**Uma nova metáfora visual escalável para dados tabulares e
sua aplicação na análise de agrupamentos**

Evinton Antonio Córdoba Mosquera

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

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Evinton Antonio Córdoba Mosquera

Uma nova metáfora visual escalável para dados tabulares e sua aplicação na análise de agrupamentos

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. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Fernando Vieira Paulovich

USP – São Carlos
Dezembro de 2017

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

C796n Cordoba Mosquera, Evinton Antonio
 Uma nova metáfora visual escalável para dados
tabulares e sua aplicação na análise de agrupamentos
/ Evinton Antonio Cordoba Mosquera; orientador
Fernando Vieira Paulovich. -- São Carlos, 2017.
92 p.

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

1. Dados Tabulares. 2. Agregação de Dados. 3.
Visualização de Dados. 4. Mineração de dados. 5.
Análise visual de agrupamentos. I. Vieira
Paulovich, Fernando , orient. II. Título.

Evinton Antonio Córdoba Mosquera

**A scalable visual metaphor for tabular data and its
application on clustering analysis**

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. *FINAL VERSION*

Concentration Area: Computer Science and
Computational Mathematics

Advisor: Prof. Dr. Fernando Vieira Paulovich

**USP – São Carlos
December 2017**

*Este trabalho é dedicado a minha mãe Marisol e meu Pai Jose, também a todas as pessoas que estão em procura da felicidade, todas aquelas que dia a dia superam as adversidade para conseguir o sucesso,
Especialmente os pesquisadores, mergulhando-se no estudo e na busca da verdade.*

AGRADECIMENTOS

Agradeço primeiramente a Deus e a todas as energias que convergiram para o sucesso. Ao meu orientador, Professor Dr. Fernando Viera Paulovich, a orientação, ajuda e pela paciência com minha desorganização, prometo melhorar mais do que já tenho melhorado. Nunca vou esquecer esse caminho no qual andei, Obrigado.

Aos meus pais, Jose Cordoba e Marisol Mosquera, pelas palavras, pelo apoio que dão na minha vida e acreditar em mim. Obrigado Leidy Yohana, por me escutar e apoiar todos os dias durante o meu mestrado.

Aos meus amigos que se preocuparam por mim, pelas alegrias que me proporcionaram quando eu voltava à casa.

A todos os meus amigos que fiz aqui no Brasil. Especialmente no ICMC, Gladys, nunca vou esquecer de seu apoio, companheirismo quando precisei, muito obrigado. Filomen, colega e amigo de turma, foram muitas noites de estudo no laboratório, muito companheirismo e cooperação. Danilo e Tácito, obrigado pelas dicas e compartilhar palavras, tudo isso foi muito importante para eu poder me adaptar. Gabriel, muito obrigado pela ajuda na escrita, Paulo, Clausius, Fabio, Diego, Lucas, Nicolás, em fim a todos os colegas do laboratório de pesquisa em Visualização, Imagens e Computação. Muito Obrigado pela amizade, minhas palavras serão sempre de gratidão.

Respeito muito as suas coragem na busca do sucesso.

Ao instituto de Ciências Matemáticas e de Computação ICMC-USP, amigos de aulas, seus professore e funcionários.

Ao governo do Chocó-Colômbia e ao Departamento Administrativo de Ciência, Tecnologia e Inovação - COLCIENCIAS, processo - acordo n° 530/2014.

A todos que de alguma forma contribuíram para a realização deste trabalho. Mi casa es su casa!

“Os campeões não são feitos nas academias. Os campeões são feitos de algo que eles têm profundamente dentro de si - um desejo, um sonho, uma visão - ”
(Muhammad Ali)

RESUMO

CORDOBA, M. E. **Uma nova metáfora visual escalável para dados tabulares e sua aplicação na análise de agrupamentos**. 2017. 92 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

A rápida evolução dos recursos computacionais vem permitindo que grandes conjuntos de dados sejam armazenados e recuperados. No entanto, a exploração, compreensão e extração de informação útil ainda são um desafio. Com relação às ferramentas computacionais que visam tratar desse problema, a Visualização de Informação possibilita a análise de conjuntos de dados por meio de representações gráficas e a Mineração de Dados fornece processos automáticos para a descoberta e interpretação de padrões. Apesar da recente popularidade dos métodos de visualização de informação, um problema recorrente é a baixa escalabilidade visual quando se está analisando grandes conjuntos de dados, resultando em perda de contexto e desordem visual. Com intuito de representar grandes conjuntos de dados reduzindo a perda de informação relevante, o processo de agregação visual de dados vem sendo empregado. A agregação diminui a quantidade de dados a serem representados, preservando a distribuição e as tendências do conjunto de dados original. Quanto à mineração de dados, visualização de informação vêm se tornando ferramental essencial na interpretação dos modelos computacionais e resultados gerados, em especial das técnicas não-supervisionados, como as de agrupamento. Isso porque nessas técnicas, a única forma do usuário interagir com o processo de mineração é por meio de parametrização, limitando a inserção de conhecimento de domínio no processo de análise de dados. Nesta dissertação, propomos e desenvolvemos uma metáfora visual baseada na *TableLens* que emprega abordagens baseadas no conceito de agregação para criar representações mais escaláveis para a interpretação de dados tabulares. Como aplicação, empregamos a metáfora desenvolvida na análise de resultados de técnicas de agrupamento. O ferramental resultante não somente suporta análise de grandes bases de dados com reduzida perda de contexto, mas também fornece subsídios para entender como os atributos dos dados contribuem para a formação de agrupamentos em termos da coesão e separação dos grupos formados.

Palavras-chave: Dados tabulares. Agregação de dados. Visualização de dados. Mineração de dados. Análise visual. Análise de agrupamentos.

ABSTRACT

CORDOBA, M. E. **A scalable visual metaphor for tabular data and its application on clustering analysis**. 2017. 92 p. Dissertação (Mestrado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

The rapid evolution of computing resources has enabled large datasets to be stored and retrieved. However, exploring, understanding and extracting useful information is still a challenge. Among the computational tools to address this problem, information visualization techniques enable the data analysis employing the human visual ability by making a graphic representation of the data set, and data mining provides automatic processes for the discovery and interpretation of patterns. Despite the recent popularity of information visualization methods, a recurring problem is the low visual scalability when analyzing large data sets resulting in context loss and visual disorder. To represent large datasets reducing the loss of relevant information, the process of aggregation is being used. Aggregation decreases the amount of data to be represented, preserving the distribution and trends of the original dataset. Regarding data mining, information visualization has become an essential tool in the interpretation of computational models and generated results, especially of unsupervised techniques, such as clustering. This occurs because, in these techniques, the only way the user interacts with the mining process is through parameterization, limiting the insertion of domain knowledge in the process. In this thesis, we propose and develop the new visual metaphor based on the *TableLens* that employs approaches based on the concept of aggregation to create more scalable representations of tabular data. As application, we use the developed metaphor in the analysis of the results of clustering techniques. The resulting framework does not only support large database analysis but also provides insights into how data attributes contribute to clustering regarding cohesion and separation of the composed groups

Keywords: Scalable Tabular Data. Data Aggregation. Data Visualization. Data Mining. Visual Analytics. Clustering Analysis.

LISTA DE ILUSTRAÇÕES

Figura 1	– Processo de integração da Mineração de dados com a Visualização de Informação. Este processo exemplifica o modelo geral de <i>visual Analytics</i>	30
Figura 2	– Visualização de dados multivariados por ponto, (a) Matriz de <i>ScatterPlot</i> , (b) <i>RadViz</i>	32
Figura 3	– Visualização Multivariada Por Linha (a) Coordenadas Paralelas, (b) Coordenadas Paralelas Circulares.	34
Figura 4	– Visualização Multivariada Por Região (a) <i>Table Lens</i> , (b) <i>Survey Plot</i>	35
Figura 5	– Junção da <i>SolarPlot</i> e <i>Polar TreeMap</i> cria uma Visualização híbrida, que apresentam um conjunto de dados de correios recebidos numa empresa.	46
Figura 6	– Coordenadas paralelas hierárquicas com vários níveis de detalhe. 6a é a representação de um agrupamento, 6b apresenta vários agrupamentos agregados, cada um em uma faixa menor, 6c apresenta vários agrupamento e usa-se a cor para identificar a similaridade entre os grupos.	47
Figura 7	– Star Glyphs Hierárquica para 392 modelos de carros produzidos entre 1970 e 1990, (a) original <i>Star Glyphs</i> , (b) hierárquica <i>Star Glyphs</i>	48
Figura 8	– Matriz de ScatterPlots Hierárquica para 392 modelos de carros produzidos entre 1970 e 1990, (a) original Matriz de ScatterPlots, (b) Hierarquical Matriz de ScatterPlots.	48
Figura 9	– Hipp para coleção de 30.000 documentos de Reuters Corpus, dados projetados usando LSP.	49
Figura 10	– MIST para coleção de documentos das conferências de visualização de informação da IEEE de 1995 até 2002, dados projetados usando LSP.	50
Figura 11	– <i>XCluSim</i> ferramenta para avaliação de resultados de agrupamentos: (A) informação de parâmetros, (B) diagrama de força, (C) vista no dendrograma das similaridade dos algoritmos de agrupamento, (D) Similaridade de grupos em linhas paralelas, (E) visualização dos dados originais.	52
Figura 12	– <i>iGPSe</i> , ferramenta de análises visual para dados de câncer.	53
Figura 13	– Analítica visual de resultados de agrupamento usando coordenadas paralelas.	54
Figura 14	– Enfoque visual para o aperfeiçoamento do parâmetro k para algoritmos de agrupamento particionais.	56
Figura 15	– Avaliação e relação visual de algoritmos de agrupamento particionais aplicados a conjuntos de dados genéticos	57
Figura 16	– <i>Heat-Map</i> para avaliar estabilidade dos agrupamentos usando <i>Microarray Data</i>	57

Figura 17 – <i>HCE</i> , Exploração hierárquica para a análise de resultados de agrupamentos.	58
Figura 18 – <i>Vetorized RadViz</i> para análise de resultados de agrupamento, baseando-se no <i>Cluster Ensemble</i> , identifica os agrupamentos semelhantes.	59
Figura 19 – Detalhamento do processo de compressão. (A) Linhas adjacentes, (B) Média dos valores compactados, (C) Subtração entre a média e o valor máximo, (D) Valor máximo (E) Desvio padrão.	62
Figura 20 – TableLens para o conjunto de dados da Qualidade de Vinho Branco com 4898 instâncias e 11 atributos.	64
Figura 21 – TableLens compactada para o conjunto de dados de Qualidade de Vinho Branco , comprimindo 9 linhas.	65
Figura 22 – Estrutura da metodologia de integração. Da esquerda para a direita, (A) Entrada do conjunto de dados, (B) Gerar diferentes resultados de agrupamento, (C) Aplicar índices de validação de agrupamentos. (D) Interpretação visual da validade do agrupamento.	66
Figura 23 – TableLens compactada do conjunto de dados de Auto MPG comprimindo com $n = 9$. Em (a), representação das cores dos grupos. Em (b) o resultado da compactação e potencial para análise de um resultado de agrupamento.	66
Figura 24 – (A) Estabilidade de grupos usando cruzando linhas, (B) Estabilidade de grupos visualizados em layout tabular.	68
Figura 25 – Abordagem da Coordenadas Paralelas como método visual que auxilia ao usuário na interpretação de estabilidade e tendências dos grupos.	68
Figura 26 – TableLens na sessão de análise sem comprimir valores de dados usando o conjunto de dados de Qualidade de Vinho Branco	71
Figura 27 – TableLens na sessão de análise comprimindo $n = 9$ linhas por unidade visual, usando o conjunto de dados de qualidade do vinho (A). (B) Validação.	72
Figura 28 – TableLens na sessão de análise comprimindo $n = 20$ linhas por unidade visual, usando o conjunto de dados de qualidade do vinho (A). (B) Validação.	72
Figura 29 – TableLens na sessão de análise comprimindo $n = 40$ linhas por unidade, usando o conjunto de dados de Qualidade de Vinho Branco (A). (B) Validação.	73
Figura 32 – TableLens na sessão de análise de agrupamento sem comprimir valores de dados, usando o conjunto de dados de Auto MPG	73
Figura 33 – <i>TableLens</i> na sessão de análise comprimindo $n = 10$ linhas ordenada de forma descendente por <i>mpg</i> , usando o conjunto de dados de Auto MPG (A). (B) Validação.	73
Figura 30 – TableLens na sessão de análise comprimindo $n = 20$ linhas e ordenada pela coluna <i>DM_stdev</i> do conjunto de dados de HTRU2 (A). (B) Validação.	74
Figura 31 – TableLens na sessão de análise comprimindo $n = 40$ linhas e ordenada pela coluna <i>DM_stdev</i> do conjunto de dados de HTRU2 (A). (B) Validação.	74

Figura 34 – Ordenação da coluna <i>mpg</i> segundo os valores de silhueta, permitindo ao usuário entender a relação dos atributos que melhor ajustam-se ao conjunto de dados, usando <i>Auto MPG</i> com KMeans $k = 3$	77
Figura 35 – Ordenação da coluna <i>horsepower</i> segundo os valores de silhueta, permitindo ao usuário entender a relação dos atributos que melhor ajustam-se ao conjunto de dados, usando o <i>Auto MPG</i> com KMeans $k = 3$	77
Figura 36 – Ordenação da coluna <i>acceleration</i> segundo os valores de silhueta, permitindo ao usuário entender a relação dos atributos com respeito do grupo, usando o <i>Auto MPG</i> com KMeans $k = 3$	78
Figura 37 – Estabilidade de grupos para mostrar os dados mais relacionados através de diferentes agrupamentos, usando <i>KMeans</i> ($k = 3, 5, 7, 10$). (A) Grupos de dados estáveis no aumento consecutivo de k para diferentes processos de agrupamento usando conjunto de dados Auto MPG	79
Figura 38 – Aplicação de Coordenadas Paralelas para auxiliar na visão de estabilidade dos grupos e dados mais relacionados.	79
Figura 39 – Estabilidade de grupos para mostrar os dados mais relacionados através de diferentes agrupamentos, usando <i>KMeans</i> ($k = 3, 5, 7, 10$). (A) Grupos de dados estáveis no aumento consecutivo de k para diferentes processos de agrupamento usando o conjunto de dados de Qualidade de Vinho Branco	80
Figura 40 – Comparação dos agrupamentos obtidos por KMeans (KM) e Expectation-Maximization (EM) com $k = 3$ usando o conjunto de dados Auto MPG	81
Figura 41 – Coordenada Paralelas para auxiliar na comparação dos agrupamentos, conjunto de dados Auto MPG	81
Figura 42 – Comparação dos agrupamentos obtidos por <i>KMeans</i> (KM) e <i>Expectation-Maximization</i> (EM) com $k = 3$ usando o conjunto de dados Qualidade de Vinho Branco	82
Figura 43 – Coordenada Paralelas para auxiliar na comparação dos agrupamentos, conjunto de dados Qualidade de Vinho Branco	82

LISTA DE TABELAS

Tabela 1 – Características dos conjuntos de dados utilizados nos experimentos realizados. 70

LISTA DE ABREVIATURAS E SIGLAS

DB	Density Based Clustering
EM	Expectation Maximization
FDL	Force Direct Graph Layout
HC	Hierarchical Clustering
HDM	Histogram Distance Measure
KM	KMeans
LAMP	Local Affine Multidimensional Projection
LSP	Least Square Projection
MDS	Multidimensional Scaling
NNM	Nearest Neighbor Measure
PCA	Principal Components Analysis
PLMP	Partial Linear Multidimensional Projection
SOM	Self-Organization Map
TL	TableLens
XM	Extended K-means

SUMÁRIO

1	INTRODUÇÃO	25
1.1	Contextualização	25
1.2	Motivação e Objetivos	26
1.3	Organização	27
2	CONCEITOS RELACIONADOS	29
2.1	Considerações Iniciais	29
2.2	Analítica Visual (<i>Visual Analytics</i>)	29
2.2.1	<i>Visualização de Dados Multivariados</i>	30
2.2.1.1	<i>Visualização Multivariada por Ponto</i>	31
2.2.1.2	<i>Visualização Multivariada Por Linha</i>	33
2.2.1.3	<i>Visualização Multivariada Por Região</i>	34
2.2.1.4	<i>Discussão</i>	34
2.2.2	<i>Agregação</i>	35
2.3	Mineração de Dados	36
2.3.1	<i>Agrupamento</i>	37
2.3.2	<i>Tipos de Agrupamento</i>	39
2.3.3	<i>Avaliação Quantitativa de agrupamento</i>	41
2.4	Considerações Finais	43
3	TRABALHOS RELACIONADOS	45
3.1	Considerações Iniciais	45
3.2	Escalabilidade Visual	45
3.3	Avaliação Visual de Agrupamento	50
3.4	Considerações Finais	60
4	METÁFORA VISUAL ESCALÁVEL PARA DADOS TABULARES E APLICAÇÃO EM ANÁLISE DE AGRUPAMENTOS	61
4.1	Considerações Iniciais	61
4.2	Agregando linhas na <i>TableLens</i>	62
4.3	Analisando Agrupamentos	65
4.3.1	<i>Metodologia de Integração com a TableLens</i>	65
4.3.2	<i>Análise Visual de Estabilidade</i>	67

5	RESULTADOS E AVALIAÇÃO	69
5.1	Considerações Iniciais	69
5.2	Avaliação da Qualidade da Agregação	70
5.3	Análise de Grupos	75
6	CONCLUSÕES	83
6.1	Limitações e trabalhos futuros	84
	REFERÊNCIAS	85

INTRODUÇÃO

1.1 Contextualização

Nas últimas décadas, a grande quantidade de dados gerados e armazenados vem tornando o processo de transformação desses em informação útil cada vez menos trivial e mais dependente de métodos computacionais. Assim, a necessidade de melhorar e/ou desenvolver novos métodos para que seja possível interpretar esses dados, apresentando-os de forma simples e intuitiva ao usuário é cada vez mais evidente (KEIM *et al.*, 2006). Duas grandes áreas da computação estudam métodos para auxiliar na interpretação de dados, a *Mineração de Dados* e a *Visualização de Informação*. A Mineração de Dados é a atividade de extrair padrões de conjunto de dados por meio de modelos computacionais e a visualização de informação está focada em apoiar o processo de interpretação de dados por meio de representações visuais (CARD; MACKINLAY; SHNEIDERMAN, 1999). Como tais estratégias são complementares é natural a união de ambas em um processo conhecido como *Mineração Visual de Dados* (WONG, 1999), atualmente também chamado de *Visual Analytics* (Keim Daniel; MANSMANN; FLORIAN, 2010).

