redundância de dimensões nos dados. Um caso prático deste problema está na construção de um classificador, onde ao evitar redundância entre os atributos de entrada pode-se fazer com que o método obtenha ganhos no tempo de execução e na precisão dos resultados (Kohavi et al., 1997).

A identificação de atributos redundantes pode ser realizada ao se observar os ângulos entre os eixos do *biplot*. Quanto menor o ângulo entre dois eixos, maior é a correlação entre as dimensões que representam e consequentemente maior a redundância. Em um exemplo para o estudo de caso, apresentado na Figura 4.2, observa-se que os eixos correspondentes às variáveis *infraestrutura* e *segurança* estão praticamente sobrepostos. Logo, o usuário poderia descartar uma delas. Além disso, quando o ângulo entre dois eixos é próximo de 180° , ou seja, os eixos apresentam direções similares mas sentidos opostos, então as dimensões apresentam alta correlação inversa, que também equivale a atributos com informações redundantes. Logo, a redundância R entre duas dimensões i e j pode ser dada por:

$$R_{ij} = |\cos(\theta(b_i, b_j))| \quad (4.2)$$

O mecanismo de combinação de atributos consiste em criar um único atributo a partir do conjunto de dimensões definido pelo usuário. Com esse novo atributo busca-se representar o máximo da informação contida nos atributos selecionados. Para isso, utiliza-se o primeiro componente resultante da análise de componentes principais (*PCA*) sobre os atributos selecionados. A Figura 4.5 ilustra uma combinação entre os atributos *gastronomia* e *clima* do conjunto de dados de países. Nota-se que o eixo resultante possui direção semelhante aos que lhe deram origem e que as relações entre os itens e atributos não foram substancialmente alteradas.

Mais formalmente, a nova dimensão resultante do mecanismo de combinação, a_{m+1} , é equivalente ao primeiro componente principal, PC_1 , resultante da análise de componentes principais (*PCA*) sobre o conjunto de variáveis selecionadas pelo usuário S_c :

$$a_{m+1} = \{a_i : i \notin S_c\} \cup PC_1(\{x_j : j \in S\}) \quad (4.3)$$

Similarmente ao mecanismo de seleção, o mecanismo de combinação também pode ser aplicado de forma automática para viabilizar transformações em lote. Porém, esta abordagem considera a medida de redundância definida em 4.2, ao invés do comprimento dos eixos. Duas dimensões a_i e a_j independentes são representadas por eixos com $R_{ij} = 0$ e

(a) Interação do usuário.

(b) Resultado.

Figura 4.5: Exemplo de combinação de atributos. O círculo azul corresponde a seleção realizada pelo usuário de quais atributos devem ser combinados. Em (b) é possível observar o novo eixo resultante da combinação das variáveis *clima* e *gastronomia*.

dimensões perfeitamente correlacionadas por $R_{ij} = 1$. Assim, o método consiste em definir conjuntos de dimensões para serem combinadas contendo eixos com R suficientemente próximos de 1 a partir de um limiar β .

A Figura 4.6 ilustra o método para combinação de dimensões em lote. Para este caso, definiu-se $\beta = 0.85$. Assim, para pertencerem ao mesmo conjunto o ângulo entre os eixos deve ser aproximadamente entre 0° e 30° ou entre 150° e 180° . O ângulo entre os eixos 1 e 2 é de 10° , por isso eles pertencem ao mesmo conjunto. Os ângulos entre os eixos 3 e 4 e entre 3 e 5 são, respectivamente, 30° e 170° , então mesmo com o ângulo de 140° entre os eixos 4 e 5 estar fora do intervalo estabelecido, eles pertencem ao mesmo conjunto devido à relação mútua com o eixo 3.

O método em lote pode definir p diferentes conjuntos $S = \{S_1, \dots, S_p\}$ de dimensões para serem combinados cada um em uma nova dimensão, a composição de S parte do limiar β :

$$S_i = \{j : R_{ij} < \beta \ \forall \ \{j, l\} \in S_i\} \quad (4.4)$$

Como o processo de combinação depende de *PCA*, que é um método automático, em certas situações esse mecanismo pode ser menos intuitivo ao usuário do que o de seleção de atributos. No entanto, em casos onde as dimensões dos dados não apresentam um significado para o usuário, o mecanismo de combinação pode ser preferível, pois não exige

(a) Os eixos em vermelho e azul indicam dois conjuntos de dimensões a serem combinados. (b) Cada conjunto resultou em uma única dimensão.

Figura 4.6: Exemplo de uso do método para combinação de dimensões em lotes utilizando $\beta = 0.85$.

que o usuário decida quais atributos devem ser descartados. De um modo geral, o uso concomitante dos dois mecanismos é o mais aconselhável. Acreditamos que em conjunto eles são suficientes para reduzir eficientemente a dimensionalidade dos conjuntos de dados, proporcionando uma análise mais eficiente.

A redução de dimensionalidade é um procedimento adequado para a remoção de dados irrelevantes e redundantes, mas não permite que o usuário lide com a falta de informações essenciais para a análise do problema. Este trabalho de mestrado se propôs a ir além da redução de dimensionalidade, o segundo objetivo aqui estabelecido consiste em permitir que o usuário agregue mais incisivamente seu conhecimento ao processo para construir novas dimensões a partir do mecanismo detalhado a seguir.

4.5 Mecanismo de Construção de Atributos

Existem métodos encontrados na literatura que permitem que o usuário agregue seu conhecimento do domínio nos conjuntos de dados. No entanto, esses trabalhos modificam todo o conjunto de dados original no processo (Liu et al., 2012; Mamani et al., 2013). O mecanismo de construção de atributos desenvolvido busca conservar as dimensões originais e representar o conhecimento do usuário em novos atributos. Esses novos atributos têm o

papel de representar aspectos subjetivos do problema em estudo que não são facilmente medidos ou capturados nas etapas de coleta e armazenamento dos dados.

Suponha que a análise do estudo de caso do conjunto de dados de países tenha como objetivo ajudar uma pessoa a traçar o destino de sua próxima viagem. Os dados coletados são importantes para a sua tomada de decisão, porém sua opinião e suas experiências prévias são fatores fundamentais para a escolha. Por exemplo, essa pessoa poderia já ter visitado ou estudado sobre algum desses países e por isso já possui uma impressão de suas *hospitalidades*. Brasil e Marrocos poderiam ser marcados como países acolhedores, e Rússia e França como países não tão receptivos. Como mostra a Figura 4.7 (a), o usuário poderia marcar esses países como respostas positivas (azul) e negativas (vermelho), respectivamente. Então, com base na relação de similaridade entre esses exemplos, seria construída uma nova dimensão, apresentada em vermelho em (b), responsável por representar esse aspecto subjetivo do problema.