A mineração visual de dados aplicada à dados multidimensionais ou multivariados é um domínio desafiador. Considere um atributo como um campo que descreve uma instância de dados, se existirem k atributos descrevendo instâncias de um conjunto multivariado, cada instância pode ser interpretada como um vetor de k -dimensões em um espaço k -dimensional. Muitas técnicas de visualização para dados multivariados têm sido propostas, sendo que a maioria mapeia os dados para pontos, linhas, regiões, ícones ou outro tipo de ícone gráfico (WARD; GRINSTEIN; KEIM, 2015). Alguns exemplos de técnicas são: *Scatterplots* (CLEVELAND, 1993), *Matrizes de ScatterPlot*, *TableLens* (RAO; CARD, 1994a), *Coordenadas Paralelas* (INSELBERG; DIMSDALE, 1990a), *Técnicas Orientadas a Pixels* (KEIM, 1996; KEIM, 2000), *RadViz* (HOFFMAN *et al.*, 1997), *Survey Plots* (LOHNINGER, 1994).

Dentre os possíveis processo de mineração apoiados pela visualização multidimensional,

um dos mais importantes é o processo não-supervisionado de agrupamento. As técnicas de agrupamento visam dividir um conjunto de dados em grupos de forma que dados semelhantes entre si permaneçam no mesmo grupo, enquanto os dados não semelhantes sejam agrupados em diferentes grupos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009a; TAN; STEINBACH; KUMAR, 2005). Os métodos de agrupamento são dependentes de informação *a priori*, como por exemplo: o número de grupos ou densidade máximo/mínimo de cada grupo. Essa informação é responsabilidade do analista, sendo que o sucesso do processo de agrupamento está intimamente ligado a parametrização empregada nas técnicas com base nessa informação.

1.2 Motivação e Objetivos

Apesar das técnicas mencionadas anteriormente serem normalmente empregadas com sucesso para pequenos conjuntos de dados, para grandes conjuntos elas tendem a apresentar problemas de sobreposição, desordem e perda do contexto visual. Assim, surge a necessidade de se conseguir definir abordagens mais compactas que possam apresentar grandes quantidades de dados em espaços visuais reduzidos, evitando perda de contexto ou visão geral. A abordagem que têm como intuito melhorar interpretabilidade e escalabilidade visual das técnicas de visualização é denominada Agregação. Agregação é o processo de reduzir a quantidade de dados por meio da derivação de dados que representem intrinsecamente o conjunto de dados original (MUNZNER; MAGUIRE, 2015).

Além desses desafios de interpretabilidade e escalabilidade visual, no que se refere a tarefas específicas de mineração, outro grande desafio envolve a interpretabilidade de modelos computacionais não-supervisionados. De forma geral, podemos dizer que se resultados gerados por processos não-supervisionados não refletirem o ponto de vista do usuário ou não forem de fácil interpretação, as tarefas de mineração tenderão a não ser satisfatórias. Isso porque, diferente das tarefas supervisionadas, como classificação e regressão, para tarefas não-supervisionadas, como as de agrupamentos, não existe uma forma definitiva de se avaliar os resultados, tornando o processo de análise demorado e dependente do ponto de vista do analista (ESTIVILL-CASTRO, 2002). Portanto, para que tarefas de mineração não-supervisionada, em especial as tarefas de agrupamento, sejam bem sucedidas, ferramental que dê suporte a interpretação dos modelos computacionais e dos resultados gerados é essencial.

Nesse contexto, o objetivo deste projeto de mestrado foi empregar modelos de agregação para aumentar a escalabilidade visual de uma técnica de visualização multivariada existente, a *TableLens* (RAO; CARD, 1994b), aplicando a nova metáfora proposta para a análise de agrupamento de dados multivariados, visando entender os resultados obtidos pelos algoritmos de agrupamento. Por meio de métricas de avaliação de agrupamento e metáforas de interação usuário-computador, nosso intuito foi possibilitar uma análise que envolve o entendimento de como cada atributo contribui para a coesão e separação dos grupos formados, além de fornecer

subsídios para a análise da estabilidade desses grupos considerando diferentes processos de agrupamento.

1.3 Organização

O restante dessa dissertação está estruturado da seguinte maneira:

- No Capítulo 2, conceitos básicos sobre técnicas de visualização de dados multivariados e mineração de dados são apresentados;
- No Capítulo 3 é feita uma revisão sobre os principais trabalhos relacionados a escalabilidade visual e à análise visual de agrupamentos;
- No Capítulo 4, apresenta-se a abordagem de escalabilidade visual da *TableLens* voltada à análise de agrupamentos;
- No Capítulo 5, a abordagem proposta é avaliada e os resultados alcançados são apresentados;
- Por fim, no Capítulo 6 apresenta-se as conclusões deste projeto de mestrado.

As opiniões, hipóteses e conclusões ou recomendação expressas neste material são responsabilidade do autor.

CONCEITOS RELACIONADOS

2.1 Considerações Iniciais

Nas seguintes seções apresentam-se conceitos relacionados a Visualização de Informação e Mineração de Dados. Na Seção 2.2 é apresentado o escopo de Analítica Visual. A Seção 2.2.1 apresenta brevemente o que é Visualização de Informação e Visualização de Dados Multivariados. Na Seção 2.3 são mostradas as ideias por trás do processo de mineração de dados, focando-se em agrupamento.

2.2 Analítica Visual (*Visual Analytics*)

A importância de entender, descobrir informações e inferir conhecimento a partir de dados tem evoluído continuamente, com o surgimento de novas formas de análise usando métodos visuais mais interativos que auxiliam na descoberta de informação útil. Abordagens de várias áreas vem sendo unidas nesse processo, tais como mineração de dados, aprendizado de máquina, visualização de informação e interação usuário-computador, resultando em uma área conhecida como Analítica Visual (*Visual Analytics*) (THOMAS; COOK, 2005a; Keim Daniel; MANSMANN; FLORIAN, 2010).

De forma geral, visa-se incluir o usuário no processo para direcionar tarefas de visualização, por meio de interação visual, a experiência do analista de dados é adicionada no processo de Mineração de Dados.

Levando em conta isso, os desafios da analítica visual são identificar os melhores algoritmos de automatização de análise de dados, aprimorar parâmetros de entrada, aperfeiçoar os métodos de interação e apresentação visual dos dados etc. Como mostrado na Figura 1 e seguindo as ideias de Silva, Xavier e Magalhães (2010), Santos *et al.* (2011), Paiva (2012), a inclusão do usuário no processo de Mineração Visual de Dados combina mediante a iteratividade

Figura 1 – Processo de integração da Mineração de dados com a Visualização de Informação. Este processo exemplifica o modelo geral de *visual Analytics*.

Fonte: Adaptada de Keim *et al.* (2008).

entre os modelos computacionais e as visualizações a flexibilidade e criatividade humana para extrair informação e gerar conhecimento. Os modelos e dados podem ser refinado através da visualização e interação do usuário. Esse enfoque é mais conhecido como **Relevance Feedback**. Embora **Relevance Feedback** seja um conceito usado mais frequentemente em sistemas de classificação, é possível aplicá-lo ao processo de agrupamento visual de dados, com o intuito de gerar um processo de mineração iterativo que se ajuste às necessidades do usuário.

Nas seções seguintes, são apresentados conceitos relevantes ao escopo deste trabalho, relacionados a Visualização de Informação para dados multivariados. No que refere-se a mineração de dados, apresenta-se conceitos relacionados a Agrupamento de Dados, também conhecido como análise não supervisionada ou descritiva.

2.2.1 Visualização de Dados Multivariados

A Visualização de informação (InfoVis) é uma área de pesquisa que visa auxiliar o usuário na tarefa de entender padrões, correlações, estrutura e tendências que existem em conjuntos de dados. Entre os desafios encontrados na área de visualização de informação, pode-se citar a apresentação de conjuntos de dados de alta dimensionalidade, sujeitos a sobreposição, oclusão ou qualquer outro tipo de dificuldade que impeça a análise correta dos dados.

No contexto de análise de grandes conjuntos de dados é necessária a criação de metáforas visuais que permitam lidar com um grande número de instâncias ou objetos. Dados consistem geralmente de um determinado número de registros, cada um deles vinculado a uma série de variáveis ou dimensões, que definem a dimensionalidade do conjunto de dados.

As instâncias podem ser interpretadas como um vetor de k -dimensões. Por exemplo, considerando dados de pacientes de um hospital, uma instância pode ser descrita por variáveis como idade, sexo, doença e assim por diante. Em análise estatística tradicional, instâncias de dados com quatro ou mais dimensões são conhecidas como multivariadas ou hiper-variadas (VIERA, 2008). Assim, por definição, $A^k = a_1^k, \dots, a_n^k$ determina um vetor para cada dimensão do conjunto de dados, onde A^k representa o rótulo no atributo k , a_n^k representa a k -ésima dimensão da instância n .

Sendo assim, técnicas de visualização para dados multivariados tem sido propostas e classificadas segundo o elemento gráfico mapeado. De acordo com Ward, Grinstein e Keim (2015) os elementos gráficos correspondentes podem ser representados por: Ponto, Linha e Região.

2.2.1.1 Visualização Multivariada por Ponto

Visualizações baseadas em pontos são projeções de instâncias k -dimensionais em um espaço visual p -dimensional com $p=\{1,2,3\}$. Entre os exemplos clássicos pode-se citar a Matriz de *Scatterplots* e a *RadViz*, descritos abaixo.

A Matriz de *Scatterplots* (UTTS, 2004) consiste em uma grade de *Scatterplots* com $N \times N$ células, onde $N = k$ é o número de atributos do conjunto de dados. Nesta visualização são apresentadas todas as combinações possíveis entre pares de atributos, com cada atributo exibido no eixo horizontal e vertical - visualizações apresentadas na diagonal representam a simetria entre atributos. A Figura 2a ilustra a matriz de *Scatterplots* para o conjunto de dados **OUT5d** de sensoriamento remoto com cinco dimensões. Nesta visualização ilustra-se claramente um problema dos *ScatterPlots* para grandes conjuntos de dados: a visualização apresenta sobreposição de dados, impedindo a interpretação adequada dos dados representados. A *RadViz* (HOFFMAN *et al.*, 1997) é uma técnica de visualização desenvolvida para mapear dados multidimensionais no espaço visual inspirada por um sistema massa-mola. Cada atributo (coordenada) é inicialmente normalizado e associado a uma Âncora Dimensional (AD). Assim, k ADs são necessárias para representar dados k -dimensionais. Estas âncoras dimensionais são posicionadas sobre a circunferência de um círculo. Cada instância n é atraída na direção das diferentes ADs, com um valor de atração correspondente aos valores de n em cada instância em cada dimensão. A Figura 2b representa a *RadViz* para o conjunto de dados Iris. É possível ver a distribuição dos dados em relação aos atributos.

No contexto de apresentar grandes conjuntos de dados, uma abordagem alternativa que tem sido aplicada com sucesso para conjuntos de dados com alta dimensionalidade são as técnicas de projeção de dados multidimensionais. Uma técnica de projeção multidimensional mapeia as instâncias de dados em um espaço uni, bi ou tridimensional, preservando alguma informação sobre as relações de distância ou similaridade contidas nos dados de forma a revelar ao máximo possível as estruturas e padrões existentes (TEJADA; MINGHIM, 2005; PAULOVICH, 2008).

Figura 2 – Visualização de dados multivariados por ponto, (a) Matriz de *ScatterPlot*, (b) RadViz.

Fonte: Cui *et al.* (2006), Ono (2015).

Na literatura, as técnicas de projeção multidimensional são divididas em quatro grupos (MAMANI, 2012; PAULOVICH *et al.*, 2011):

- Decomposição espectral: as técnicas de decomposição espectral calculam coordenadas cartesianas para cada instância de dados no espaço visual realizando uma decomposição espectral da matriz de similaridade entre todos os pares de instâncias de dados. A mais conhecida é a *Classical Scaling* (MDS) (TORGERSON, 1965), algumas derivações da MDS são: *Landmarks* MDS (SILVA; TENENBAUM, 2004), ISOMAP (TENENBAUM; SILVA; LANGFORD, 2000), FastMap (FALOUTSOS; LIN, 1995).
- Minimização não linear: as técnicas de minimização não linear, definem uma função de perda com base na diferença entre as distâncias no espaço de características e as instâncias no espaço visual, minimizando tal função utilizando o método de gradiente descendente para encontrar um mínimo. Um exemplo de otimização não linear é a técnica *Sammon's Mapping* (SAMMON, 1969).
- Modelos baseados em força: para as técnicas de modelos baseados em força, cada instância é modelada como uma partícula sobre o efeito de força de atração ou repulsão em relação às demais partículas. Essas técnicas empregam modelos de molas tentando chegar a um sistema em estado de equilíbrio: os objetos são posicionados inicialmente de forma aleatória ou usando alguma heurística, e então as forças geradas pelas molas são aplicadas até atingir uma posição de equilíbrio. As forças no sistema são proporcionais à diferença entre as dissimilaridade entre objetos $\delta(x_i, x_j)$ e a distância entre os pontos no *layout* $d(y_i, y_j)$. Algumas abordagens propostas nesse sentido são: Chalmers (CHALMERS, 1996), *Force*

Scheme (TEJADA; MINGHIM; NONATO, 2003), *Least Square Projection* (LSP) (PAULOVICH *et al.*, 2008), *Piecewise Laplacian-based Projection* (PLP) (PAULOVICH *et al.*, 2011), *Glimmer* (INGRAM; MUNZNER; OLANO, 2009).

- Operador linear: as técnicas de operador linear buscam encontrar uma representação de menor dimensão, usando-se critérios de redução e preservação de relações de distâncias. Algumas abordagens desenvolvidas nesse contexto são a *Random Projection* (BINGHAM; MANNILA, 2001), que utiliza uma matriz cujas colunas tem tamanho unitário preservando a relação de similaridade entre os dados. A *Part Linear Multidimensional Projection* (PLMP) (PAULOVICH; SILVA; NONATO, 2010), que substitui a matriz empregada na *Random Projection* por uma construída levando-se em consideração relações de distâncias - a PLMP primeiro projeta uma amostra de dados, para então mapear linearmente as instâncias restantes, visando preservar a estrutura de relações entre amostras. A *Principal Component Analysis* (PCA) (JOLLIFFE, 1986) transforma um conjunto de variáveis correlacionadas em um conjunto menor de variáveis com a finalidade de reter tanto quanto possível a variação dos dados no conjunto empregado, em tal grau que se a métrica de distância no espaço multidimensional é Euclideana, o resultado alcançado pela PCA é o mesmo da *Classical Scaling*. Por fim, a *Local Affine Multidimensional Projection* (LAMP) (JOIA *et al.*, 2011) usa uma amostra como ponto de controle e constrói um mapeamento ortogonal que projeta os dados no espaço visual. Para cada ponto a ser projetado é construído um mapeamento ortogonal afim que projeta os dados do espaço multidimensional para o visual 2D ou 3D, sendo que o mapeamento segue o *layout* com referência aos pontos de controle.