Nota-se pela Figura 4.7 que nova dimensão faz com que o posicionamento dos elementos e dos eixos seja rearranjado para se adaptar às novas relações de similaridade entre os elementos e dimensões. Observa-se que o eixo que representa o novo atributo apresenta uma direção significativamente distinta e tem uma relação inversa aos fatores padrão, *infraestrutura* e *segurança*, sendo mais relacionado ao *clima* e *gastronomia* de um país. Essa mesma abordagem poderia ser utilizada para agregar outros aspectos de interesse do usuário ao conjunto de dados.

(a) Interação do usuário.

(b) Resultado.

Figura 4.7: Exemplo de construção de atributos. Os exemplos positivos (azuis) e negativos (vermelhos) escolhidos pelo usuário são utilizados para a construção da nova dimensão (vermelho).

O método utilizado para a criação dos novos atributos se baseia em LDA (*Linear Discriminant Analysis*) (Ripley et al., 1995) e consiste em criar um modelo que busque maximizar a distância entre elementos de respostas diferentes e minimizar a distância entre elementos com as mesmas respostas. De um modo geral, o modelo pode ser entendido com uma matriz de transformação, que ao ser aplicada sobre a matriz de dados, retorna um vetor de comprimento igual ao número de linhas da matriz de entrada.

Mais formalmente, a matriz de transformação L é obtida a partir da decomposição das matrizes de espalhamento intra e inter-classes, S_W e S_B , respectivamente. Seja X_1 é um subconjunto de X com n_1 elementos e respostas positivas, e X_2 com n_2 elementos e respostas negativas, $X_1 \cap X_2 = \emptyset$. A média das colunas de X , X_1 e X_2 são, respectivamente, μ , μ_1 e μ_2 . As matrizes centralizadas X_i^θ são dadas por:

$$X_i^\theta = X_i - \mu_i \quad (4.5)$$

A matriz de espalhamento intra-classes (*within-class scatter*) S_W é dada por:

$$S_W = \sum_{i=1}^2 S_i \quad (4.6)$$

onde

$$S_i = \sum_{x \in X_i}^n (x - \mu_i)(x - \mu_i)^T \quad (4.7)$$

A matriz de espalhamento inter-classes (*between-class scatter*) S_B é calculada pela seguinte equação:

$$S_B = \sum_1^2 n_i(\mu_i - \mu)(\mu_i - \mu)^T \quad (4.8)$$

Então, decompõe-se $S_W^{-1}S_B$ em seus autovalores λ e autovetores v :

$$(\lambda, v) = S_W^{-1}S_B v \quad (4.9)$$

A matriz de transformação L corresponde ao autovetor com o maior autovalor. Assim, o novo atributo D é o resultado da matriz de transformação L aplicada sobre X :

$$D_{n \times 1} = X_{n \times m} L_{m \times 1} \quad (4.10)$$

Para a criação do modelo, exige-se ao menos um exemplar de resposta positiva e outro de resposta negativa. Contudo, devem ser fornecidos exemplares suficientes para capturar

as diferentes características da dimensão a ser construída. De qualquer modo, como as dimensões são construídas com base em informações parciais de alguns exemplos de entrada, pode haver situações em que certas características dos dados não sejam levadas em consideração para a criação do modelo.

Para dar visibilidade à qualidade do modelo construído, foi desenvolvido um método para expressar a incerteza do mecanismo de construção. Esse método consiste em apresentar por meio de um *heatmap* quão bem os modelos conseguem separar elementos de respostas diferentes e aproximar elementos de mesma resposta. A Figura 4.8 apresenta um exemplo dessa visualização para o conjunto de dados *Wine*.

Figura 4.8: Visualização da incerteza do mecanismo de construção de atributos. Regiões em branco e azul indicam baixa incerteza, enquanto regiões em laranja e vermelho representam alta incerteza.

Dado um conjunto de dados com n elementos $X = \{x_1, \dots, x_n\}$, a incerteza de cada elemento $U = \{u_1, \dots, u_n\}$ é calculada a partir de:

$$U_i = 1 - \max(P_i) \quad (4.11)$$

Onde P_{ij} corresponde às probabilidades de X_i pertencer às respostas $R = \{r_1, \dots, r_n\}$, $R \in \{-1, 0, 1\}$, com -1 e 1 para respostas negativas e positivas, respectivamente, e 0 para respostas indefinidas. Essa probabilidade é calculada com base nos z vizinhos mais próximos de X_i , $V_i = \{v_1, \dots, v_z\}$, de acordo com uma função de distância δ :

$$P_{ij} = \frac{|V \in R, R = j|}{z} \quad (4.12)$$

Uma vez obtidos os valores de U para todos os n elementos do conjunto de dados, cria-se a visualização da incerteza do modelo de construção de atributos. Dado as posições do elementos no espaço bidimensional, $Y = \{y_1, \dots, y_n\}$, define-se os valores correspondentes às coordenadas, c_x e c_y , de um gradiente com base em Gaussianas ao redor de cada elemento projetado no plano bidimensional:

$$f(c_x, c_y) = \sum_{i=1}^n u_i \exp \left(- \left(\frac{(c_x - y_{i0})^2}{2\sigma_x^2} + \frac{(c_y - y_{i1})^2}{2\sigma_y^2} \right) \right) \quad (4.13)$$

Os valores acumulados do gradiente podem ser convertidos para uma escala de cor para auxiliar a identificar regiões de maior incerteza. No exemplo da Figura 4.8, utilizou-se a tradicional escala de temperatura, onde regiões em branco e azul indicam baixa incerteza, enquanto regiões em laranja e vermelho representam alta incerteza. O mapa da incerteza pode ser utilizado como uma referência para as transformações realizadas e também como um guia para a escolha de novos exemplos para o processo de construção de atributos.

Neste capítulo foram apresentados os conceitos e detalhes que envolvem os mecanismos interativos de transformação. No próximo capítulo, apresenta-se uma avaliação desses mecanismos por meio de uma série de simulações do comportamento dos usuários.

Experimentos e Resultados

NESTE capítulo, apresenta-se a avaliação dos mecanismos interativos de transformação de dados multidimensionais e descreve-se a ferramenta desenvolvida. A avaliação assume que com dados mais representativos, os métodos que operam sobre eles, como algoritmos de classificação de dados, devem apresentar melhores resultados. Então, os critérios para a avaliação deste trabalho foram baseados nessa melhoria. Mais especificamente, utilizou-se a acurácia dos classificadores e o tempo de execução como medidas do desempenho de tais tarefas.

A avaliação foi dividida em duas etapas. Inicialmente, realizou-se uma exaustiva avaliação quantitativa, na qual buscou-se medir o desempenho dos mecanismos desenvolvidos a partir da simulação de possíveis interações realizadas pelo usuário. Em seguida, em uma abordagem qualitativa, buscou-se medir o desempenho dos mecanismos ao serem aplicados de forma integrada.

5.1 Avaliação quantitativa

O procedimento adotado na avaliação quantitativa consiste em simular as possíveis interações realizadas pelos usuários, nas quais busca-se tornar os dados mais representativos para o problema abordado. Diferentemente de avaliações baseadas em experimentos com usuários, essas simulações permitem que um elevado número de testes seja realizado de modo controlado, viabilizando uma investigação exaustiva dos diferentes casos de uso e a reprodutibilidade dos resultados.

Ao todo, foram realizados testes com 12 conjuntos de dados consolidados das áreas de aprendizado de máquina e de mineração de dados. Buscou-se escolher conjuntos com características diversas que permitissem avaliar as técnicas em cenários variados. As principais características desses conjuntos são apresentadas na Tabela 5.1. Esses conjuntos foram obtidos por meio do pacote *mlbench* (Leisch et al., 2010) da ferramenta R (R Core Team, 2014) e no repositório *UCI* (Newman et al., 1998).

Tabela 5.1: Características dos conjuntos de dados utilizados na avaliação.

Dataset	Elementos	Dimensões	Classes
BreastCancer	683	9	2
Diabetes	392	8	2
DNA	3186	180	3
Glass	214	9	6
Ionosphere	351	32	2
Iris	150	4	3
Sonar	208	60	2
Soybean	562	35	3
Vehicle	846	18	4
WDBC	569	30	2
Wine	178	13	3
Zoo	100	16	7

Para cada um desses conjuntos, aplicou-se as técnicas transformação utilizando diferentes parametrizações e mediu-se a acurácia e tempo de execução do classificador seguindo a abordagem de validação-cruzada (*cross-validation*). Esse procedimento é descrito com mais detalhes no Algoritmo 1, apresentado a seguir:

O método de classificação adotado, $\mathcal{C}$, foi o kNN (k *Nearest Neighbors*), justamente por ser um dos métodos de classificação mais afetados pelos efeitos da maldição da dimensionalidade (Beyer et al., 1999). Definiu-se $k = 3$, assim, para prever o rótulo de uma instância desconhecida, o classificador realiza uma votação sobre os rótulos dos 3 elementos mais similares do conjunto de treinamento. O valor de nf foi estabelecido como 10, assim a cada iteração do classificador, utiliza-se 10% do conjunto de dados para treinamento e o restante para teste.

Dado esse escopo geral, a seguir discute-se com mais detalhes os procedimentos adotados para avaliar cada um dos mecanismos interativos de transformação.

5.1.1 Mecanismo de Seleção

Nesta seção, discute-se sobre a eficácia do mecanismo de seleção de atributos. Não foram considerados testes no contexto de seleção de itens, pois trata-se de uma tarefa

---

Algoritmo 1 Avaliação das Técnicas

$\mathcal{X}$: conjunto de dados

$\mathcal{T}$: transformação

$\mathcal{C}$: método de classificação

nf : números de *folds*

função ($\mathcal{X}, \mathcal{T}, \mathcal{C}, nf$)

$\mathcal{X}' \leftarrow \mathcal{T}(\mathcal{X})$

▷ aplica a transformação aos dados

$folds \leftarrow \text{REPARTE}(\mathcal{X}', nf)$

▷ divide os dados em nf folds

para todos $i \in (1 : nf)$ **faça**

$\mathcal{X}'_{treino} \leftarrow folds \in folds_i$

$\sigma \leftarrow \mathcal{C}(\mathcal{X}'_{treino})$

▷ constrói o classificador

$\mathcal{X}'_{teste} \leftarrow folds \notin folds_i$

$(a, t) \leftarrow (a, t) + \sigma(\mathcal{X}'_{teste})$

▷ aplica o classificador ao conjunto de teste

fim para

retorna $(a, t)/nf$

fim função

a : acurácia do classificador

t : tempo de execução

---

muito específica para cada situação, o que impede uma avaliação comparativa. Buscou-se simular o comportamento de um usuário que deseja remover atributos irrelevantes dos conjuntos de dados. Como apresentado no Capítulo 4, o mecanismo de seleção pode ser utilizado para este propósito por meio da remoção de variáveis que sejam representadas por eixos de curto comprimento. Assim, a simulação consiste em avaliar se o comprimento dos eixos pode ser considerado como um critério adequado para a seleção de atributos.

Foram realizados diversos testes, buscando-se avaliar os possíveis pontos de corte. A abordagem adotada consiste em variar o valor de α definido na Equação 4.1, que equivale a proporção do comprimento total do maior eixo. Variou-se esse parâmetro de 0.00 a 0.99, de modo que todas as dimensões representadas por eixos com comprimentos correspondentes menores que α fossem descartadas. Para cada valor do parâmetro, executou-se 100 vezes o procedimento descrito pelo Algoritmo 1, onde mediu-se a acurácia e o tempo de execução do classificador.

Os resultados obtidos nas simulações é apresentado nas Figuras 5.1 e 5.2. Para cada gráfico, o eixo horizontal acompanha o parâmetro de α utilizado, sendo que inicia em 0.00 e termina em 0.99. Por sua vez, no eixo horizontal são apresentadas três informações que podem variar de 0% a 100%. Em vermelho, apresenta-se a média da acurácia do classificador. Em azul, indica-se o tempo de execução do classificador, relativo ao maior tempo. O número de dimensões em relação ao conjunto inicial é representado pela cor verde, sendo que, por exemplo, um valor de 50% indica que o número de dimensões após a transformação é metade do original.

(a) BreastCancer

(b) Diabetes

(c) DNA

(d) Glass

(e) Ionosphere

(f) Iris

Figura 5.1: Avaliação do mecanismo de seleção (1 de 2). Apresenta-se a acurácia em vermelho, o tempo relativo em azul e a dimensionalidade relativa em verde.

(a) Sonar

(b) Soybean

(c) Vehicle

(d) WDBC

(e) Wine

(f) Zoo

Figura 5.2: Avaliação do mecanismo de seleção (2 de 2). Apresenta-se a acurácia em vermelho, o tempo relativo em azul e a dimensionalidade relativa em verde.

A análise dos resultados das simulações do mecanismo de seleção de atributos deve levar em consideração uma observação importante. Primeiramente, a medida que o valor de α aumenta, mais dimensões serão removidas, até o ponto que restará somente a dimensão representada pelo eixo de maior comprimento. Assim, a linha verde nas figuras, que indica a proporção do número de dimensões, deve ser utilizada como uma referência para as comparações de tempo e acurácia.

Em relação ao tempo de execução dos classificadores, é possível observar uma redução para praticamente todos os conjuntos de dados. Em alguns casos essa redução pode ser mais drástica devido ao tamanho e complexidade dos conjuntos de dados. Para o conjunto de dados *Ionosphere*, por exemplo, que apresenta 351 elementos e 32 dimensões, o decaimento é mais acentuado do que para o conjunto *Diabetes*, que apesar de conter 392 elementos, possui apenas 8 dimensões.