2.2.1.2 Visualização Multivariada Por Linha

Este modelo de visualização caracteriza-se por utilizar linhas cujos padrões podem ser percebidos por meio de cruzamento ou curvatura. Como exemplo mais tradicional: coordenadas paralelas (INSELBERG; DIMSDALE, 1990a), visualização em que os atributos são representados em eixos paralelos e as instâncias são polilinhas que cruzam e conectam esses eixos em uma posição proporcional do intervalo entre o menor e maior valor de cada atributo. A Figura 3a ilustra a técnica de coordenadas para o conjunto de dados com especificações de 392 modelos de carros produzidos entre 1970 e 1990. Outra abordagem são as Curvas de Andrews, que apresentam os dados em forma de curva seguindo o padrão de transformadas de Fourier, baseando-se em termos das funções senoidal e cossenoidal para os dados (CHAN, 2006). As Coordenadas Paralelas Circulares estendem o conceito de coordenadas paralelas mapeando os dados para eixos radiais (Figura 3b). A ordenação destes eixos-dimensões costuma ser um problema comum a essas visualizações, assim como conjuntos de dados com muitas instâncias e atributos geram problemas de sobreposição.

Figura 3 – Visualização Multivariada Por Linha (a) Coordenadas Paralelas, (b) Coordenadas Paralelas Circulares.

(a) Coordenadas Paralelas, para 392 modelos de carros produzidos entre 1970 e 1990.

(b) Coordenada Paralelas Circulares para o conjunto de dados da Iris.

Fonte: Ward, Grinstein e Keim (2015), Hoffman (2000).

2.2.1.3 Visualização Multivariada Por Região

Visualizações baseadas em região apresentam os dados empregando polígonos, linhas mapeadas para tamanho, cor, forma e textura. Alguns exemplos mais clássicos: Histogramas, gráficos de barra, *HeatMaps*, (LOHNINGER, 1994), *Survey Plots*, *Table Lens* (RAO; CARD, 1994a; TENEV; RAO, 1997). A Figura 4a ilustra a *Table Lens* para o conjunto de dados de vinhos com 12 dimensões. Esta técnica se baseia no modelo de planilhas eletrônicas: para o mapeamento de dados multivariados em uma tabela, os atributos são representados por colunas, as linhas representam instâncias com uma altura de um pixel e cada célula é apresentada por uma barra horizontal de tamanho proporcional ao valor do atributo. Outro exemplo é a Figura 4b, que apresenta *Survey Plots* do (LOHNINGER, 1994). A área da visualização é dividida em k colunas retangulares que representam os atributos, cada linha horizontal do retângulo representando uma instância, cujo tamanho varia de acordo com o valor.

2.2.1.4 Discussão

Os trabalhos de visualização de dados multivariados descritos apresentam diferentes representações gráficas. No entanto, tais técnicas podem ser empregadas com sucesso somente em pequenos conjuntos de dados. Para conjuntos de dados maiores ou com alta dimensionalidade, elas tendem a apresentar problemas de sobreposição ou oclusão, não conseguindo identificar satisfatoriamente grupos de dados correlacionados ou determinar anomalias ou tendências (LU *et al.*, 2012; VIERA, 2008). Uma abordagem que tem sido combinada com as técnicas de visualização para dados multivariados é a agregação. Um método de agregação lida com

Figura 4 – Visualização Multivariada Por Região (a) *Table Lens*, (b) *Survey Plot*.

Fonte: Elaborada pelo autor.

Fonte: Demšar *et al.* (2013).

problemas de sobreposição e oclusão possibilitando uma visão geral do conjunto de dados sem alterar o modelo de visualização.

2.2.2 Agregação

Uma forma de reduzir a quantidade de dados apresentados e facilitar a análise de grandes conjuntos de dados na visualização está em usar métodos de agregação. Esses métodos envolvem tipicamente agrupamento de dados baseado na similaridade, filtragem e amostragem. Por definição, a agregação é a apresentação de dados derivados que representam intrinsecamente um grupo ou conjunto de dados (MUNZNER; MAGUIRE, 2015; WARD; GRINSTEIN; KEIM, 2015). Munzner e Maguire (2015), Ward, Grinstein e Keim (2015) classificaram agregação da seguinte forma:

Agregação de Dados: a agregação de dados proporciona uma visão global de um conjunto de dados através do uso de dados derivados que são calculados a partir dos dados originais. Estes dados derivados podem ser calculados usando informação descritiva - métodos estatísticos descritivos.

Agregação Espacial: O principal objetivo da agregação espacial é representar a natureza dos dados espaciais considerando mudanças no tamanho da visualização ou posição dos dados.

Agregação de Atributo: a agregação de atributos também pode ser entendida como redução de dimensionalidade. Esta é a abordagem mais complexa de agregação, cujo objetivo é preservar a estrutura significativa de um conjunto de dados utilizando menos atributos para representá-lo. Portanto, é definida como o processo de projeção de dados de alta dimensionalidade para um espaço de menor dimensão 2D ou 3D.

A representação incorporada nos dados derivados apresentados nos trabalhos de agregação ([ELMQVIST; FEKETE, 2010](#); [WARD; GRINSTEIN; KEIM, 2015](#)) pode ser gerada de diferentes maneiras:

- Contagem: número de dados no grupo;
- Média: valor que aponta para onde mais se concentram os dados no grupo;
- Mediana: valor numérico que separa a metade superior da metade inferior de uma amostra de dados no grupo;
- Moda: valor que surge com mais frequência no grupo;
- Amplitude: O valor mínimo e máximo no grupo;
- Percentis: medidas que dividem a amostra no grupo;
- Distribuição: medidas de distribuição que podem ser apresentadas com histogramas, gráficos de barra, etc.

Para alguns casos os dados de agregação são derivados calculando a informação descritiva utilizando métodos estatísticos e podem ser representados por tamanho, textura, cor, etc. Conforme [Shneiderman \(1996\)](#), no contexto de explorar dados agregados, o *Seeking Mantra* (Visão geral, Filtro e Zoom, Detalhe) tem demonstrado ser útil na intervenção do usuário no processo de visualização de informação.

Nesta seção, foram apresentadas técnicas e metáforas visuais comumente utilizadas para exploração de conjuntos de dados. Na seção seguinte, são apresentados conceitos de mineração de dados e algoritmos de agrupamento.

2.3 Mineração de Dados

A mineração de dados é o processo de descoberta automática de informações úteis ou extração de relações ocultas em grandes depósitos de dados. A mineração é parte integral do processo **KDD** *Knowledge Discovery in Database*, que envolve a conversão de dados brutos em informações úteis desde o pré-processamento dos dados (transformação dos dados) até pós-processamento (exploração de dados) ([TAN; STEINBACH; KUMAR, 2005](#)). Assim como ocorre com a visualização de dados, na mineração de dados enfrenta-se desafios como escalabilidade, dimensionalidade e análise de dados heterogêneos.

Segundo o apresentado por diversos pesquisadores ([Faceli et al. \(2011\)](#), [Hastie, Tibshirani e Friedman \(2009b\)](#), [Tan, Steinbach e Kumar \(2005\)](#)) a mineração de dados geralmente pode ser dividida em diferentes tipos de tarefas: Tarefas de Previsão, também conhecidas como

Supervisionadas, e Tarefas Descritivas, ou Não Supervisionadas. O objetivo da Tarefa de Previsão é prever o valor de uma determinada instância baseando-se nos valores de outras instâncias. Existem duas formas de modelagem de previsão: Classificação e Regressão. Nas Tarefas Descritivas, o objetivo é encontrar relações, anomalias, grupos, tendências ou trajetórias, o que implica derivar padrões de tendência que resumam os relacionamentos nos dados. Geralmente as tarefas descritivas são exploratórias e requerem pós-processamento ou visualização para explicar os resultados. As tarefas descritivas mais conhecidas são: agrupamento, análise de associação e detecção de anomalias. O agrupamento é o processo de colocar dados similares em grupos; A análise de associação é mais usada para descobrir padrões que descrevam variáveis ou características associadas dentro dos dados; a detecção de anomalias pode simplificar a tarefa de identificar instâncias cujas características sejam diferentes dos resto do conjunto de dados, também conhecidas como fatores estranhos.

A seguir, são apresentados com mais detalhes conceitos de mineração de dados relevantes ao escopo deste projeto, referentes a agrupamento, como técnicas, métricas de avaliação, desafios para os algoritmos de agrupamentos, etc. A mineração de dados ainda é uma área relativamente nova e explorável. Para aprofundar sobre estudo da mineração de dados, pode-se usar como informação de referencia (FACELI *et al.*, 2011; HASTIE; TIBSHIRANI; FRIEDMAN, 2009b; TAN; STEINBACH; KUMAR, 2005).

2.3.1 Agrupamento

Algoritmos de agrupamento são em geral usados para analisar dados que não requerem uma variável alvo, com o objetivo de encontrar a similaridade entre instâncias. A análise de agrupamento também é denominada segmentação de dados, na qual as instâncias são separadas em grupos intra-relacionados e inter-relacionados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009b). Na literatura os algoritmos de agrupamento tem sido propostos e utilizados para diferentes tipos de conjuntos de dados e objetivos de pesquisa. Na maioria das aplicações, assume-se que os dados não fornecem tendências e tem uma estrutura aleatória. Isto implica que, durante a execução, dados com maior similaridade fiquem no mesmo grupo, separando-se dos dados espúrios e não similares (JAIN; MURTY; FLYNN, 1999).

Na aplicação das técnicas de agrupamento, a forma como a similaridade ou dissimilaridade entre os objetos é calculada pode desempenhar um papel central. Por isso, todos os algoritmos de agrupamento recebem como entrada uma matriz de similaridade para agrupar os dados. Assim, o primeiro passo para uma aplicação efetiva desses algoritmos é entender o que é mensurado como a distância $D(x_i, x_j)$. A distância pode ser descrita como similaridade ou dissimilaridade. Um problema importante no cálculo de distâncias é que nem sempre os atributos em um conjunto de dados são homogêneos, existindo dois tipos de cenários que podem distorcer os resultados de similaridade: os vetores possuem diferentes normas Euclidianas, ou os atributos das instâncias possuem uma escala diferente, gerando um maior peso na função de

distância. Para evitar estes cenários, existem processos de **Normalização ou Padronização** que podem ser aplicados para obter melhores resultados. Tanto o processo de normalização quanto a padronização tem como objetivo fazer o conjunto de valores de um atributo ter uma determinada propriedade. A **Normalização**, que pode ser aplicada ao primeiro cenário, faz com que todos os dados da variável tenham um valor normal ou Gaussiano, isto é $\|x_i\| = \sum_{k=1}^n x_{ik}^2 = 1$. Assim, cada dado i é dividido pela normal, $x'_{ji} = \frac{x_{ji}}{\|x_i\|}$. Já o caso do segundo cenário pode ser evitado aplicando-se a **Padronização**. Se a media dos valores de um atributo é $\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$ e seu desvio padrão $\sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$, portanto, a transformação para os dados é $x'_{ij} = \frac{(x_{ij} - \bar{x}_j)}{\sigma_j}$ para $1 \leq i \leq n$ e $1 \leq j \leq m$, gerando assim uma nova variável que tem uma média 0 e um desvio padrão de 1.

No que diz respeito a dissimilaridade, dentre os diferentes tipos de funções de distância existentes a distância de *Minkowski* é a de maior abrangência. Em realidade, ela cria a generalização de métricas de distancias D , onde r é um parâmetro que corresponde às dimensões. A Equação 2.1 mostra esta definição.

$$d_r(x_i, x_j) = \left(\sum_{k=1}^n |x_{ik} - x_{jk}|^r \right)^{1/r} \quad (2.1)$$

Com $r = 1$ obtém-se a distância *Manhattan* (*City Block*), com $p = 2$ obtém-se a distância Euclidiana, com $p = \infty$ tem-se a distancia suprema também conhecida como L_{max} ou norma ∞ , mais conhecida como $L_\infty(x_i, x_j) = |x_{ik} - x_{jk}|$. Para ser classificada como métricas de distância, uma função de distância deve obedecer três propriedades (TAN; STEINBACH; KUMAR, 2005): Positividade, Simetria e Diferença Triangular.

- **Positividade:** $\forall x_i, x_j \in X : d(x_i, x_j) > 0$
- **Simetria:** $\forall x_i, x_j \in X : d(x_i, x_j) = d(x_j, x_i)$
- **Diferença Triangular:** $\forall x_i, x_j, x_z \in X : d(x_i, x_j) \leq d(x_i, x_z) + d(x_z, x_j)$

Uma relação de similaridade refere-se ao inverso da dissimilaridade, sendo uma medida de quão semelhantes são as instâncias. Dos exemplos gerais de distância, pode-se citar: $s(x_i, x_j) = \frac{1}{d(x_i, x_j) + 1}$ e $s(x_i, x_j) = e^{-d(x_i, x_j)}$.

Os métodos descritos são um resumo da possibilidade de ajuste para um domínio geral. Em cada domínio de aplicação é preferível o emprego de métricas e métodos de transformação próprios, dependendo das características do conjunto a ser mensurado. Uma discussão mais detalhada sobre essas métricas pode ser encontrada em (WARD; GRINSTEIN; KEIM, 2015; ESTIVILL-CASTRO, 2002; TAN; STEINBACH; KUMAR, 2005; HASTIE; TIBSHIRANI; FRIEDMAN, 2009b).

O paradigma de avaliar os grupos ou divisões em estruturas de dados é conhecido como análise de agrupamento. A análise de agrupamento tem como objetivo minimizar a distância entre as instâncias de dados dentro de um grupo e maximizar a distância entre diferentes grupos. Além disso, a análise de agrupamento se relaciona com a classificação pelo fato de gerar rótulos de classe para os objetos de acordo com os grupos atribuídos, embora os rótulos sejam obtidos baseando-se puramente nos dados e não em modelos construídos a partir de variáveis alvos.

Existem diversos critérios para classificar algoritmos de agrupamento. [Aggarwal e Reddy \(2013\)](#), [Arora e Chana \(2014\)](#) afirmam que os algoritmos de agrupamento devem ser classificados em Hierárquicos, Particionais e Baseados em Densidade. [Tan, Steinbach e Kumar \(2005\)](#) discordam, propondo classificá-los baseando-se em como os grupos se comportam: Bem separados, Baseados em Protótipos, Baseados em Grafos, Baseados em Densidade e Baseados em Grupos Conceituais. A seguir são estes conceitos são apresentados em mais detalhes.

2.3.2 Tipos de Agrupamento

Em alguns cenários de classificação de algoritmos de agrupamento, é possível que haja distinção entre o conceito e a forma como o algoritmo de fato gera grupos.

No que diz respeito à forma como grupos são organizados, pode-se classificar algoritmos de agrupamento da seguinte forma:

- Agrupamentos **Bem Separados**: refere-se a um conjunto de dados no qual cada objeto está mais próximo a cada um dos outros objetos no grupo do que qualquer outro objeto que não esteja nesse grupo.
- Agrupamentos **Baseados em Protótipos**: os objetos estão mais próximos a um protótipo do grupo; O protótipo pode ser definido como a média, mediana, moda ou dado central.
- Agrupamentos **Baseados em Grafos**: os grupos podem ser definidos como dados ou componentes conectados, isto é, um grupo está conectado entre si e não possui conexão com objetos externos. Os algoritmos baseados em grafos são particularmente suscetíveis a ruído e dados anômalos.
- Agrupamentos **Baseados em Densidade**: refere-se a grupos que são criados em uma região densa de objetos, rodeada por uma região de baixa densidade. Os algoritmos baseados em densidade são capazes de lidar com ruídos e detectar anomalias, uma vez que esses dados são as pontes entre os grupos.
- **Grupos Conceituais**: define-se como grupos que compartilham objetos.

No que diz respeito à forma como os agrupamentos são gerados, pode-se classificá-los como: **Particionais, Hierárquicos e Baseados em Densidade**.

- **Algoritmos de Agrupamento Particional:** algoritmos que simplesmente geram uma estrutura de divisão do conjunto de dados em subconjuntos não interseccionados são conhecidos como particionais. Existem diferentes técnicas que usam essa abordagem, a maioria baseada em dois algoritmos mais usados *K-Means* e *K-Medoid*. O algoritmo *K-Means* gera a partição por meio de protótipos ou centroides que são geralmente dados pela média do grupo. O algoritmo *K-Medoid* define protótipos em termos de medóide, que são os dados mais representativos para o grupo ou faixa de dados. As técnicas de agrupamento particionais dependem do parâmetro k , que define a quantidade de grupos para se gerar. Primeiramente, o usuário fornece os k centroides iniciais, sendo então cada objeto atribuído ao centroide mais próximo, cada atribuição gerando uma coleção de objetos (grupo). O centroide de cada grupo é atualizado com a média dos objetos atuais do grupo. Repete-se os passos de atribuição de objetos e atualização de centroides até que nenhum objeto mude de grupo. A seção de resultados oferece uma descrição mais detalhada desses algoritmos.
- **Algoritmos de Agrupamento Hierárquico:** Os agrupamentos hierárquicos podem ser vistos como a sequência dos agrupamentos particionais. A estrutura do agrupamento hierárquico permite aos grupos ter subconjuntos (subgrupos), gerando um conjunto de grupos aninhados como árvore onde cada nó (grupo) na árvore é uma união dos filhos (subgrupo). A raiz da árvore é o grupo contendo todo o conjunto de dados. Os algoritmos de agrupamento hierárquico podem ser gerados usando uma abordagem **aglomerativa** ou **divisiva**. A primeira abordagem começa com grupos representando dados individuais, e em cada etapa, une grupos mais próximos. A abordagem **divisiva** começa com todo o conjunto de dados, e a cada etapa divide os dados em grupos distintos até que restem apenas grupos únicos de pontos individuais. A construção de tais abordagens tanto para aglomerar quanto para dividir depende da relação de distâncias utilizada. Exemplos de relações são *Única*, *Completa* ou *Média*. A conexão única define proximidade de grupos como a menor distância entre dois dados que estejam em diferentes grupos. A conexão completa é definida pela maior distância entre dados de diferentes grupos. A conexão média define a proximidade como a média das distâncias entre pares para todas as instâncias de dois grupos.
- **Algoritmos de Agrupamento Baseado em Densidade:** algoritmos baseados em densidade localizam regiões de alta densidade que estejam separadas entre si por regiões de baixa densidade. Embora os algoritmos de densidade não sejam baseados em métricas de similaridade, a densidade é avaliada para um determinado ponto no conjunto de dados contando-se o número de pontos dentro de um determinado raio, *Eps*, em aquele ponto. Para grupos nesta abordagem, o raio *Eps* é normalmente especificado pelo usuário. Para cada grupo, algoritmos baseados em densidade interpretam pontos de acordo com sua função no agrupamento. **Pontos Centrais:** são pontos no interior de um grupo, que possuem

um dado número mínimo de outros pontos dentro uma determinada distância (vizinhança) de si. **Ponto de Limite:** um ponto limite fica na vizinhança de um ponto central, mas não possui vizinhos suficientes para ser classificado como tal. **Ponto de Ruído:** qualquer ponto que não seja nem ponto central nem de limite. O algoritmo mais conhecido por trabalhar com agrupamento por densidade é o **DBSCAN**: este algoritmo trabalha com as definições de pontos dadas, sendo que cada dois pontos de centro que estejam suficiente próximos dentro de uma distância Eps entre si são colocados no mesmo grupo, qualquer ponto de limite que esteja o suficientemente próximo ao ponto de centro é colocado no grupo e os pontos ruído são descartados.

As breves definições apresentadas são geralmente as mais utilizadas ou mais abrangentes para os algoritmos de agrupamento. Contudo, algoritmos de agrupamento tem sido desenvolvidos para diferentes domínios e aplicações. Nem todos os algoritmos apresentados anteriormente são apropriados para todos os tipos de dados (texto, categórico, grafos, séries temporais), o motivo disso é que o que constitui um bom resultado de agrupamento está aberto a uma interpretação subjetiva. Outros tipos de agrupamento segundo a área podem gerar modelos probabilísticos, modelos espectrais, para dados *streaming* ou series temporais (AGGARWAL; REDDY, 2013; TAN; STEINBACH; KUMAR, 2005).

2.3.3 Avaliação Quantitativa de agrupamento

A avaliação de grupos muitas vezes é executada como parte de uma análise exploratória de dados, tornando-se subjetiva. No entanto, uma motivação pela qual a avaliação deve fazer parte de uma análise de agrupamento está em entender por que os algoritmos retornam diferentes resultados e reduzir a necessidade de tentativa e erro para procurar grupos ótimos. Portanto, as questões importantes para a validação de grupos estão baseadas na tendência de agrupamento, determinar o número correto de grupos, avaliar a qualidade dos resultados, comparar diferentes resultados e comparar grupos.

As medidas ou índices de validação de agrupamento auxiliam nessa análise, podendo variar dependendo do campo de aplicação. Contudo, pesquisas realizados por (ARBELAITZ *et al.*, 2013; YERA *et al.*, 2017; TAN; STEINBACH; KUMAR, 2005) apresentam uma visão geral dos índices de validação aplicados a algoritmos de agrupamento. Considera-se que em geral os índices de validação de agrupamento são aplicados para três aspectos do agrupamento: **Internos**, **Externos** e **Relativos**.

Índices Internos: também conhecidas como não supervisionados porque medem a qualidade de uma estrutura do agrupamento utilizando dados presentes nos grupos e sem dados externos ou relações externas. Mede tanto a coesão, para determinar o quão intrinsecamente os dados de um grupo estão relacionados, quanto a separação, que determina o quão distintos ou bem separados os grupos estão uns dos outros. Algumas medidas de coesão e separação

são: SSE (soma dos erros quadrados, mostrado na equação 2.2) que mede o quão disperso os objetos estão de um centroide ou protótipo; SBB (soma de quadrados intermediário) que mede a separação do agrupamento com a soma das distâncias entre os centroides/protótipos; Coeficiente de Silhueta (equação 2.4), que mede a contribuição dos dados em um grupo considerando coesão e separação.

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} dist(c_i, x)^2 \quad (2.2)$$

$$SSB = \frac{1}{2K} \sum_{i=1}^K \sum_{j=1}^K \frac{m}{k} dist(c_i, c_j)^2 \quad (2.3)$$

$$S_i = \frac{b(X_i, C_k) - a(X_i, C_k)}{\max \{b(X_i, C_k), a(X_i, C_k)\}}$$

$$S(C) = \frac{1}{N} \sum_{c_k \in C} \sum_{X_i \in c_k} \frac{b(X_i, C_k) - a(X_i, C_k)}{\max \{b(X_i, C_k), a(X_i, C_k)\}} \quad (2.4)$$

$$a(X_i, C_k) = \frac{1}{C_k} \sum_{x_j \in C_k} D(x_i, x_j)$$

$$b(X_i, C_k) = \min_{c_i \in C \setminus c_k} \left\{ \frac{1}{C_k} \sum_{x_j \in C_k} D(x_i, x_j) \right\}$$

Onde k é o número de grupos e c_i representa o grupo para instância i . O valor do coeficiente de silhueta consiste de três passos: para objeto de índice i , é calculada sua distância média até outros objetos no seu grupo a . Para o objeto de índice i e qualquer grupo que não contenha esse objeto é calculada a distância média do objeto até todos os objetos nesse grupo, encontrando o menor entre esses valores com relação a todos os grupos b . Para cada objeto de índice i , o coeficiente de silhueta é Sil_i , para o grupo é a somatória S_i . O $S(C)$ representa a somatória dos coeficientes de silhueta para o agrupamento. Os valores de silhueta variam entre -1 e 1. Neste trabalho de mestrado é utilizado o coeficiente de silhueta.

Índices Externos: conhecidos como supervisionados por medir a estrutura de um agrupamento segundo dados externos ou alguma estrutura externa. Esses índices externos usam informações que não estão presentes nos grupos. O procedimento habitual é medir o grau de correspondência entre os rótulos de grupos e os rótulos das classes para comparar ou avaliar o quanto o processo de classificação manual corresponde ao processo automático da análise de agrupamento. Os índices de maior uso são: **Entropia** (equação 2.5), que mede o grau no qual cada grupo consiste de objetos de uma única classe; A **pureza** (equação 2.6), que mede o quanto um grupo contém dados de uma única classe; A **Precisão** (equação 2.7), que determina quantos dados de um grupo pertencem a uma classe específica; A **Lembrança** (equação 2.8) é a extensão

na qual um grupo contém todos os dados de uma classe específica e a **Medida F** (equação 2.9) é a combinação de precisão e lembrança que mede a extensão na qual um grupo contém apenas dados de uma determinada classe e todos os objetos dessa classe.

$$e_i = - \sum_{j=1}^L P_{ij} \log_2 P_{ij} \quad (2.5)$$

$$pureza = \sum_{i=1}^K \frac{m_i}{m} P^i \quad (2.6)$$

$$precisao(i, j) = P_{ij} \quad (2.7)$$

$$lembranca = m_{ik}/m_j \quad (2.8)$$

$$F(i, j) = (2 \times precisao(i, j) \times lembranca(i, j)) / (precisao(i, j) + lembranca(i, j)) \quad (2.9)$$

Onde P_{ij} é a probabilidade que um membro de i pertença à classe j , m é a quantidade de dados em um grupo, m_i ou m_j representam a quantidade de dados em um grupo da classe i ou j . Para a entropia L é o numero da classes.

Índices Relativos: no geral os índices relativos são utilizados para comparar dois ou vários agrupamentos já avaliados usando índices internos ou externos.

2.4 Considerações Finais

O principal enfoque nesta seção foi estabelecer os conceitos fundamentais de visualização de dados multivariados e agrupamento, a fim de facilitar o entendimento das áreas envolvidas neste trabalho de mestrado. Tanto os processos de mineração descritiva de dados quanto os de visualização de dados multivariados atacam os grandes desafios da Analítica Visual (THOMAS; COOK, 2005b).

Contudo, os processos de mineração de dados muitas vezes ainda são entendidos como caixas pretas, fornecendo resultados mas não apresentando como tais resultados obtidos ou os processos de decisão envolvidos. As projeções multidimensionais não apresentam informações sobre os atributos/dimensões, não sendo possível entender de forma explícita a lógica por trás da formação do layout, tornando o processo de exploração da projeção mais difícil. No próximo capítulo são apresentadas as pesquisas relacionadas que unem visualização com análise de agrupamento para conjuntos de dados.

TRABALHOS RELACIONADOS

3.1 Considerações Iniciais

Este capítulo apresenta pesquisas realizadas referentes aos tópicos principais deste projeto de mestrado. Em particular, na seção 3.2 são apresentadas pesquisas relacionadas à escalabilidade visual, que é parte integral desse projeto. Na seção 3.3 são apresentados trabalhos relacionados à analítica visual que envolvem a avaliação de agrupamentos auxiliada com técnicas de visualização de informação. Também são descritas pesquisas feitas pelos pesquisadores do grupo VICG-ICMC no tópico de seleção de dimensões para a avaliação de aprendizagem supervisionada. Finalmente, a seção 3.4 apresenta as considerações finais levantando vantagens e limitações das abordagens descritas nestes trabalhos.

3.2 Escalabilidade Visual

No geral as técnicas de visualização de dados multivariados consistem em analisar conjuntos de dados grandes e/ou complexos. No entanto, tais técnicas não são escaláveis visualmente já que tendem a apresentar oclusão, desordem e perda do contexto da visualização, (VIERA, 2008; LU *et al.*, 2012). Como apresentado na seção anterior (2.2.2), uma abordagem que tem sido utilizada juntamente com as técnicas de visualização é a agregação - a apresentação de dados derivados que representam intrinsecamente um grupo ou o conjunto de dados. Esta seção descreve técnicas estudadas capazes de melhorar a representação de grandes conjuntos de dados.

O método proposto por (CHUAH, 1998) apresenta uma visualização híbrida, combinando a técnica de histograma circular *SolarPlot* (CHUAH, 1998) e a *Polar TreeMap* (JOHNSON, 1993) em uma representação agregada. O processo de agregação baseia-se nos seguintes conceitos: agregação automática, possibilidade de alteração do nível da agregação, identificação de diferentes níveis de agregação para manter o contexto, retroalimentação dos níveis de agregação. A Figura 5 ilustra sua aplicação para um conjunto de dados da frequência de correios recebidos

em uma empresa. Esta visualização apresenta a estrutura organizacional de uma empresa, sendo a raiz da hierarquia representada pelo CEO, o primeiro nível por vice-presidentes e assim por diante. O histograma em cada nó representa a quantidade de correios recebidos por pessoa no dia. A cor dos nós representa o tempo de trabalho das pessoas na empresa. Um problema desta visualização híbrida (Figura 5) é que cada nó tem um ângulo diferente de apreciação, no qual o histograma é apresentado. Portanto, os níveis de agregação são diferentes em cada nível da estrutura visual.

Figura 5 – Junção da *SolartPlot* e *Polar TreeMap* cria uma Visualização híbrida, que apresentam um conjunto de dados de correios recebidos numa empresa.

Fonte: Chuah (1998).

Algumas abordagens mais sofisticadas para apresentar agregação fazem uso de operações de interação e navegação na visualização, geralmente baseadas em similaridade dos dados. FUA; WARD; RUNDENSTEINER fornecem agrupamento hierárquico para dados multivariados por meio de coordenadas paralelas hierárquicas (do inglês *Hierarchical Parallel Coordinates*) (FUA; WARD; RUNDENSTEINER, 1999). Neste trabalho a técnica de visualização coordenadas paralelas foi estendida inserindo agrupamento hierárquico nos dados multivariados para criar múltiplas visões e poder apresentar dados agregados na visualização. Diferentes estatísticas sobre cada agrupamento são calculadas e visualizadas: o número de dados que compõem o grupo, a média, mínimo e máximo e a profundidade na hierarquia. Os grupos são representados por uma faixa de largura e opacidade variáveis, a média do grupo por um valor representativo mostrado como uma polilinha que atravessa os eixos paralelos e a largura em cada eixo dada pelos valores mínimo e máximo para esta dimensão no grupo. A polilinha e a faixa que representa o agrupamento são coloridas baseando-se na similaridade da hierarquia entre os grupos.

Figura 6 – Coordenadas paralelas hierárquicas com vários níveis de detalhe. 6a é a representação de um agrupamento, 6b apresenta vários agrupamentos agregados, cada um em uma faixa menor, 6c apresenta vários agrupamento e usa-se a cor para identificar a similaridade entre os grupos.

Fonte: Fua, Ward e Rundensteiner (1999).

A Figura 6 ilustra um conjunto de dados com oito dimensões e 230.000 instâncias em diferentes níveis de detalhe. Estes níveis de detalhe são entendidos como níveis da estrutura hierárquica, controlados pela interação do usuário. Algumas técnicas de interação que estão integradas nesta visualização são: *Brushing*, para a localização de um subconjunto, *Drill-down and Roll-up*, para a visualização dos dados a um dado nível de hierarquia, e *Dynamic Masking*, para controlar o nível de opacidade das áreas identificadas com *Brushing*.

YANG; WARD; RUNDENSTEINER propuseram outras técnicas de visualização multi-variada. Eles definiram o framework *Iterative Hierarchical Display* (IHD), que aplica conceitos de agregação para matrizes de *Scatterplots*, *Star Glyphs* e *Dimensional Staking*. Na Figura 7b é ilustrada a *Star Glyphs* hierárquica aplicada para um conjunto de dados de avaliação de carros produzidos entre 1970 e 1990, onde cada glifo representa um grupo de carros. A média de cada grupo é utilizada para gerar a forma básica. A faixa ao redor do polígono tem duas bordas, uma

fora do polígono e outra dentro do polígono. A borda interna de cada eixo intersecta o valor mínimo da sua respectiva dimensão, enquanto a borda exterior de cada eixo intersecta o valor máximo.

Figura 7 – Star Glyphs Hierárquica para 392 modelos de carros produzidos entre 1970 e 1990, (a) original *Star Glyphs*, (b) hierárquica *Star Glyphs*.

Fonte: Yang, Ward e Rundensteiner (2003).

Figura 8 – Matriz de ScatterPlots Hierárquica para 392 modelos de carros produzidos entre 1970 e 1990, (a) original Matriz de ScatterPlots, (b) Hierarquical Matriz de ScatterPlots.

Fonte: Yang, Ward e Rundensteiner (2003).

Modelos de agregação também foram aplicados a projeções de dados. Uma ideia desenvolvida por PAULOVICH; MINGHIM é a HiPP *Hierarchical Point Placement Strategy* (PAULOVICH; MINGHIM, 2008). A HiPP é uma técnica para a visualização, interação e organização hierárquica de conjuntos de dados multidimensionais, com especial atenção para coleções de documentos. O objetivo é a preservação de relações de similaridade nos vários níveis de

Figura 9 – Hipp para coleção de 30.000 documentos de Reuters Corpus, dados projetados usando LSP.

Fonte: [Paulovich e Minghim \(2008\)](#).

detalhe, agrupando instâncias similares enquanto também separa visualmente grupos dissimilares. A *HiPP* oferece uma estratégia de reorganização dos dados com base na representação visual gerada, criando agrupamentos unidos ou separados, reorganizando os dados de acordo com necessidades específicas do usuário.