Ao analisar os resultados de acurácia, nota-se comportamentos interessantes. Primeiramente, nota-se que existe um claro limite para o uso do mecanismo de seleção, sendo que a remoção de dimensões representadas por eixos de tamanho significativo, acarreta em uma redução na acurácia do classificador. Esse decaimento é um indicador de um ponto positivo para este mecanismo, pois indica que o tamanho dos eixos tem uma relação direta com a importância das dimensões que representam. Outro ponto de interesse é o aumento da acurácia nos intervalos iniciais do mecanismo, isto é, o mecanismo está sendo capaz de remover dimensões irrelevantes, fazendo com que tarefas subjacentes sejam executadas com mais eficácia.

Uma análise conjunta de tempo de execução e acurácia destaca ainda mais a utilidade do mecanismo de seleção. Para grande parte dos conjuntos de dados, nota-se que pode ser obtida uma redução no tempo de execução sem qualquer prejuízo à acurácia. Mais do que isso, em alguns casos, como o do conjunto *DNA*, nota-se uma drástica queda no tempo de execução e, ao mesmo tempo, um aumento substancial na acurácia do classificador.

5.1.2 Mecanismo de combinação

Na seção anterior, avaliou-se a capacidade do mecanismo de seleção em remover atributos irrelevantes dos conjuntos de dados. Aqui, avalia-se a eficácia do mecanismo de combinação para tratar o problema de atributos redundantes. O procedimento adotado é similar ao adotado anteriormente, variou-se o valor de β do método para combinação de dimensões em lotes, descrito na Equação 4.4.

Para a bateria de testes, variou-se β de 0.00 a 0.99. Para cada valor de β executou-se 100 vezes o procedimento descrito pelo Algoritmo 1, medindo-se a acurácia e o tempo de execução do classificador. O resultado obtido nas simulações pode ser observado. Os resultados obtidos nas simulações é apresentado nas Figuras 5.3 e 5.4. O eixo horizontal

indica o parâmetro de β utilizado, que varia de 0.00 a 0.99. Exatamente como nos resultados do mecanismo de seleção, o eixo horizontal apresenta três informações que podem variar de 0% a 100%: a média da acurácia do classificador em vermelho, o tempo de relativo de execução do classificador em azul e o número relativo de dimensões em verde.

O resultado das simulações do mecanismo de combinação é muito similar ao do mecanismo de seleção. Novamente, observa-se um decaimento do tempo de execução do classificador, que é mais ou menos acentuado de acordo com a complexidade do conjunto de dados. Para a maioria dos conjuntos é possível obter níveis significativos de redução de tempo de processamento sem perda na acurácia do classificador. Também observa-se, em alguns casos, um aumento considerável na acurácia do classificador nos intervalos iniciais da transformação.

No entanto, nota-se uma diferença entre os resultados dos mecanismos. Nos intervalos finais do mecanismo de seleção, nota-se uma redução significativa na acurácia do classificador. Isso ocorre, pois dimensões relevantes estão sendo descartadas. Já para o mecanismo de combinação essa redução é menos acentuada. Esse comportamento pode ser explicado pelo fato que para este mecanismo as informações não são completamente perdidas como no caso do de seleção. Por mais que duas dimensões possam ser distintas, ao combina-las mantém-se parte da informação contida em cada uma delas, isso faz com que o caso extremo dos intervalos finais dos parâmetros da simulação não prejudiquem demasiadamente o resultado da classificação.

Nessas duas seções, avaliou-se a capacidade dos mecanismos interativos em lidar com dimensões irrelevantes e redundantes do conjunto de dados. Na seção a seguir será avaliada a eficácia do mecanismo de construção de dimensões.

5.1.3 Mecanismo de Construção

A avaliação do mecanismo de construção segue a mesma abordagem dos outros mecanismos. Desta vez, a simulação consiste em observar como o número de exemplos de entrada para o modelo afeta a representatividade das dimensões construídas. Para viabilizar a realização dos testes, a classe dos conjuntos de dados foi utilizada como variável resposta para o modelo. Foi estabelecido um parâmetro s correspondente à proporção do dessa variável que era utilizada como entrada. Variou-se o valor de s de 0% a 100%, com 100% todos os elementos da variável resposta eram utilizados.

O resultado obtido nas simulações é apresentado nas Figuras 5.5 e 5.6. A apresentação dos resultados deste mecanismo difere dos anteriores. Como o objetivo da análise é verificar a capacidade do mecanismo em representar o conhecimento do usuário em novos atributos, então mediu-se apenas a acurácia do classificador. A linha vermelha indica a

(a) BreastCancer

(b) Diabetes

(c) DNA

(d) Glass

(e) Ionosphere

(f) Iris

Figura 5.3: Avaliação do mecanismo de combinação (1 de 2). Apresenta-se a acurácia em vermelho, o tempo relativo em azul e a dimensionalidade relativa em verde.

(a) Sonar

(b) Soybean

(c) Vehicle

(d) WDBC

(e) Wine

(f) Zoo

Figura 5.4: Avaliação do mecanismo de combinação (2 de 2). Apresenta-se a acurácia em vermelho, o tempo relativo em azul e a dimensionalidade relativa em verde.

acurácia do classificador ao utilizar o atributo construído e a pontilhada em preto indica a acurácia do classificador utilizando os dados originais. Nota-se que foi utilizado um intervalo variável para o eixo vertical, buscando dar mais destaque para as diferenças entre as medidas. Por sua vez, o eixo horizontal indica a proporção dos dados que foram utilizados como entrada para a criação do modelo de construção.

De um modo geral, o resultado da avaliação quantitativa é positivo, pois para praticamente todos os conjuntos de dados há um aumento significativo na acurácia do classificador. O único caso em que o valor obtido é menor do que ao se utilizar os dados originais é para o conjunto *Soybean*. Porém, este conjunto já apresenta um valor altíssimo de acurácia, então a adição de novas dimensões não parece ser necessária.

Um dos comportamentos que pode ser observado com frequência é um decaimento na acurácia logo no início do gráfico. Isso ocorre, pois, inicialmente, o tamanho do conjunto de exemplos de entrada do mecanismo de construção é muito pequeno, sendo que o modelo não consegue capturar as características presentes nos dados. A medida que esse conjunto cresce, também aumenta a qualidade do modelo e, consequentemente, a acurácia do classificador. Nota-se que, na maioria dos casos, o ponto ótimo é alcançado com menos de 25% dos dados.