Para representar hierarquia em dados não intrinsecamente hierárquicos, usa o algoritmo *bisecting KMeans*. Os elementos são apresentados em um círculo. Primeiramente é mapeado o nó raiz, com o tamanho de cada círculo representando o numero de elementos ou instâncias de dados que pertencem a ele. Os grupos e subgrupos de instâncias de dados são representados usando círculos dentro de círculos, cuja posição é dada pela semelhança entre os subgrupos que o representam.

A Figura 9 mostra um conjunto de dados de uma coleção de 30.000 documentos coloridos segundo a frequência da palavra "Clinton". Nesta visualização os grupos de documentos são explicitamente representados e suas densidades são mapeadas por tamanho do circulo e cor. Esta técnica pode ser classificada como uma técnica de agregação híbrida, por combinar agregação de dados e agregação de atributos.

O trabalho recentemente proposto por ([PAGLIOSA et al., 2013](#)) denominado Informação Multiescala e Sumarização de Texto (MIST) faz junção de duas abordagens, a *word cloud* e a *rigid-body (Box2D)*, para exibir coleções de documentos. Esta metáfora visual utiliza círculos para representar grupos de documentos, onde o tamanho de cada círculo define a importância dos documentos no conjunto de dados. Dados menos relevante são exibidos como pontos menores para reforçar a ideia de densidade, podendo indicar quais regiões na visualização incluem dados ocultos e devem ser mais exploradas. Geralmente, cada grupo de documentos vizinhos formado gera uma *word cloud* que é visualizada usando *rigid-body*, a fim de aproximar grupos similares

Figura 10 – MIST para coleção de documentos das conferências de visualização de informação da IEEE de 1995 até 2002, dados projetados usando LSP.

Fonte: Pagliosa *et al.* (2013).

e evitar a sobreposição dos dados. No caso deste trabalho, os dados derivados que apresentam agregação são visualizados pelos tamanhos dos círculos e por pontos ao redor dos círculos que mostram a densidade de dados ocultos. A Figura 10 ilustra a MIST usando a técnica de projeção LSP para uma coleção de 616 artigos publicados na conferência de visualização de informação da IEEE de 1995 até 2002, composta por metadados, título, palavras-chave, resumo e referências. Quando os dados são mapeados com a LSP, cada um representa um *rigid-body* ou círculo. Uma vez que os dados são mapeados, o conteúdo do círculo é representado usando a *word cloud*: palavras-chaves com o maior numero de ocorrências são mapeados no centro do círculo e com uma fonte de letras maior que das palavras menos relevantes. Nesta visualização o usuário pode selecionar uma região para explorar dados ocultos por meio de zoom, esta seleção é ilustrada com retângulo azul da Figura 10. As setas azuis representam a transição da navegação zoom.

Finalmente, podemos concluir que o uso de agregação é visado tanto para melhorar a representação visual quanto para apresentar informações detalhadas, características que facilitam a tarefa de análise visual (ELMQVIST; FEKETE, 2010). Neste projeto de mestrado empregamos a metáfora de agregação como método de escalabilidade visual. A seguir apresentamos um estudo de técnicas de análise visual de agrupamento.

3.3 Avaliação Visual de Agrupamento

Segundo Keim Daniel, Mansmann e Florian (2010), embora tenham sido desenvolvidas metáforas visuais para avaliar a qualidade e eficiência de resultados de algoritmos de agrupamento, determinar a validade destes resultados ainda continua sendo uma tarefa subjetiva. Isto ocorre principalmente devido à mineração de dados, em geral, ser uma tarefa intensiva em termos

de precisão, sendo dependente da configuração dos parâmetros dos métodos utilizados (L'YI *et al.*, 2015). Desse modo, métodos e métricas para a avaliação tanto de qualidade da visualização quanto das técnicas de mineração de dados foram propostos por diversos autores.

Além de técnicas quantitativas de avaliação de resultados de agrupamento apresentadas na seção 2.3.3, existem técnicas visuais com o mesmo intuito. Elas são apresentadas nesta seção em termos de coesão, separação, detecção de dados anômalos, estabilidade do grupo, etc.

A análise de agrupamento na área de bioinformática tem sido utilizada também como instrumento de avaliação e comparação de diversos resultados apresentados pelos algoritmos de agrupamento. Por exemplo, a recente pesquisa apresentada por (L'YI *et al.*, 2015) denominada *XCluSim* permite avaliar agrupamentos de forma visual, permitindo ao usuário interativamente comparar diferentes resultados dos agrupamentos e explorá-los baseando-se no paradigma de visualização (Visão Geral, filtragem, zoom e detalhe) mais conhecido como *Seeking Mantra* (SHNEIDERMAN, 1996). Nessa ferramenta, é fornecida uma visão geral das similaridades dos grupos e agrupamentos; desta forma, avaliar diferentes agrupamentos se torna uma tarefa mais fácil. Para identificar as similaridades entre os agrupamentos é utilizada a metáfora de codificação de cor conhecida como *Tree Colors*. A estrutura *Tree Colors* é usada para colorir a similaridade entre agrupamentos: primeiramente, é aplicada sobre os resultados uma técnica de agrupamento hierárquico (*HAC*) usando como métrica de similaridade o coeficiente de correlação entre os grupos. A partir dessa estrutura, a *Tree Colors* cria uma estrutura de árvore colorida, na qual os grupos que compartilham instâncias similares em cada agrupamento são coloridos com a mesma cor.

A Figura 11 ilustra a interface gráfica da ferramenta *XCluSim* para quatro diferentes algoritmos de agrupamento (*HAC*, *self-organizing map (SOM)*, *KMeans* e *expectation-maximization (EM) clustering*), dividida em cinco seções. A seção (A) apresenta um resumo dos parâmetros usados e a quantidade de grupos criados em cada resultado da execução dos algoritmos. A seção (B) mostra um esquema de força de grafos *FDL* entre os algoritmos, usando como métrica de força a *Medida F*. A finalidade é apresentar a adjacência dos resultados obtidos pelos algoritmos, conetando-os como grafos. Na seção (C) é visualizada no dendrograma a similaridade hierárquica dos resultados apresentados em cada execução dos algoritmos, sendo a métrica de similaridade também a *Medida F*. Para visualizar a precisão com a qual as instâncias são atribuídas aos mesmos grupos em cada algoritmo de agrupamento, a seção (D) mostra uma metodologia baseada em conjuntos de visões paralelas denominado *Stable Group*, onde cada linha paralela representa um algoritmo de agrupamento, as barras representam os grupos resultantes do algoritmo e a conexão entre as linhas representam as instancias em diferentes posições segundo o grupo.

Na mesma área, a pesquisa realizada por DING *et al.* apresenta uma ferramenta de análise visual utilizando as técnicas *Parallel Set*, *Silhouette Plot* e *Heat Map*. Esta pesquisa é mais conhecida como *interactive Genomics Patient Stratification explorer (iGPSe)*, uma ferramenta web baseada na análise de dados genômicos usando conjuntos de dados de casos de câncer. A

Figura 11 – *XCluSim* ferramenta para avaliação de resultados de agrupamentos: (A) informação de parâmetros, (B) diagrama de força, (C) vista no dendrograma das similaridade dos algoritmos de agrupamento, (D) Similaridade de grupos em linhas paralelas, (E) visualização dos dados originais.

Fonte: L'Yi *et al.* (2015).

iGPSe é apresentada como uma forma de explorar, avaliar e comparar a qualidade dos grupos nos agrupamentos. Visando aproveitar a habilidade humana para interagir com a ferramenta, é permitido ao usuário escolher os algoritmos de agrupamento e executar a análise *On-the-fly*. Adicionalmente, é provida uma fase denominada *cluster refinement*, que avalia o rendimento e acurácia dos algoritmos, proporcionando ao usuário um entendimento amplo sobre o acerto ou equívoco dos parâmetros escolhidos.

A Figura 12 ilustra parte da ferramenta de visualização *iGPSe*, que faz uso de 4 técnicas de visualização de informação. Na Figura 12a é apresentado um *heat map* e o gráfico de silhueta. Nessas visualizações foi utilizado como algoritmo de agrupamento o *KMeans* para $k = 4$. O *heat map* é usado para mostrar os resultados de agrupamento e o gráfico de silhueta para avaliar a qualidade do agrupamento: valores próximos a 1 são grupos bem compactos e próximos a -1 são grupos cujos objetos estão espalhados. Para a representação da silhueta, usa-se o gráfico de linha, na qual as linhas para a direita são valores positivos e linhas para a esquerda, valores negativos. A Figura 12b ilustra a visualização de grafos, usando-se como método de visualização a *FDL*: nesta visualização, cada nó (vértice) é um paciente e as conexões (arestas) entre nós mostram relações de similaridade entre os pacientes no grupo. A Figura 12c ilustra a *Parallel Set* para avaliar a coesão e estabilidade para dois agrupamentos usando o mesmo conjunto de dados, mas

Figura 12 – *iGPSe*, ferramenta de análises visual para dados de câncer.

Fonte: Ding *et al.* (2014).

com diferentes atributos. Nessa visualização os grupos são apresentados como colunas onde a altura da coluna depende da quantidade de dados no grupo. A largura de linha que junta os resultados dos dois agrupamentos indica a quantidade de pacientes que compartilham os grupos. Na Figura 12d é mostrada a visualização denominada *Kaplan-Meier Survival Plot* que permite entender a tendência e a relação entre grupos selecionados pelo usuário. Nessa visualização por exemplo, a relação entre grupo é realizada pelos atributos probabilidade de sobrevivência e o tempo de sobrevivência. A ferramenta permite a interatividade com o usuário, proporcionando-lhe duas métricas de similaridade para serem escolhidas: Coeficiente de correlação de Pearson e Distância Euclidiana. Também são oferecidos diferentes algoritmos de agrupamento: *KMeans*, *Spectral KMeans* e Detecção de comunidades.

Outra ferramenta para a análise visual de agrupamentos foi desenvolvida por (TURKAY *et al.*, 2013), denominada *CIVA* (*Clustering Interactive Visual Analysis*). Esta ferramenta permite interativamente comparar as tendências dos agrupamentos, assim como estabilidade e coesão dos grupos. Na *CIVA*, utiliza-se diferentes visões para criar as conexões de associação entre os resultados obtidos pelos algoritmos de agrupamento. Nota-se que o foco nessa pesquisa foi utilizar diferentes metáforas visuais e técnicas de visualização, como *scatterplots*, *heat maps*,

Histograma e Coordenadas Paralelas. Estas diferentes visões auxiliam o usuário a entender, verificar e aprimorar os resultados e, caso necessário, fornecer novos parâmetros. As fases desta ferramenta são: formação de grupos e avaliação dos agrupamentos. Estas fases são repetidas conforme se realiza a interação do usuário. Na fase de formação de grupos, é fornecida ao usuário a possibilidade de escolher a exploração das instâncias, dimensões a serem agrupadas e o algoritmo de agrupamento.

Figura 13 – Análise visual de resultados de agrupamento usando coordenadas paralelas.

Fonte: Turkay *et al.* (2013).

Na Figura 13 são ilustradas as fases da ferramenta CIVA. A Figura 13a mostra como a representação das coordenadas paralelas proporciona uma visão da estabilidade das instâncias nos agrupamentos. Os eixos verticais paralelos representam os diferentes algoritmos de agrupamento C e as representações dos grupos. C_0 é *KMeans* com $k = 5$, C_1 é *HC* com $k = 5$ e C_3 é *HC*

com $k = 3$. As instâncias são linhas curvas coloridas de acordo com os grupos, que cruzam os eixos, juntando-os segundo a posição do grupo em cada eixo. A importância dessa representação visual está em visualizar a variação das instâncias para os grupos e a relação dos agrupamentos. Similarmente, mediante a seleção de um grupo, a Figura 13b mostra a *Heat-Map*, a qual é utilizada para visualizar a matriz de dissimilaridade de um grupo, neste caso o grupo selecionado pelo usuário é C_1 . Desta maneira, a ferramenta possibilita o uso da interação usuário-computador para explorar sub-conjuntos de grupos, baseando-se na formação, comparação de grupos e a estabilidade das instâncias. demonstra como a junção de duas técnicas de visualização apoia a exploração da formação dos grupos. Nessa imagem

A ferramenta de visualização para comparar agrupamentos CComViz (ZHOU; KONECNI; GRINSTEIN, 2009) (*Cluster Comparision Visualization*), que possibilita analisar a estabilidade dos grupos, foi proposta e posteriormente estendida por ZHOU *et al.*. Neste trabalho, representa-se visualmente os diferentes resultados obtidos na análise de dados genéticos utilizando algoritmos de agrupamento particionais *KMeans* e *EM*. A principal ideia deste trabalho é fornecer uma ferramenta visual para aperfeiçoar a escolha dos pontos iniciais de controle e o numero de grupos. A Figura 14 ilustra os resultados desta abordagem usando 7 algoritmos diferentes de agrupamento com $k = 6$: *KM* (*KMeans*), *EM* (*Expectation Maximization*), *XM* (*Extended KMeans*), *DB* (*Density-based clustering built on KMeans*), *SOM* (*Self-Organizing Map*), *HC* (*Hierarchical Clustering with tree cutoff*) e *FF* (*Farthest First*). Para analisar os resultados da aplicação de vários algoritmos de agrupamento, os autores utilizam a metáfora visual das coordenadas paralelas para dados categóricos. Cada barra colorida representa um grupo do agrupamento que é representado pelos eixos. A última coluna da visualização representa a métrica de estabilidade dos dados para o conjunto de agrupamento. Nesta visualização pode-se observar que existe uma variação das instancias para cada eixo, possivelmente representando o parâmetro exato k . Porém, estas variações também criam desordem e oclusão, que é um problema presente na visualização de dados multi-variados.

Figura 14 – Enfoque visual para o aperfeiçoamento do parâmetro k para algoritmos de agrupamento particionais.

Fonte: Zhou *et al.* (2010).

Geralmente, quando se trata da análise de dados multidimensionais, fazer uma analogia entre o resultado obtido e os grupos criados é uma tarefa fundamental. Em função disso, o trabalho proposto por (LEX *et al.*, 2010) desenvolveu uma ferramenta visual denominada *Matchmaker*, que usando a metáfora visual das coordenadas paralelas permite ordenar e comparar as dimensões, grupos e algoritmos de agrupamento. Neste trabalho são apresentadas duas etapas de avaliação visual. A primeira etapa aborda as relações das dimensões, os grupos e os dados de cada grupo. Um eixo pode conter uma ou varias dimensões, uma dimensão pode estar em vários eixos. A segunda etapa mostra visualmente o resultado de aplicar três diferentes algoritmos de agrupamento para um conjunto de dados de expressão genética.

A Figura 15a apresenta a primeira etapa da ferramenta *Matchmaker* para um conjunto de dados com 78 dimensões e 400 instâncias, onde as dimensões são apresentadas como eixos paralelos, um eixo podendo conter varias dimensões agrupadas. Cada *heat map* em um eixo representa um grupo gerado pela aplicação de um algoritmo de agrupamento para o eixo. Os eixos são unidos por fitas com uma largura que depende dos grupos em cada eixo. Esta ferramenta também permite diferentes formas de interação do usuário, como por exemplo ordenar os valores dos eixos. A abordagem (*Overview - Detail*) permite ver detalhes dos dados nos grupos. O (*Brushing - Highlight*) feito nesta ferramenta possibilita selecionar um grupo de instancias e ver a variação para cada eixo. A Figura 15b ilustra a outra etapa, que corresponde à análise visual de resultados dos diferentes algoritmos de agrupamento particionais. Os resultados dos algoritmos são apresentados em eixos paralelos, cada eixo é dividido por quadrados na qual a altura depende da quantidade de dados no grupo. Os eixos são unidos por fitas que mostram as fragmentações

Figura 15 – Avaliação e relação visual de algoritmos de agrupamento particionais aplicados a conjuntos de dados genéticos

Fonte: [Lex et al. \(2010\)](#).

conforme ao algoritmo.

Heat maps ainda são utilizados por outras visualizações de agrupamento, como a abordagem de ([SHARKO et al., 2007](#)) para comparar diversos algoritmos de agrupamento usando *Forelimbs Microarray Data*. Essa metodologia visual está focada na métrica estabilidade de grupo (*Cluster Stability*). Esta métrica fundamenta-se em analisar quantas vezes os dados aparecem em um mesmo grupo ao longo de diferentes processos de agrupamentos, já seja usando diferentes algoritmos ou parâmetros. Assim, essa métrica auxilia identificar a variação dos agrupamentos e grupos iguais.

Figura 16 – *Heat-Map* para avaliar estabilidade dos agrupamentos usando *Microarray Data*

Fonte: [Sharko et al. \(2007\)](#).

A Figura 16 ilustra as formas como a *heat map* foi utilizada por [SHARKO et al.](#). A

Figura 16a mostra a métrica de matriz de estabilidade para 4 agrupamentos (*KMeans*, *Expectation Maximization*, *Extend KMeans*, *Density base clustering KMeans* com $k = 11$, os eixos estão ordenados pelos rótulos obtidos do *KMeans*. O eixo X e eixo Y representam os dados de genes. Ao invés de relacionar cada gene, nessa pesquisa cada pixels representam a quantidade de vez que os genes ficam no mesmo grupo. As cores brilhantes são valores que apresentam maior estabilidade e as cores escuras menor estabilidade para o gene no eixo X e Y. Ao respeito de comprar dois algoritmos de agrupamento diferentes. A Figura 16b ilustra a comparação entre dois algoritmos: *DB* e *KMeans*, ao invés de relacionar genes, nessa figura são relacionados os grupos. Isto permite entender quantas vezes os dois algoritmos geram os grupo com a mesmas instâncias. Também são utilizados outros algoritmos como *Expectation Maximization*, *Extended KMeans*, *Density Based Clustering*.

Seo e Shneiderman (2002) introduziram diferentes conceitos em relação de estabilidade e coesão visual no trabalho (*HCE - Interactively Exploring Hierarchical Clustering Result*), exploração interativa hierárquica do usuário-computador para a análise de resultados de agrupamentos (SEO; SHNEIDERMAN, 2002). Nessa pesquisa foi utilizado o dendrograma como metáfora visual para visualizar os diferentes resultados dos algoritmos de agrupamentos hierárquicos. Nessa abordagem o enfoque é encontrar o par de genes com os perfis de expressões mais semelhantes. Iterativamente é construída uma hierarquia agrupando genes que são mais semelhantes. A fim de lidar com grandes conjuntos de dados, folhas próximas no dendrograma são agregadas, apresentando a média dos dados agregados como uma nova folha. Quando uma metáfora de agregação é utilizada, deve-se fornecer a exploração dos dados agregados. Assim, a *HCE* usa exploração de dados hierárquicos por meio da barra de rolagem. Também é possível a escolha entre diferentes algoritmos de agrupamento hierárquico.

Figura 17 – *HCE*, Exploração hierárquica para a análise de resultados de agrupamentos.

Fonte: Seo e Shneiderman (2002).

Na Figura 17 é mostrada especificamente a abordagem de comparação e relação dos diferentes agrupamentos hierárquicos. Mediante a interação, permite-se selecionar um grupo e entender a relação das instâncias desse grupo com os grupos de outro agrupamento. A forma de

Figura 18 – *Vectorized RadViz* para análise de resultados de agrupamento, baseado-se no *Cluster Ensamble*, identifica os agrupamentos semelhantes.

Fonte: [Sharko, Grinstein e Marx \(2008\)](#).

representar a relação são linhas que cruzam entre os agrupamentos e os grupos inter-relacionados são mostrados no *heat map* e o dendrograma usando a matriz de similaridade.

O trabalho denominado *Vectorized RadViz* (VRV) por ([SHARKO; GRINSTEIN; MARX, 2008](#)) analisa resultados de diversos agrupamentos utilizando a metáfora de projeções de dados multidimensionais RadViz (ver seção 2.2.1.1). Desta forma, auxilia-se a encontrar diferentes relacionamentos entre os diferentes agrupamentos. Para obter esta representação, a RadViz trabalha com valores de instâncias de 1 e 0, o valor 1 indica que a instância foi agrupada nessa dimensão (grupo) corretamente e o valor 0 que a instância não pertence ao grupo. De forma geral, quando um processo de avaliação de resultados de agrupamento é aplicado a um conjunto de dados, gera-se um novo conjunto de dados onde os diferentes agrupamentos são representados como novos atributos e as instâncias representam os valores dos rótulo do grupo daquele atributo. Para visualizar as similaridades entre os grupos e agrupamentos, as ADs são ordenadas levando em conta a métrica de distância *Class Discrimination Layout algorithm (t-Statistic)*, que é usada como métrica de similaridade das dimensões segundo os dados nos grupos. A (*t-Statistic*) é usada para gerar a similaridade e atribui o rótulo do grupo para cada dimensão ao maior valor da média dos grupos naquela dimensão. Assim, as dimensões (ADs) que tenham o mesmo rótulo do grupo são apresentadas adjacentes, nomeadas setor de classificação.

A Figura 18 ilustra duas visões da VRV, o conjunto de dados usado primeiramente para avaliar o método foi o Iris. Similarmente às pesquisas descritas anteriormente, o método VRV tenta visualizar a estabilidade dos grupos e a similaridade entre os agrupamentos. A Figura 18a ilustra estes conceitos: os agrupamentos sendo representados pelas ADs são visualizados adjacentes devido a proximidade das instâncias que foram agrupadas nos mesmos grupos. Na Figura 18a são apresentadas linhas que mostram a relação e conexão das instâncias com os grupos. Uma instância com poucas ou nenhuma conexão foi agrupada de forma correta no grupo, enquanto muitas conexões indicam inconsistência para aqueles grupos. Círculos com diferentes

raios são apresentados nessa visualização indicando que instâncias próximas são agregadas: quanto maior for a quantidade de instâncias adjacentes, maior o tamanho do círculo.

3.4 Considerações Finais

Neste capítulo foram apresentadas diversas pesquisas relacionadas às técnicas de visualização de informação que envolvem mineração de dados não supervisionada, em especial, técnicas de agrupamento (*clustering*), mostrando como a interatividade do usuário pode auxiliar a análise visual de dados um processo mais amigável de compreensão. Também foram apresentadas pesquisas que usam agregação para auxiliar na escalabilidade visual.

As representações visuais dos processos de mineração de dados apresentam um conteúdo mais interativo, permitindo ao usuário ser uma fonte ativa, interagindo com o layout e concentrando-se em inferir das ocorrências particulares da visualização. Porém, nem sempre o resultado apresentado pelo processo de mineração de dados, agrupamento, pode ser de um todo considerado válido, portanto que avaliar esses demanda de longa experiência. Com relação à validação de resultados, para avaliar agrupamentos não existe uma métrica geral que possa identificar o melhor resultado de forma visual e determinar de forma correta os melhores parâmetros. Isto ocorre devido à diversidade nos dados que estão sendo analisados e na subjetividade do analista.

Finalmente, obter escalabilidade visual e facilitar as tarefas de análise visual de resultados de agrupamento são tarefas que podem ser aprimoradas utilizando os paradigmas de visualização descritos. Com isso, possibilita-se construir uma visualização que auxilia ao usuário na análise de agrupamento com maior quantidade de dados. A abordagem proposta neste texto é capaz de apresentar graficamente uma maior quantidade de dados usando métodos de agregação, otimizando o espaço visual. Ainda que uma limitação observada na avaliação visual de agrupamento é a desordem, quando se avalia estabilidade de grupo, nossa abordagem apresenta uma forma simples e limpa de apresentar estabilidade e comparar agrupamentos.

METÁFORA VISUAL ESCALÁVEL PARA DADOS TABULARES E APLICAÇÃO EM ANÁLISE DE AGRUPAMENTOS

4.1 Considerações Iniciais

Dentre os diferentes tipos de modelos de dados usados para visualização, os modelos de organização tabular são os mais populares (TELEA *et al.*, 2006). A principal razão dessa popularidade está na facilidade deles serem apresentados em uma visualização. De forma geral, esse tipo de dados é também chamados de dados multivariados ou multidimensionais, onde as colunas representam as dimensões ou atributos e as linhas representam as instâncias de dados. Na maioria dos casos, as visualizações utilizadas na análise de dados tabulares visam fornecer subsídios para o entendimento de como as instâncias e as dimensões estão relacionadas ou encontrar valores espúrios (MUNZNER; MAGUIRE, 2015).

Uma das técnicas de visualização mais populares para dados tabulares é a *TableLens* (RAO; CARD, 1994a). A *TableLens* é uma extensão e melhoria da metáfora visual para as usuais planilhas eletrônicas (*spreadsheets*) que dá suporte a análise visual de correlação entre atributos e a descoberta de valores espúrios. Na *TableLens*, cada valor de um determinado atributo em uma instância é representado por meio de uma linha com altura de um pixel e largura proporcional ao valor do atributo. Apesar da sua popularidade, a *TableLens* apresenta limitações quanto a escalabilidade visual, isto é, a capacidade de representar os dados considerando o espaço visual disponível. O problema é que a quantidade máxima de dados apresentados esta relacionada à resolução do dispositivo de saída. Por exemplo, se a resolução de um monitor for 2000 pixels de altura, somente podem ser visualizados 2000 instâncias de dados. Barras de rolagem pode ser usadas como solução, mas perde-se o contexto ou visão geral dos dados (JERDING; STASKO, 1998; CUI *et al.*, 2006).

Assim, um dos desafios deste trabalho de mestrado foi definir uma estratégia baseada na metáfora visual da *TableLens* que não seja limitada pela resolução vertical dos dispositivos de saída e que não perca o contexto geral dos dados. Desenvolvemos uma nova abordagem visual para representar maiores quantidades de dados por meio da compressão de linhas, usando o conceito de agregação. Como aplicação, empregamos essa estratégia na análise de agrupamentos, fornecendo subsídios para entender como as dimensões estão relacionadas à qualidade dos agrupamentos gerados, o grau de estabilidade das instâncias de dados dentro dos agrupamentos e a relação entre agrupamentos. A seguir o processo de agregação e compactação da *TableLens* é detalhado.

4.2 Agregando linhas na *TableLens*

Para ilustrar a situação de perda da visão geral da *TableLens*, considere a Figura 20 como a representação completa de um conjunto de dados. Como pode ser observado, quantidade de dados apresentadas na visualização é relacionada à resolução vertical do dispositivo de saída. Por exemplo, na Figura 20 é visualizado o conjunto de dados de Qualidades de Vinho Branco (CORTEZ *et al.*, 1998), com 4898 instâncias e 11 dimensões. Em um dispositivo padrão, com resolução vertical de 1024 pixels, essa visualização é mostrada acompanhada por uma barra de rolagem para ser possível ver as linhas ocultas, levando a perda de contexto. O método proposto neste trabalho para escalar visualmente a *TableLens* está baseado na compressão de linhas adjacentes, representando-as em uma só linha usando agregação de dados. A Figura 19 ilustra o processo de compressão de linhas.

Figura 19 – Detalhamento do processo de compressão. (A) Linhas adjacentes, (B) Média dos valores compactados, (C) Subtração entre a média e o valor máximo, (D) Valor máximo (E) Desvio padrão.

Fonte: Elaborada pelo autor.

No processo de compressão, primeiro deve-se definir o número n de linhas adjacentes para agrupar. Este número n é definido pelo usuário. Em seguida, calcula-se a largura da linha comprimida. Para isso devemos saber como na *TableLens* são mapeadas os valores das instâncias. Os valores dos atributos contínuos são normalizados com relação ao maior valor do atributo, esse valor é representado por uma linha colorida com largura escalada à largura da coluna:

$$lw_i^k = \frac{a_i^k \times cw^k}{\max\{a_1^k, \dots, a_N^k\}} \quad (4.1)$$

onde cw^k é a largura da coluna referente ao atributo k , a_i^k é o valor da instância i em k e $\max\{a_1^k, \dots, a_n^k\}$ é o maior valor do atributo k . Para os dados discretos ou categóricos, a largura da coluna exibida na tela é dividida pela quantidade de categorias, para cada categoria estabelece-se uma cor diferente. Esta é representada por uma linha colorida que é posicionada na coluna dependendo do índice da categoria. Neste projeto não serão considerados os dados categóricos, por isso não aprofundaremos nesse tópico.

A Figura 19(A) ilustra a compressão de linhas como $n = 5$. O processo de agrupar as linhas é realizado para cada coluna com as mesmas instâncias adjacentes. Para cada grupo de n linhas selecionadas calcula-se o valor médio (equação 4.2), desvio padrão (equação 4.3) e o máximo valor do grupo. O desvio padrão é calculado com intuito de entender a distribuição dos dados comprimidos, isto é, o quanto as linhas comprimidas estão com valores similares ou diferentes. Com isto, identifica-se atributos com valores parecidos.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i \quad (4.2)$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2} \quad (4.3)$$

De forma a preservar valores espúrios, os maiores valores identificados para cada n linhas agrupadas no passo anterior definem a largura da linhas agregadas, as quais são coloridas em verde (Figura 19(B)). Depois, a média das n linhas é calculada e mapeada para a cor azul, caso a média seja menor do que o maior valor (pode ser igual) é calculada a subtração entre o maior valor e o valor da média (Figura 19(C)). Por fim, o desvio padrão é mapeado para um pixel vermelho escalado com respeito ao valor da média e posicionado desde o começo da linha. Se os valores do desvio padrão são baixos os pixels tendem a estar próximos ao começo da linha, caso contrário eles tendem mais para a direita da linha (Figura 19(E)).

A Figure 21 ilustra a nossa abordagem de escalabilidade visual para a *TableLens* usando o mesmo conjunto de dados de **Qualidade de Vinho Branco** da Figura 20. Em relação ao valor do desvio padrão, percebe-se que as linhas comprimidas adjacentes têm valores parecidos já que os pontos vermelhos estão alinhados no começo das colunas. Observa-se nessa visualização que os padrões visuais da *TableLens* originais são mantidos, mas que os dados são representados em uma única visualização sem perda de contexto. Além disso, alterar o valor n permite que a escalabilidade visual seja orientada para as necessidades do usuário. Por exemplo, se o usuário estabelecer uma quantidade de compressão de linha $n = 5$ e a altura do gráfico for $h = 900$, esta abordagem conseguiu representar $h * n = 4500$ instâncias de dados sem a necessidade de uma barra de rolagem e sem perder o contexto.

A seguir explica-se como usar essa ideia de compressão de linhas para auxiliar na análise visual de agrupamentos.

Figura 20 – TableLens para o conjunto de dados da **Qualidade de Vinho Branco** com 4898 instâncias e 11 atributos.

Fonte: Elaborada pelo autor.

Figura 21 – TableLens compactada para o conjunto de dados de **Qualidade de Vinho Branco**, comprimindo 9 linhas.

Fonte: Elaborada pelo autor.

4.3 Analisando Agrupamentos

Apesar do sucesso dos processos de agrupamento de dados, considera-se que resultados satisfatórios, a partir do ponto de vista de um determinado usuário, só serão atingidos se for possível avaliar o comportamento, importância ou influência dos dados e atributos nos resultados obtidos, bem como entender se os grupos são similares ou não similares (HALKIDI; VAZIRGIANNIS, 2002). Assim, desenvolvemos uma metodologia que possibilita a aplicação da metáfora compacta apresentada anteriormente na análise de agrupamentos. A ideia foi gerar um processo no qual o usuário possa interagir com a visualização para analisar o comportamento dos grupos, instâncias e dimensões nos agrupamentos. Para isso, utiliza-se como métrica de avaliação o coeficiente de silhueta (ROUSSEEUW, 1987). A seguir essa metodologia e processo são detalhados.

4.3.1 Metodologia de Integração com a TableLens

A motivação principal para nosso processo de integração visual para a compreensão do comportamento das técnicas de agrupamento está no fato de que diferentes modelos de agrupamento envolvem conceitos matemáticos abstratos e alguns desses podem ser bastante complexos. Na análise de agrupamento, entender quais são os atributos mais relevantes para os grupos formados potencialmente ajuda na compressão desses modelos. A Figura 22 ilustra a sequência de passos empregados para avaliar diferentes resultados de agrupamentos.

No primeiro passo (Figura 22(A)), os dados de entrada são pré-processados e organizados de forma tabular e, em seguida, grupos são formados usando diferentes algoritmos de agrupamento (Figura 22(B)). Depois, sobre o resultado obtido é computado o coeficiente de silhueta, considerando os atributos individualmente e considerando todo os atributos juntos (Figura 22(C)).

Figura 22 – Estrutura da metodologia de integração. Da esquerda para a direita, (A) Entrada do conjunto de dados, (B) Gerar diferentes resultados de agrupamento, (C) Aplicar índices de validação de agrupamentos. (D) Interpretação visual da validade do agrupamento.

Fonte: Elaborada pelo autor.

Figura 23 – TableLens compactada do conjunto de dados de Auto MPG comprimindo com $n = 9$. Em (a), representação das cores dos grupos. Em (b) o resultado da compactação e potencial para análise de um resultado de agrupamento.

Fonte: Elaborada pelo autor.

O Coeficiente de Silhueta foi selecionado como métrica de validação uma vez que esse mede tanto coesão quanto separação dos grupos (ARBELAITZ *et al.*, 2013; TAN; STEINBACH; KUMAR, 2005). Com a silhueta, nosso intuito foi permitir a identificação dos atributos mais relevantes para os agrupamentos formados. Por fim, no passo é gerada a visão tabular usando a técnica de visualização *TableLens* compacta de modo que o usuário possa interagir com a ferramenta por meio da ordenação, seleção de atributos e de agrupamentos (Figura 22(D)). Neste passo também é usado a técnica de visualização Coordenadas Paralelas (INSELBERG; DIMSDALE, 1990b) para auxiliar o usuário na interpretação de estabilidade, tendências e mudanças de dados nos grupos considerando diferentes resultados de agrupamento.