Apesar de, em alguns casos, 25% representar um número elevado de elementos a serem rotulados pelo usuário, acreditamos que o valor ótimo de acurácia poderia ser mais rapidamente alcançado, caso a amostragem não fosse realizada de forma aleatória, mas sim pelo próprio usuário. Vale lembrar que desenvolvemos um método visual para apresentar a incerteza do mecanismo de construção justamente para essas situações, buscando ajudar o usuário a encontrar quais elementos devem ser mais significativos para a criação do modelo.

Os resultados apresentados nesta seção dão indícios da eficiência e eficácia dos mecanismos desenvolvidos, mas aborda cada mecanismo de forma individual. Como uma das grandes motivações deste trabalho é justamente propor um arcabouço integrado para as transformações, é importante avaliar o comportamento dos métodos desenvolvidos ao serem aplicados concomitantemente. Esta avaliação é o objetivo da próxima seção.

(a) BreastCancer

(b) Diabetes

(c) DNA

(d) Glass

(e) Ionosphere

(f) Iris

Figura 5.5: Avaliação do mecanismo de criação (1 de 2). Apresenta-se a acurácia do classificador utilizando (vermelho) e sem utilizar (preto) o novo atributo.

(a) Sonar

(b) Soybean

(c) Vehicle

(d) WDBC

(e) Wine

(f) Zoo

Figura 5.6: Avaliação do mecanismo de criação (2 de 2). Apresenta-se a acurácia do classificador utilizando (vermelho) e sem utilizar (preto) o novo atributo.

5.2 Avaliação qualitativa

A avaliação qualitativa consiste em ilustrar o uso dos três mecanismos de transformação de forma integrada. A medida para avaliação é a acurácia do classificador kNN , calculada como na seção anterior. Escolheu-se o conjunto de dados *DNA* para essa demonstração, pois é o que contém o maior número de elementos e maior complexidade em termos de número de dimensões. Este conjunto possui 3.186 elementos que representam conformações de *DNA*, sendo que cada um é descrito por 180 variáveis binárias. O conjunto é dividido em três classes: conformações que correspondem a um *exão*, a um *intrão* ou outro segmento (Leisch et al., 2010).

Inicialmente, projeta-se os dados utilizando *PCA* e cria-se o *biplot*. A visualização resultante é apresentada na Figura 5.7, a cor dos círculos indica a classe do elemento que representam. Para o conjunto de dados original, a acurácia do classificador é 0.68.

Figura 5.7: *Biplot* inicial do conjunto de dados *DNA* ao se projetar os dados com o método *PCA*. Os 3.186 elementos do conjunto de dados são representados por círculos, sendo que suas cores indicam a classe a qual pertencem. As 180 variáveis são representadas pelos eixos do *biplot*. A acurácia do classificador sobre os dados é 0.68.

Nota-se pela Figura 5.7 que há um aglomerado de dimensões pouco relevantes no centro do *biplot*. O mecanismo de seleção de dimensões é ideal para essas situações. Ao utilizar o método de seleção em lotes com $\alpha = 0.1$, 145 dimensões são removidas. A acurácia do classificador aplicado sobre o conjunto de dados transformado é 0.79, um aumento significativo em relação aos dados originais. A visualização resultante é apresentada na Figura 5.8, onde é possível notar o motivo do aumento na acurácia do classificador, há uma melhor separação dos elementos de classes distintas.

Figura 5.8: Resultado da seleção em lote com $\alpha = 0.1$. Das 180 variáveis originais, 145 variáveis são removidas. A acurácia do classificador sobre os dados transformados é 0.79.

Após remover as variáveis menos relevantes, é interessante reduzir a redundância das variáveis restantes. Isso pode ser alcançado por meio do mecanismo de combinação de variáveis. Ao se aplicar o método de combinação em lotes com $\beta = 0.3$ sobre o resultado da transformação anterior, obtém-se um conjunto de dados com 15 dimensões. Das 35 dimensões resultantes da seleção realizada anteriormente, apenas 5 são conservadas e 10 novas são obtidas a partir de combinações das restantes. O novo valor da acurácia do classificador sobre os dados transformados é 0.83. O resultado visual da transformação é

apresentado na Figura 5.9, novamente é possível observar que há uma melhor separação entre elementos de classes distintas.

Figura 5.9: Ao se aplicar o método de combinação em lotes com $\beta = 0.3$ sobre o resultado da transformação anterior, com 35 variáveis, obtém-se um conjunto de dados com 15 dimensões. A acurácia do classificador sobre os dados transformados é 0.83.

No caso deste exemplo, a classe de todos os elementos do conjunto de dados é conhecida, mas esse não é o cenário usual em situações reais. Normalmente, apenas uma pequena parcela dos dados é rotulada. Por um momento, suponha que a informação de classe do conjunto *DNA* é desconhecida. Assume-se apenas que o usuário tem conhecimento sobre o domínio e é capaz de definir a classe de uma conformação ao analisar seus atributos. Neste caso, o usuário poderia rotular uma amostra do conjunto de dados e por meio do mecanismo de construção, poderia criar novas variáveis para representar essa preciosa informação.

A Figura 5.10 exemplifica o resultado do mecanismo de construção considerando que o usuário rotulou 5% do conjunto de dados, aproximadamente 160 conformações, obtendo-se 3 novas dimensões, uma para cada classe. Nota-se pela figura, que os eixos correspondentes às novas dimensões construídas (em vermelho) apontam para as regiões

de maior concentração de cada classe. A acurácia do classificador para os novos dados é 0.86.

Figura 5.10: São construídas 3 novas dimensões a partir de uma amostra rotulada pelo usuário de equivalente a 5% do conjunto de dados. A acurácia do classificador sobre os dados transformados é 0.86.

Ao longo do processo de transformação dos dados, o usuário pode utilizar o mapa de incerteza como um guia para suas transformações. A Figura 5.11 apresenta esse mapa para o conjunto de dados original e para os dados após as transformações realizadas. Inicialmente, a incerteza estava espalhada homogeneamente, dificultando que o usuário pudesse identificar regiões de interesse. Após as transformações, foram reveladas regiões com maior concentração de elementos incertos, permitindo ao usuário analisar essas regiões com mais detalhes e, eventualmente, rotular novos elementos para o mecanismo de construção, a fim de reduzir a incerteza na região.

Nesta seção, mostrou-se que ao se utilizar os mecanismos desenvolvidos de forma integrada é possível transformar complexos conjuntos de dados de forma eficiente. A seguir, apresenta-se a ferramenta desenvolvida que une os conceitos e métodos apresentados, viabilizando a utilização dos mecanismos de forma integrada.

(a) Mapa de incerteza antes das transformações.

(b) Mapa de incerteza após as transformações.

Figura 5.11: Mapa de incerteza antes e após as transformações.

5.3 Ferramenta

Os três mecanismos de transformação foram integrados em uma ferramenta, na qual podem ser usados concomitantemente. A Figura 5.12 ilustra essa ferramenta. O *back-end* foi desenvolvido na linguagem *R* e para o *front-end* utilizou-se *Javascript* e a biblioteca *D3.js*. A aplicação roda em qualquer *browser* moderno, o que facilita sua divulgação. Atualmente, não se encontra disponível, mas após sua publicação, poderá ser livremente acessada.

Figura 5.12: Ilustração da ferramenta desenvolvida.

Pela Figura 5.12 é possível observar as principais características da ferramenta. No início do painel esquerdo, o usuário pode navegar pelas pastas e arquivos disponíveis. Mais abaixo, pode escolher alguma das variáveis para ser considerada como a classe do conjunto, fazendo com que a cor dos círculos indique a qual classe pertencem. Em seguida, há uma caixa com todas as dimensões do conjunto de dados que permite visualizar e escolher quais variáveis estão sendo utilizadas. Ainda no painel da esquerda, permite-se que o usuário mantenha ou remova um conjunto de itens ou dimensões selecionadas ou que restaure o conjunto de dados ao estado inicial. O último item do painel esquerdo dá ao usuário duas opções para realizar a combinação das variáveis selecionadas, por meio de PCA ou, simplesmente, calculando uma média entre elas.

O painel central corresponde à visualização dos dados e possui recursos interativos como *drag* para selecionar conjuntos de itens ou dimensões, *click* para mostrar detalhes

de um elemento e definir a entrada do mecanismo de construção, e *hover* para apresentar uma breve descrição do item ou da dimensão sob o *mouse*.

O painel direito contém um filtro para os elementos visuais, no qual o usuário pode mostrar ou esconder as setas do *biplot*, o *heatmap* da incerteza ou os rótulos dos itens e dimensões. Há opções para definir o escopo dos elementos selecionados com o *mouse*, as quais permitem que o usuário selecione somente itens, apenas dimensões ou ambos. Este painel contém alguns controles para alterar a aparência do *heatmap*, como a escala de cor adotada. Finalmente, o último item do painel direito permite que o usuário escolha qual técnica de projeção multidimensional deseja utilizar, sendo que, atualmente, as opções são PCA, *Force-Scheme* (Tejada et al., 2003b) e *tSNE* (Maaten et al., 2008).

Neste capítulo, apresentamos uma avaliação exaustiva dos mecanismos interativos de transformação desenvolvidos. Os resultados obtidos por meio das simulações dão indícios da eficácia dos métodos desenvolvidos e do sucesso deste trabalho. No próximo capítulo apresentamos a conclusão deste projeto, discutindo-se sobre as principais contribuições do trabalho.

Conclusão

A exploração de conjuntos de dados é um problema abordado com frequência em diversos domínios de aplicação e tem como objetivo uma melhor compreensão de fenômenos simulados ou medidos. Tal atividade é precedida pelas etapas de coleta e armazenamento de dados que buscam registrar o máximo de detalhes sobre algum fenômeno observado. Porém, a exploração efetiva dos dados envolve uma série de desafios. Um deles é a dificuldade em identificar quais dados são realmente relevantes para as análises. Outro problema está relacionado com a falta de garantias de que os fatores fundamentais para a compreensão do problema tenham sido coletados.

Neste trabalho, apresentou-se como as técnicas de transformação interativa de dados podem ser utilizadas para abordar os desafios do processo exploratório. Mostrou-se que os trabalhos disponíveis na literatura possuem limitações, como interfaces demasiadamente complexas e mecanismos de interação pouco flexíveis. Assim, este projeto de mestrado teve como objetivo desenvolver novas técnicas visuais interativas para a transformação de dados multidimensionais em busca de eliminar variáveis irrelevantes ou redundantes e construir novas dimensões para representar informações ausentes. Esse objetivo foi alcançado por meio do desenvolvimento de três mecanismos de interação: seleção, combinação e construção. Os três mecanismos foram integrados em uma ferramenta, na qual podem ser usados concomitantemente.

O mecanismo de seleção pode agir tanto sobre os itens quanto sobre as dimensões de um conjunto de dados. Do ponto de vista de itens, seu propósito é viabilizar análises locais por meio da criação de subconjuntos dos dados e para a remoção de *outliers*. Em

relação aos atributos, o mecanismo de seleção tem como objetivo auxiliar o usuário a remover dimensões irrelevantes para a compreensão do fenômeno observado.

O mecanismo de seleção de dimensões auxilia o usuário a encontrar todos os atributos que são de algum modo relevantes para a compreensão do fenômeno observado. Porém, este mecanismo não é adequado para construir um subconjunto dos atributos de entrada evitando ao máximo a redundância entre eles. Para isso, foi desenvolvido o mecanismo de combinação de atributos.

Os mecanismos de seleção e combinação de atributos são adequados para a remoção de dados irrelevantes e redundantes, mas não permitem que o usuário lide com a falta de informações essenciais para a análise do problema. O mecanismo de construção de atributos desenvolvido busca conservar as dimensões originais e representar o conhecimento do usuário em novos atributos. Esses novos atributos têm o papel de representar aspectos subjetivos do problema em estudo que não são facilmente medidos ou capturados nas etapas de coleta e armazenamento dos dados.

A eficiência desses mecanismos foi confirmada por meio de intensas baterias de testes que buscaram simular o comportamento dos usuários em diversos cenários. Os resultados dos experimentos realizados sobre diversos conjuntos de dados dão indícios de que os métodos desenvolvidos possibilitam a obtenção de conjuntos de dados mais representativos. Mais especificamente, foram obtidos melhores resultados em tarefas de classificação de dados ao utilizar os métodos desenvolvidos.

Acreditamos que a grande contribuição deste projeto foi estabelecer um balanceamento adequado entre a interação do usuário e o uso de métodos automáticos. Os métodos para transformações em lote se mostrou muito eficiente para lidar com conjuntos de alta dimensionalidade. Conseguimos abordar problemas que não eram bem resolvidos pelos trabalhos do estado-da-arte e desenvolvemos uma abordagem promissora que não sobrecarrega o usuário com interfaces demasiadamente complexas e que não segue o paradigma *black-box*, permitindo a compreensão e interação ao longo do processo de transformação dos dados de forma eficaz e eficiente.

Ao longo deste projeto, foi tomada a decisão de modificar o comportamento do mecanismo de construção de atributos. Inicialmente, foi proposto que o mecanismo deveria permitir que o usuário alterasse o posicionamento dos pontos em busca de modificar as relações de similaridade dos itens do conjunto de dados. Essa alteração deveria ser a base para a construção das novas dimensões. No entanto, consideramos que esse tipo de interação é muito subjetiva e dificulta a descrição do atributo construído. Por exemplo, no estudo de caso apresentado no Capítulo 4, era intuitivo compreender que a nova variável correspondia à hospitalidade dos países. Porém, imagine que, ao invés de selecionar respostas positivas e negativas, fosse necessário re-posicionar os pontos. As

novas relações de similaridade não seriam suficientes para estabelecer uma semântica para essa transformação. Além disso, ao alterar as relações de similaridade, o usuário faria com que os atributos conhecidos também sejam modificados, este é uma consequência inerente deste tipo de transformação.