A Figura 23b apresenta a técnica *TableLens* modificada com o objetivo de apresentar os rótulos dos grupos e os valores dos índices de validação de agrupamento. Em cada coluna os

valores dos atributos são mostrados à esquerda (A). O valor do Coeficiente de Silhueta referente ao atributo é mostrado à direita (C). Os valores de silhueta são compactados considerando a maior quantidade de instâncias com o mesmo rótulo. Os valores de silhueta podem assumir valores positivos e negativos entre $[-1, 1]$. Para se construir a visualização, esses valores são normalizados entre $[0, 1]$, fazendo $S_n = |(S + 1)/2|$. Se o valor do Coeficiente de Silhueta for negativo, ele é mapeado para uma linha azul, caso contrário é mapeado para uma linha verde. Com o propósito de ter uma melhor interpretação dos grupos, cada grupo é mapeado para uma cor e no centro das colunas são mostradas as cores dos grupos a qual a instância pertence (B).

4.3.2 Análise Visual de Estabilidade

A fim de se comparar diferentes resultados de agrupamento, uma coluna extra é adicionada à visualização. Essa é dividida em partes representando cada resultado de agrupamento. Cada resultado define uma barra retangular mapeada lado a lado. Cada barra retangular é dividida por caixas coloridas que representam os rótulos dos grupos. Além disso, a última coluna contém o valor de silhueta de um determinado resultado de agrupamento selecionado pelo usuário considerando todos os atributos.

O conceito de barras retangulares mapeadas lado a lado já foi utilizado em diferentes trabalhos relacionados, por exemplo, na técnica de Visualização de Conjuntos Paralelos *Parallel set Visualization* (BENDIX; KOSARA; HAUSER, 2005) (Figura 24(A)). A Visualização de Conjuntos Paralelos especificamente foi desenvolvida para analisar dados categóricos se baseando na metáfora visual das Coordenadas Paralelas. Os atributos são representados por eixos paralelos, as instâncias são agrupadas em blocos para cada eixo. Essa visualização permite a interação do usuário mediante a ordenação dos eixos de forma independente. Um dos problemas apresentados na Visualização de Conjuntos Paralelos é o excessivo cruzamento de linhas. Por exemplo em (A), as instâncias foram ordenadas com base no eixo $K3$, porém, como a ordenação de cada eixo é independente, o eixo $K5$ apresenta ordenação diferente o que causa os cruzamentos. A nossa abordagem é realizada aninhando os resultados em uma coluna extra, a qual não apresenta problemas de desordem visual ou cruzamento de linhas entre as barras paralelas. A ideia é simplesmente ordenar os dados de acordo com alguma coluna que representa um agrupamento. A Figura 24(B) ilustra como a abordagem de organizar os rótulos no modelo tabular não apresenta cruzamento e possibilita uma melhor comparação de grupos formados por diferentes algoritmos/parametrizações, com mudanças de cor indicando mudanças de grupo entre dois agrupamentos diferentes.

Na Figura 24(B) é possível analisar a estabilidade dos grupos formados. A estabilidade dos grupos fundamenta-se em analisar se as instâncias de dados são mantidas no mesmo grupo ou mudam de grupo considerando diferentes processos de agrupamento (SEO; SHNEIDERMAN, 2002), permitindo analisar relações de similaridade entre dados.

Figura 24 – (A) Estabilidade de grupos usando cruzando linhas, (B) Estabilidade de grupos visualizados em layout tabular.

No processo de análise de estabilidade, a ferramenta permite ao usuário selecionar grupos de instâncias que mudam de rótulo, mostrando esses grupos por meio da técnica de Coordenadas Paralelas. A Figura 25 apresenta como nossa ferramenta integra a técnica de Coordenadas Paralelas na TableLens, alinhando os eixos das coordenadas paralelas com o início das colunas da TableLens (A). A mudança do rótulo das instâncias são representadas pelas cores. Nesse exemplo, percebe-se que o grupo foi dividido em dois considerando diferentes processos de agrupamento, um colorido em azul e outro em rosa, possibilitando a análise de qual(ais) atributos influenciam na divisão do grupo.

Figura 25 – Abordagem da Coordenadas Paralelas como método visual que auxilia ao usuário na interpretação de estabilidade e tendências dos grupos.

Em relação a visão fornecida, o usuário pode selecionar atributos para gerar um novo processo de agrupamento, tornando a análise de agrupamentos um processo iterativo auxiliado pelo usuário. Vale a pena ressaltar que o método proposto para a validação visual de grupos está baseado em vários princípios de validação de grupos, tais como: determinar a tendência de agrupamento, determinar o número de k grupos que melhor se ajusta ao conjunto de dados, avaliar quão bem os resultados de grupos se encaixam aos dados, determinar os dados mais relacionados em um conjunto de dados ao longo de diferentes processos de agrupamentos. Além disso, facilitar a comparação de diferentes agrupamentos obtidos por diferentes algoritmos. No próximo capítulo são apresentados os resultados obtidos na compressão de linhas e na análise de agrupamento utilizando-se diferentes conjuntos de dados como teste.

RESULTADOS E AVALIAÇÃO

5.1 Considerações Iniciais

A avaliação da técnica proposta é apresentada nesta seção, com validação conduzida por meio de resultados experimentais usando diferentes conjuntos de dados, e mensurada de acordo com a qualidade do processo de agregação (quantitativa) e a capacidade de suportar análise de agrupamentos (qualitativa). A expectativa é que tais visualizações sejam boas compressões da representação original da *TableLens*, preservando a distribuição original dos dados e mantendo valores espúrios. Além disso, a visualização proposta visa permitir, por exemplo, a compreensão do comportamento dos algoritmos de agrupamento em relação ao papel de atributos e da estabilidade dos grupos, mostrando como tais atributos se comportam ou influenciam os grupos gerados.

Três conjuntos de dados foram utilizados para essa análise: **Qualidade de Vinhos Brancos**; **HTRU2**; e **Auto MPG**. O conjunto **Qualidade de Vinhos Brancos** contém 4898 instâncias representando diferentes vinhos brancos produzidos no norte de Portugal. Cada instância tem 11 atributos (*fixed acidity*, *volatile acidity*, *citric acid*, *residual sugar*, *chlorides*, *free sulfur dioxide*, *total sulfur dioxide*, *density*, *pH*, *sulphates*, *alcohol*) (CORTEZ *et al.*, 2009; LICHMAN, 2013). O segundo conjunto, **HTRU2**, descreve uma amostra de corpos celestes candidatos a serem estrelas, contendo 17898 instâncias com 8 atributos (*Mean*, *Standard Deviation*, *Excess kurtosis*, *Skewness*, *Mean of the DM-SNR*, *Standard deviation of the DM-SNR*, *Excess kurtosis of the DM-SNR*, *Skewness of the DM-SNR*) (KEITH *et al.*, 2010; LICHMAN, 2013). Por fim, o conjunto **Auto MPG** contém 398 instâncias com 8 atributos (*mpg*, *cylinders*, *displacement*, *horsepower*, *weight*, *acceleration*, *model year*, *oring*, *car name*) descrevendo diferentes tipos de carros. No caso do conjunto **Auto MPG**, o atributo *car name* foi descartado por não ser contínuo nem categórico (QUINLAN, 1993; LICHMAN, 2013). A Tabela 1 sumariza essa informação.

Tabela 1 – Características dos conjuntos de dados utilizados nos experimentos realizados.

Conjunto de dados	Instâncias	Atributos
Qualidade de Vinho Branco	4898	11
HTRU2	17898	8
Auto MPG	398	8

5.2 Avaliação da Qualidade da Agregação

A técnica de compressão da *TableLens* pode ser validada considerando-se quão bem os dados agregados ou abstraídos representam os dados originais. Nesse contexto existem duas técnicas que avaliam o nível de preservação dos dados do conjunto original no conjunto agregado, conhecidas como *Histogram Difference Measure (HDM)* e *Nearest Neighbor Measure (NNM)* (CUI *et al.*, 2006). A HDM é baseada no conceito de histograma e tenta avaliar a diferença da distribuição entre dois conjuntos de dados mediante a minimização da densidade da distribuição. Assim, usamos a HDM para avaliar se a distribuição do conjunto original de dados é preservada no conjunto de dados agregados. A NNM, por sua vez, é utilizada para verificar a qualidade de vizinhança e distância entre os dados do conjunto original e os dados no conjunto agregado, pelo fato de que sua prioridade é avaliar a preservação e a mudança dos dados espúrios mediante a média normalizada de distâncias entre cada instância do conjunto original e sua representação no conjunto de dados abstraído. Ambas as medidas são normalizadas entre $[0, 1]$ e quanto maior o valor, melhor é o resultado de preservação da informação. Além dessas, a métrica Dal (CUI *et al.*, 2006) representa a porcentagem de dados agregados em relação ao conjunto original de dados.

Inicialmente os valores de HDM e NNM foram avaliados conforme o nível de compressão aumenta (n). A Figura 26 ilustra o conjunto de dados das qualidades de vinho sem comprimir linhas, enquanto as Figuras 28a e 29a com $n = \{20, 40\}$ respectivamente ilustram os resultados. Pode-se perceber que, enquanto os valores de linhas n aumentam, o HDM se mantém estável indicando que a distribuição e a densidade dos dados é mantida. Entretanto, o NNM apresenta um aumento (Figura 29a), indicando melhor preservação de dados espúrios. Mas isso acontece com uma compactação de 40 vezes dos dados originais e o mapeamento dos valores estatísticos ajudam a entender a informação perdida.

A mesma análise foi realizada para o conjunto de dados **HTRU2**. A Figura 30 apresenta os resultados para o conjunto **HTRU2** com $n = 20$ e a Figura 31 apresenta os mesmos dados com $n = 40$. Observe que nesses casos os valores de HDM e a NNM também se mantiveram em um patamar alto, indicando boa qualidade de compressão, mesmo para maiores valores de linhas comprimidas n . Visualmente essa qualidade também pode ser verificada já que o formato das colunas se mantém (padrão de desenho) e os dados espúrios são visíveis. Aqui não mostramos o formato do conjunto original uma vez que esse é muito grande, impossibilitando gerar uma

Figura 26 – TableLens na sessão de análise sem comprimir valores de dados usando o conjunto de dados de **Qualidade de Vinho Branco**.

Figura 27 – TableLens na sessão de análise comprimindo $n = 9$ linhas por unidade visual, usando o conjunto de dados de qualidade do vinho (A). (B) Validação.

Figura 28 – TableLens na sessão de análise comprimindo $n = 20$ linhas por unidade visual, usando o conjunto de dados de qualidade do vinho (A). (B) Validação.

Figura 29 – TableLens na sessão de análise comprimindo $n = 40$ linhas por unidade, usando o conjunto de dados de Qualidade de Vinho Branco (A). (B) Validação.

imagem contendo o mesmo em sua totalidade.

Figura 32 – TableLens na sessão de análise de agrupamento sem comprimir valores de dados, usando o conjunto de dados de **Auto MPG**.

Figura 33 – TableLens na sessão de análise comprimindo $n = 10$ linhas ordenada de forma decendente por *mpg*, usando o conjunto de dados de **Auto MPG** (A). (B) Validação.

Figura 30 – TableLens na sessão de análise comprimindo $n = 20$ linhas e ordenada pela coluna DM_stdev do conjunto de dados de **HTRU2** (A). (B) Validação.

Figura 31 – TableLens na sessão de análise comprimindo $n = 40$ linhas e ordenada pela coluna DM_stdev do conjunto de dados de **HTRU2** (A). (B) Validação.

A nova metáfora para escalabilidade visual da *TableLens* permite, por meio do uso de médias simples, uma visão representativa do conjunto de dados e uma melhor visão geral. Além da representação da média, a definição do tamanho da linha como o maior valor dentro de um grupo de linhas comprimidas permite a identificação de valores espúrios de uma maneira simples e compacta. Além disso, representa dados em um layout tabular sem perder o contexto geral do conjunto de dados. Uma maneira tradicional de organizar os dados nas visualizações tabulares é ordenar os dados pelos valores dos atributos. Por meio do clique nas colunas, a *TableLens* ordena os valores dos atributos agrupando valores similares, colocando-os em linhas adjacentes como é mostrado na Figura 32. Apesar do método de compressão de linhas não estabelecer explicitamente uma forma de permutação para obter uma melhor compressão, considera-se que o usuário possa fazer uso da ordenação de colunas para comprimir linhas com valores parecidos. Um exemplo desse caso é ilustrado na Figura 33, onde o conjunto de dados **Auto MPG** é compactado com $n = 10$ e ordenado pelo atributo *mpg* de forma descendente, fazendo com que o desvio padrão seja mapeado para o começo da coluna. Isso indica que os valores desse atributo são parecidos e podem ser agrupados, o que não prejudica a tendência do atributo. No entanto, para as outras colunas, o desvio padrão é mostrado com maior relevância (pontos vermelhos), facilitando inferir a relação direta entre instâncias se os pontos não estiverem desordenados. Assim como ordenar pode auxiliar na identificação de instâncias com alta relação, ordenar as colunas também pode melhorar ou prejudicar as métricas de avaliação NNM e HDM.

A subtração entre o maior valor do grupo de linhas comprimidas e a média facilita a compreensão da distribuição de dados como a principal característica na exploração de dados. Grandes faixas verdes nas linhas indicam que os dados estão espalhados com respeito à média (linhas azuis). Esta característica pode auxiliar na busca de dados peculiares ou dados que estão fora de uma distribuição normal para o atributo. A homogeneidade entre as larguras das linhas azuis permitirá identificar os valores abaixo e acima da média. Um exemplo desse padrão pode-se perceber na Figura 31 do conjunto de dados de **HTRU2**, em que a coluna *Profile_mean* mostra que esse atributo tem a maioria dos valores parecidos, e as últimas linhas mostram uma variação dos valores com respeito à média, indicando heterogeneidade das linhas formadas. Por outro lado, pequenas faixas verdes indicam que a maioria dos valores são parecidos em relação à média do atributo. Assim, informações como média, desvio padrão, maior valor, subtração entre a media e o maior valor são úteis para especificar a diferença entre os dados e como os dados se comportam em cada atributo. Baseando-se nas avaliações feitas, é importante ressaltar que a abordagem proposta consegue escalar visualmente quarenta (40) vezes mais que a *TableLens* original, permitindo assim representar maior quantidade de dados sem perder o contexto.

5.3 Análise de Grupos

Diferentes processos de agrupamento podem gerar diferentes rótulos para as instâncias, tanto que comparar diferentes agrupamentos sempre é uma tarefa subjetiva dependente do

analista (TAN; STEINBACH; KUMAR, 2005; HASTIE; TIBSHIRANI; FRIEDMAN, 2009b). Nesse contexto, identificar o algoritmo de agrupamento mais apropriado para um conjunto de dados tornou-se um desafio. Embora existam diferentes metáforas para analisar agrupamentos, perguntas como “porque os grupos foram gerados?” ou “qual a influência de um atributo no grupo ou agrupamento?” ainda não foram apropriadamente respondidas.

Portanto, com o intuito de responder essas perguntas, o coeficiente de silhueta foi utilizado como métrica de avaliação nesse projeto. Uma das características do coeficiente de silhueta é determinar o número k de agrupamentos para um conjunto de dados (TAN; STEINBACH; KUMAR, 2005), indicando o quão coesos e separados são os grupos gerados. Assim, surgiu a ideia de desenvolver uma metáfora visual que permita a análise dos agrupamentos em termos de atributos e instâncias nos grupos, baseado-se nos valores do Coeficiente de Silhueta. Essa metáfora visual possibilita identificar o número de grupos k que ajustam-se melhor ao conjunto de dados, assim como uma rápida compreensão dos atributos em cada grupo. Por exemplo, as Figuras 34 e 35 mostram uma característica da abordagem proposta, que é facilitar a compreensão da relação entre atributos ao ordenar as colunas pelo valor do coeficiente de silhueta de cada atributo. Desta forma, o usuário pode observar quais são os atributos com os mesmos comportamentos (correlacionados) segundo os valores de silhueta obtidos em cada processo de agrupamento.