Um possível trabalho futuro para este projeto consiste em incrementar o mecanismo de construção para permitir que lide com respostas com valores contínuos. Atualmente, o método que serve como base, LDA, permite somente que valores categóricos sejam utilizados. Entretanto, para algumas tarefas parece mais intuitivo permitir que o usuário estabeleça valores em um intervalo, ao invés, de resposta positiva ou negativa. Para o mesmo estudo de caso apresentado no Capítulo 4, poderia ser mais interessante possibilitar ao usuário definir, do seu ponto de vista, quão acolhedor são os países, por exemplo, fornecendo uma nota de 0 a 10, assim como os outros atributos do conjunto de dados.

Bibliografia

- Achlioptas, D. (2003). “Database-friendly random projections: Johnson-Lindenstrauss with binary coins”. Em: *Journal of Computer and System Sciences* 66.4, pp. 671 –687.
- Amaldi, E. e V. Kann (1998). “On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems”. Em: *Theoretical Computer Science* 209.1, pp. 237–260.
- Artero, A., M. de Oliveira e H. Levkowitz (2006). “Enhanced High Dimensional Data Visualization through Dimension Reduction and Attribute Arrangement”. Em: *Information Visualization, 2006. IV 2006. Tenth International Conference on*, pp. 707–712.
- Bellman, R. E. (1961). *Adaptive control processes - A guided tour*. Princeton, New Jersey, U.S.A.: Princeton University Press, p. 255.
- Beyer, K. et al. (1999). “When Is ”Nearest Neighbor”Meaningful?” Em: *In Int. Conf. on Database Theory*, pp. 217–235.
- Brandes, U. e C. Pich (2007). “Eigensolver methods for progressive multidimensional scaling of large data”. Em: *Proceedings of the 14th international conference on Graph drawing*. GD’06. Karlsruhe, Germany: Springer-Verlag, pp. 42–53.
- Brandoli, B. et al. (2010). “Visual Data Exploration to Feature Space Definition”. Em: *Proceedings of the 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images*. SIBGRAPI ’10. Washington, DC, USA: IEEE Computer Society, pp. 32–39.
- Choo, J. et al. (2010). “iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction”. Em: *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pp. 27–34.
- Cox, M. A. A. e T. F. Cox (2008). “Multidimensional Scaling”. English. Em: *Handbook of Data Visualization*. Springer Handbooks Comp.Statistics. Springer Berlin Heidelberg, pp. 315–347.
- Dill, J. et al. (2012). *Expanding the Frontiers of Visual Analytics and Visualization*. Springer.

- Dy, J. G. e C. E. Brodley (2000). “Visualization and interactive feature selection for unsupervised data”. Em: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '00. Boston, Massachusetts, USA: ACM, pp. 360–364.
- Faloutsos, C. e K.-I. Lin (1995). “FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets”. Em: *Proceedings of the 1995 ACM SIGMOD international conference on Management of data*. SIGMOD '95. San Jose, California, USA: ACM, pp. 163–174.
- Fodor, I. (2002). *A Survey of Dimension Reduction Techniques*. Rel. téc. Lawrence Livermore National Lab., CA (US).
- Friendly, M. (2002). “Corrgrams: Exploratory Displays for Correlation Matrices”. Em: *The American Statistician* 56.4, pp. 316–324.
- Fukunaga, K. (1990). *Introduction to statistical pattern recognition (2nd ed.)* San Diego, CA, USA: Academic Press Professional, Inc.
- Gabriel, K. R. (1971). “The biplot graphic display of matrices with application to principal component analysis”. Em: *Biometrika* 58.3, pp. 453–467.
- Greenacre, M. (2010). *Biplots in Practice*. Fundación BBVA.
- Guo, D. (2003). “Coordinating computational and visual approaches for interactive feature selection and multivariate clustering”. Em: *Information Visualization* 2.4, pp. 232–246.
- Guyon, I. e A Elisseeff (2003). “An introduction to variable and feature selection”. Em: *The Journal of Machine Learning Research* 3, pp. 1157–1182.
- Harding, J. a., M. Shahbaz e a. Kusiak (2006). “Data Mining in Manufacturing: A Review”. Em: *Journal of Manufacturing Science and Engineering* 128.4, p. 969.
- Healey, C. G., K. S. Booth e J. T. Enns (1995). “Visualizing real-time multivariate data using preattentive processing”. Em: *ACM Trans. Model. Comput. Simul.* 5.3, pp. 190–221.
- Ingram, S. et al. (2010). “DimStiller: Workflows for dimensional analysis and reduction”. Em: *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pp. 3–10.
- Jeong, D. H. et al. (2009). “iPCA: An Interactive System for PCA-based Visual Analytics”. Em: *Computer Graphics Forum* 28.3, pp. 767–774.
- Johansson, S. e J. Johansson (2009). “Interactive Dimensionality Reduction Through User-defined Combinations of Quality Metrics”. Em: *IEEE Transactions on Visualization and Computer Graphics* 15.6, pp. 993–1000.
- Jolliffe, I. T. (2002). *Principal Component Analysis*. Second. Springer.
- Kalman, D. (1996). “A Singularly Valuable Decomposition: The SVD of a Matrix”. Em: *The College Mathematics Journal* 27.1, pp. 2–23.

- Kaski, S. e J. Peltonen (2011). “Dimensionality Reduction for Data Visualization”. Em: March, pp. 100–104.
- Keim, D. (2002). “Information visualization and visual data mining”. Em: *Visualization and Computer Graphics, IEEE Transactions on* 8.1, pp. 1–8.
- Kohavi, R. e G. H. John (1997). “Wrappers for feature subset selection”. Em: *Artif. Intell.* 97.1-2, pp. 273–324.
- Kohonen, T. (1990). “The self-organizing map”. Em: *Proceedings of the IEEE* 78.9, pp. 1464–1480.
- König, A. (2000). “Dimensionality reduction techniques for multivariate data classification, interactive visualization, and analysis-systematic feature selection vs. extraction”. Em: *Knowledge-Based Intelligent Engineering Systems and Allied Technologies, 2000. Proceedings. Fourth International Conference on*. Vol. 1, 44–55 vol.1.
- Kouiroukidis, N. e G. Evangelidis (2011). “The Effects of Dimensionality Curse in High Dimensional kNN Search”. Em: *Informatics (PCI), 2011 15th Panhellenic Conference on*, pp. 41–45.
- Krause, J., A. Perer e E. Bertini (2014). “INFUSE: Interactive Feature Selection for Predictive Modeling of High Dimensional Data”. Em: *Visualization and Computer Graphics, IEEE Transactions on* 20.12, pp. 1614–1623.
- Leisch, F. e E. Dimitriadou (2010). *mlbench: Machine Learning Benchmark Problems*. R package version 2.1-1.
- Liu, J. et al. (2012). “Dis-function: Learning Distance Functions Interactively”. Em: *Proceedings of the 2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. VAST '12. Washington, DC, USA: IEEE Computer Society, pp. 83–92.
- Maaten, L. van der e G. Hinton (2008). “Visualizing High-Dimensional Data Using t-SNE”. Em:
- Maaten, L. Van der, E. Postma e H. Van den Herik (2009). “Dimensionality reduction: A comparative review”. Em: *Technical Report TiCC TR 2009-005*.
- MacEachren, A. et al. (2003). “Exploring high-D spaces with multiform matrices and small multiples”. Em: *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pp. 31–38.
- Mamani, G. M. H. et al. (2013). “User-driven Feature Space Transformation”. Em: Eurographics Association.
- May, T, J Davey e T Ruppert (2011a). “SmartStripes-Looking under the Hood of Feature Subset Selection Methods”. Em: *International Workshop on Visual Analytics (EuroVA)*, pp. 13–16.
- May, T. et al. (2011b). “Guiding feature subset selection with an interactive visualization”. Em: *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 111–120.

- Milojević, S. et al. (2012). “Information visualization state of the art and future directions”. Em: *Proceedings of the American Society for Information Science and Technology* 49.1, pp. 1–3.
- Newman, D. et al. (1998). *UCI Repository of machine learning databases*.
- Ngai, E., L. Xiu e D. Chau (2009). “Application of data mining techniques in customer relationship management: A literature review and classification”. Em: *Expert Systems with Applications* 36.2, pp. 2592–2602.
- Nilsson, R. et al. (2007). “Consistent Feature Selection for Pattern Recognition in Polynomial Time”. Em: *The Journal of Machine Learning Research* 8, pp. 589–612.
- Pagliosa, P. et al. (2015). “Projection inspector: Assessment and synthesis of multidimensional projections”. Em: *Neurocomputing* 150, Part B.0, pp. 599–610.
- Paiva, J. G. S. et al. (2012). “Semi-Supervised Dimensionality Reduction based on Partial Least Squares for Visual Analysis of High Dimensional Data”. Em: *Computer Graphics Forum* 31.3pt4, pp. 1345–1354.
- Paulovich, F. V. e R. Minghim (2008). “HiPP: A Novel Hierarchical Point Placement Strategy and its Application to the Exploration of Document Collections”. Em: *Visualization and Computer Graphics, IEEE Transactions on* 14.6, pp. 1229–1236.
- Paulovich, F. V., C. T. Silva e L. G. Nonato (2010). “Two-Phase Mapping for Projecting Massive Data Sets”. Em: *IEEE Transactions on Visualization and Computer Graphics* 16.6, pp. 1281–1290.
- Pearson, K. (1901). “On lines and planes of closest fit to systems of points in space”. Em: *Philosophical Magazine* 2.6, pp. 559–572.
- Pekalska, E. et al. (1999). “A new method of generalizing Sammon mapping with application to algorithm speed-up”. Em: *Annual Conf. Advanced School for Comput. Imag.* Ed. por M. Boasson et al., pp. 221–228.
- Pudil, P. e J. Hovovicova (1998). “Novel methods for subset selection with respect to problem knowledge”. Em: *Intelligent Systems and their Applications, IEEE* 13.2, pp. 66–74.
- R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
- Ripley, B. D. e N. L. Hjort (1995). *Pattern Recognition and Neural Networks*. 1st. New York, NY, USA: Cambridge University Press.
- Sammon, J. W. (1969). “A Nonlinear Mapping for Data Structure Analysis”. Em: *IEEE Transactions on Computers* 18.5, pp. 401–409.
- Schreck, T. et al. (2008). “Visual cluster analysis of trajectory data with interactive Kohonen Maps”. Em: *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, pp. 3–10.

- Seo, J. e B. Shneiderman (2004). “A Rank-by-Feature Framework for Unsupervised Multi-dimensional Data Exploration Using Low Dimensional Projections”. Em: *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pp. 65–72.
- Silva, V. d. e J. B. Tenenbaum (2004). *Sparse multidimensional scaling using landmark points*. Rel. téc. Stanford.
- Tan, P.-N., M. Steinbach e V. Kumar (2005). *Introduction to Data Mining, (First Edition)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.
- Tejada, E., R. Minghim e L. G. Nonato (2003a). “On Improved Projection Techniques to Support Visual Exploration of Multidimensional Data Sets”. Em: *Information Visualization 2.4*, pp. 218–231.
- (2003b). “On improved projection techniques to support visual exploration of multidimensional data sets”. Em: *Information Visualization 2.4*, pp. 218–231.
- Torgerson, W. S. (1965). “Multidimensional scaling of similarity”. English. Em: *Psychometrika* 30.4, pp. 379–393.
- Turkay, C., P. Filzmoser e H. Hauser (2011). “Brushing Dimensions - A Dual Visual Analysis Model for High-Dimensional Data”. Em: *Visualization and Computer Graphics, IEEE Transactions on* 17.12, pp. 2591–2599.
- Wang, J. et al. (2003). “Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets”. Em: *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pp. 105–112.
- Williams, M. e T. Munzner (2004). “Steerable, Progressive Multidimensional Scaling”. Em: *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pp. 57–64.
- Wong, P. C. (1999). “Visual data mining”. Em: *Computer Graphics and Applications, IEEE* 19.5, pp. 20–21.
- Yang, J. et al. (2003). “Visual hierarchical dimension reduction for exploration of high dimensional datasets”. Em: *Proceedings of the symposium on Data visualisation 2003. VISSYM '03*. Grenoble, France: Eurographics Association, pp. 19–28.
- Yang, J. et al. (2007). “Value and Relation Display: Interactive Visual Exploration of Large Data Sets with Hundreds of Dimensions”. Em: *Visualization and Computer Graphics, IEEE Transactions on* 13.3, pp. 494–507.
- Yang, J., M. Ward e E. Rundensteiner (2002). “InterRing: an interactive tool for visually navigating and manipulating hierarchical structures”. Em: *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, pp. 77–84.
- Yang, J. et al. (2004). “Value and Relation Display for Interactive Exploration of High Dimensional Datasets”. Em: *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on*, pp. 73–80.

Zhang, J. e L. Gruenwald (2006). “Opening the Black Box of Feature Extraction: Incorporating Visualization into High-Dimensional Data Mining Processes”. Em: *Data Mining, 2006. ICDM '06. Sixth International Conference on*, pp. 1188–1192.