Ainda nesse contexto, as Figuras 34 e 35 ilustram a metáfora visual aplicada ao conjunto de dados **Auto MPG**. Os valores de silhueta mostrados nessas visualizações correspondem a agrupamentos obtidos pelo algoritmo KMeans com $k = 3$. Mediante a ordenação das colunas *mpg* e *horsepower*, percebe-se que a maioria dos dados no grupo de cor verde possui uma boa coesão e separação com o atributo. Além disso, mostra-se que ambos atributos tem uma relação com respeito nos valores de silhueta, com isso, poderia entender-se que para o agrupamento desses atributos tem a mesma relevância nesse grupo. Portanto, levando em consideração esta relevância, a metáfora poderia auxiliar o usuário na tarefa de seleção de atributos, melhorando a geração de grupos. Quando a ordenação é realizada pela coluna *acceleration*, a Figura 36 ilustra que o atributo não possui uma relevância ou uma relação com o grupo de cor verde, mas sim para algumas instâncias do grupo de cor azul claro.

Figura 34 – Ordenação da coluna *mpg* segundo os valores de silhueta, permitindo ao usuário entender a relação dos atributos que melhor ajustam-se ao conjunto de dados, usando *Auto MPG* com KMeans $k = 3$.

Fonte: Elaborada pelo autor.

Figura 35 – Ordenação da coluna *horsepower* segundo os valores de silhueta, permitindo ao usuário entender a relação dos atributos que melhor ajustam-se ao conjunto de dados, usando o *Auto MPG* com KMeans $k = 3$.

Fonte: Elaborada pelo autor.

Figura 36 – Ordenação da coluna *acceleration* segundo os valores de silhueta, permitindo ao usuário entender a relação dos atributos com respeito do grupo, usando o *Auto MPG* com *KMeans* $k = 3$.

Fonte: Elaborada pelo autor.

Mais uma característica acrescentada neste projeto de mestrado foi permitir analisar de forma tabular a estabilidade de grupos mediante a ordenação dos rótulos dos agrupamentos. Como dito anteriormente, a análise de estabilidade tem o intuito de visualizar e entender como os dados se mantêm juntos no decorrer de diferentes processos de agrupamento. A Figura 37(A) ilustra um caso da análise de estabilidade usando o algoritmo de agrupamento *KMeans* com diferentes números de grupo $k = \{3, 5, 7, 10\}$. Os rótulos foram ordenados clicando em cada execução do agrupamento desde $k = 3$ até $k = 10$. Embora este modelo não revele muitas ideias sobre tendências nos dados, as linhas que representam os valores de silhueta facilitam o entendimento desses. Nota-se que os valores de silhueta estão positivos na maioria das colunas, significando que os atributos tem influência positiva na separação e na coesão. Por exemplo, a Figura 37 mostra a ordenação por $k - means10$ com $k = 10$, em que o algoritmo colocou as instâncias em um grupo diferente, mas as manteve juntas. Sendo assim, verifica-se que o motivo principal da similaridade acontece devido aos atributos *cylinders*, *displacement*, *horsepower*.

Quando o usuário deseja ter detalhes de algum grupo, a ferramenta permite selecionar grupos individualmente. Com o objetivo de aumentar o poder analítico, o grupo selecionado é então visualizado usando coordenadas paralelas (Figuras 38 e 39), mostrando a visão geral do conjunto selecionado e dando maior destaque para o grupo de instâncias selecionadas pelo usuário. Assim, tem-se um contexto geral do conjunto e é possível entender melhor a razão de a coesão e a separação dos grupos serem influenciadas por certos atributos. Também, a análise de estabilidade de grupo apresentada na visualização de coordenadas paralelas justifica-se por permitir a análise de similaridade dos dados considerando os atributos. A Figura 38 ilustra a análise de estabilidade para o conjunto de dado **Auto MPG**, em que as instancias com maior relação têm uma tendência direta entre o atributo *mpg* e *cylinders*. Por outro lado, na Figura 39 é mostrado a análise de estabilidade para o conjunto de dados **Qualidade de Vinho Branco** comprimindo 9 linhas. A compreensão dos vinhos mais relacionados é facilitada com o uso de coordenadas paralelas, permitindo inferir que os vinhos com valores menores tendem a estar

Figura 37 – Estabilidade de grupos para mostrar os dados mais relacionados através de diferentes agrupamentos, usando $KMeans(k = 3, 5, 7, 10)$. (A) Grupos de dados estáveis no aumento consecutivo de k para diferentes processos de agrupamento usando conjunto de dados **Auto MPG**

Fonte: Elaborada pelo autor.

Figura 38 – Aplicação de Coordenadas Paralelas para auxiliar na visão de estabilidade dos grupos e dados mais relacionados.

Fonte: Elaborada pelo autor.

mais relacionados nos primeiros 6 atributos durante diferentes processos de agrupamento.

Figura 39 – Estabilidade de grupos para mostrar os dados mais relacionados através de diferentes agrupamentos, usando $KMeans(k = 3, 5, 7, 10)$. (A) Grupos de dados estáveis no aumento consecutivo de k para diferentes processos de agrupamento usando o conjunto de dados de **Qualidade de Vinho Branco**

Fonte: Elaborada pelo autor.

Mediante a ordenação dos rótulos ao respeito dos grupos de cada agrupamento, a metáfora visual facilita visualizar e interpretar os dados que mudaram de rótulo. Sendo assim, o uso do layout tabular facilita comparar diferentes resultados de algoritmos e visualizar porque as instâncias mudaram de grupo. As Figuras 40 e 42 mostram os resultados obtidos comparando resultados dos algoritmos $KMeans$ e $Expectation Maximization$ usando como parâmetro $k = 3$ para ambos. Por exemplo, na Figura 40 é ilustrado a ordenação pelos rótulos do agrupamento $KMeans$ o qual esta indicado pela seta na última coluna. Considera-se que os valores de coeficiente de silhueta em cada coluna auxiliam na compreensão dessas mudanças. Na Figura 40, as instâncias marcadas em (A), que são rotuladas de forma diferente pelos dois algoritmos, apresentam valores positivos na maioria das colunas da silhueta, mas para as colunas *mpg*, *weight* e *cylinder*, os valores tende a ser negativos.

Com o propósito de entender essa mudança, o usuário pode interagir com a visualização selecionando o grupo de instâncias no qual ocorre a mudança de rótulo, sendo auxiliada pela visualização de coordenadas paralelas apresentada na Figura 41(A)), a qual mostra que os dados que mudaram de rótulos são bastante diferentes dos demais, os valores para esses considerando os atributos *weight* e *cylinder* estão separados da maioria dos valores no grupo. Portanto, o algoritmo EM tende agrupar melhor os dados que estão mais coesos em relação ao próprio grupo, enquanto $KMeans$ é mais suscetível a criar grupos mais separados. Vale lembrar que o $KMeans$ é um algoritmo de agrupamento divisivo baseado em um modelo de partição dos dados segundo a métrica de distância empregada (JAIN, 2010), enquanto, o EM é baseado em

Figura 40 – Comparação dos agrupamentos obtidos por KMeans (KM) e Expectation-Maximization (EM) com $k = 3$ usando o conjunto de dados **Auto MPG**.

Fonte: Elaborada pelo autor.

Figura 41 – Coordenada Paralelas para auxiliar na comparação dos agrupamentos, conjunto de dados **Auto MPG**

Fonte: Elaborada pelo autor.

modelos probabilísticos de estimação da distribuição dos dados para gerar grupos (AGGARWAL; REDDY, 2013).

Em relação a esses modelos de separação e probabilidade usados pelos algoritmos *KMeans* e EM respectivamente, as Figuras 42 e 43 possibilitam reforçar as conclusões expostas anteriormente sobre o relacionamento dos dados e os atributos com respeito ao agrupamento, nas quais o conjunto **Qualidade de Vinho Branco** foi utilizado com maior quantidade de dados. Na Figura 42(A), as instancias do grupo verde gerado por EM tendem a ter os maiores valores.

Figura 42 – Comparação dos agrupamentos obtidos por *KMeans* (KM) e *Expectation-Maximization* (EM) com $k = 3$ usando o conjunto de dados **Qualidade de Vinho Branco**.

Fonte: Elaborada pelo autor.

Figura 43 – Coordenada Paralelas para auxiliar na comparação dos agrupamentos, conjunto de dados **Qualidade de Vinho Branco**

Fonte: Elaborada pelo autor.

CONCLUSÕES

A *TableLens* é uma técnica de visualização de informação bastante popular que mapeia para o espaço visual dados tabulares, também chamados de multivariados ou multidimensionais, e possibilita a exploração e análise das correlações entre diferentes atributos e a identificação de elementos espúrios. Apesar dessa popularidade, o número de dados que podem ser analisados sem perda de contexto está relacionado à resolução vertical do dispositivo de saída, limitando a escalabilidade visual da técnica.

Nesse contexto, neste trabalho de mestrado foi proposta e desenvolvida uma nova metáfora visual baseada na técnica *TableLens* que permite a análise de maiores quantidades de dados por meio de uma estratégia de agregação visual. A nova metáfora permite que o usuário controle o nível de abstração da visualização gerada controlando o nível de agregação desejado. A ideia é mapear várias linhas da *TableLens* em uma única linha apresentando no modelo compacto informação como média e desvio padrão, resultando em uma visualização que não somente preserva a distribuição original dos dados mas que também permite a identificação de dados espúrios.

A metáfora desenvolvida foi avaliada em termos da preservação de vizinhança e distribuição dos dados (quantitativo). Os resultados mostraram que nossa abordagem preserva bem a vizinhança, a distância entre os dados, a densidade e a distribuição. Mesmo para valores altos de compactação os resultados foram bastante satisfatórios, sendo capaz representar grandes quantidade de dados – em nossos testes até 40 vezes mais dados sem perder muita informação.

Empregamos a nova metáfora produzida no contexto de análise de técnicas de agrupamento. A ideia foi prover mecanismos para analisar como cada atributo dos dados influencia no resultado de algoritmos de agrupamento em termos da coesão e separação dos grupos formados. Além disso, mecanismos para analisar a estabilidade dos grupos formados são providos, facilitando o entendimento de diferentes modelos de agrupamento. Nos testes realizados, comparamos os algoritmos de agrupamento *K-Means* e *Expectation Maximization* e diferentes parametrizações

desses algoritmos, e a técnica se mostrou bastante útil (qualitativamente) para o entendimento das diferenças entre os modelos gerados.

6.1 Limitações e trabalhos futuros

No tocante a criação de uma aplicação interativa, o tempo para gerar novos agrupamentos é uma limitação. Não somente a aplicação dos algoritmos de agrupamento é demorada, mas também o cálculo dos valores de silhueta são dispendiosos. Além disso, com relação à análise de agrupamento, os valores de silhuetas mapeados correspondem a um só agrupamento selecionado pelo usuário. Na estratégia apresentada não é possível fazer uma comparação entre os diferentes valores de silhueta para diferentes agrupamentos. Como trabalho futuro, considera-se aprimorar a visualização para comparar também diferentes valores de silhueta para diferentes agrupamentos.

A metáfora proposta somente mapeia dados contínuos e categóricos. Novas abordagens para mapear dados textuais devem ser buscadas em trabalhos futuros. Além disso, como trabalho futuro fica também a realização de testes formais com usuários para verificar se o que se consegue extrair, em termos de padrões visuais, da *TableLens* original se consegue extrair da nossa abordagem compacta. Apesar dos testes quantitativos indicarem que a distribuição dos dados e a identificação de elementos espúrios são preservados, somente testes com usuários finais podem corroborar a eficácia da nossa abordagem de compactação.

REFERÊNCIAS

AGGARWAL, C. C.; REDDY, C. K. **Data clustering: algorithms and applications**. [S.l.]: Chapman and Hall/CRC, 2013. Citado nas páginas 39, 41 e 81.

ARBELAITZ, O.; GURRUTXAGA, I.; MUGUERZA, J.; PÉREZ, J. M.; PERONA, I. An extensive comparative study of cluster validity indices. **Pattern Recognition**, v. 46, n. 1, p. 243 – 256, 2013. ISSN 0031-3203. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S003132031200338X>>. Citado nas páginas 41 e 66.

ARORA, S.; CHANA, I. A survey of clustering techniques for big data analysis. **5th International Conference on Confluence 2014 - The Next Generation Information Technology Summit**, p. 59–65, 2014. Disponível em: <<http://www.scopus.com/inward/record.url?eid=2-s2.0-84914141572{%&}partnerID=40{%&}md5=6b56c138d99fff678b0f6d03af>>. Citado na página 39.

BENDIX, F.; KOSARA, R.; HAUSER, H. Parallel sets: visual analysis of categorical data. In: **IEEE Symposium on Information Visualization, 2005. INFOVIS 2005**. [S.l.: s.n.], 2005. p. 133–140. ISSN 1522-404X. Citado na página 67.

BINGHAM, E.; MANNILA, H. Random projection in dimensionality reduction: Applications to image and text data. In: **Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY, USA: ACM, 2001. (KDD '01), p. 245–250. ISBN 1-58113-391-X. Disponível em: <<http://doi.acm.org/10.1145/502512.502546>>. Citado na página 33.

CARD, S. K.; MACKINLAY, J. D.; SHNEIDERMAN, B. (Ed.). **Readings in information visualization: using vision to think**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999. ISBN 1-55860-533-9. Citado na página 25.

CHALMERS, M. A linear iteration time layout algorithm for visualising high-dimensional data. In: **VIS '96: Proceedings of the 7th conference on Visualization '96**. Los Alamitos, CA, USA: IEEE Computer Society Press, 1996. p. 127–ff. ISBN 0-89791-864-9. Citado na página 32.

CHAN, W. W.-Y. A survey on multivariate data visualization. **Department of Computer Science and Engineering, Hong Kong University of Science and Technology**, v. 8, n. 6, p. 1–29, 2006. Citado na página 33.

CHUAH, M. C. Dynamic aggregation with circular visual designs. In: **Information Visualization, 1998. Proceedings. IEEE Symposium on**. [S.l.: s.n.], 1998. p. 35–43, 151. Citado nas páginas 45 e 46.

CLEVELAND, W. S. **Visualizing Data Why do Data Visualization ?** [s.n.], 1993. v. 36. 314 p. ISSN 00401706. ISBN 9780596514556. Disponível em: <<http://www.amazon.com/Visualizing-Data-William-S-Cleveland/dp/0963488406>>. Citado na página 25.

CORTEZ, P.; CERDEIRA, A.; ALMEIDA, F.; MATOS, T.; REIS, J. Modeling wine preferences by data mining from physicochemical properties. **Decision Support Systems**, v. 47, n. 4, p. 547–553, 1998. Citado na página 62.

_____. Modeling wine preferences by data mining from physicochemical properties. **Decis. Support Syst.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 47, n. 4, p. 547–553, nov. 2009. ISSN 0167-9236. Disponível em: <<http://dx.doi-org.ez67.periodicos.capes.gov.br/10.1016/j.dss.2009.05.016>>. Citado na página 69.

CUI, Q.; WARD, M.; RUNDENSTEINER, E.; YANG, J. Measuring data abstraction quality in multiresolution visualizations. **IEEE Transactions on Visualization and Computer Graphics**, v. 12, n. 5, p. 709–716, Sept 2006. ISSN 1077-2626. Citado nas páginas 32, 61 e 70.

DEMŠAR, J.; CURK, T.; ERJAVEC, A.; GORUP, Č.; HOČEVAR, T.; MILUTINVIČ, M.; MOŽINA, M.; POLAJNAR, M.; TOPLAK, M.; STARIČ, A.; ŠTAJDOHAR, M.; UMEK, L.; ŽAGAR, L.; ŽBONTAR, J.; ŽITNIK, M.; ZUPAN, B. Orange: Data mining toolbox in python. **J. Mach. Learn. Res.**, JMLR.org, v. 14, n. 1, p. 2349–2353, jan. 2013. ISSN 1532-4435. Disponível em: <<http://dl.acm.org.ez67.periodicos.capes.gov.br/citation.cfm?id=2567709.2567736>>. Citado na página 35.

DING, H.; WANG, C.; HUANG, K.; MACHIRAJU, R. iGPSe: a visual analytic system for integrative genomic based cancer patient stratification. **BMC bioinformatics**, v. 15, n. 1, p. 203, jul 2014. ISSN 1471-2105. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/25000928http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-203http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4227100>>. Citado nas páginas 51 e 53.

ELMQVIST, N.; FEKETE, J.-D. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. **IEEE Transactions on Visualization and Computer Graphics**, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 16, n. 3, p. 439–454, maio 2010. ISSN 1077-2626. Disponível em: <<http://dx.doi.org/10.1109/TVCG.2009.84>>. Citado nas páginas 36 e 50.

ESTIVILL-CASTRO, V. Why so many clustering algorithms: a position paper. **ACM SIGKDD Explorations Newsletter**, v. 4, n. 1, p. 65–75, 2002. ISSN 19310145. Citado nas páginas 26 e 38.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. Inteligência artificial: Uma abordagem de aprendizado de máquina. **Livros Técnicos e Científicos**, 2011. Citado nas páginas 36 e 37.

FALOUTSOS, C.; LIN, K.-I. Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. **SIGMOD Rec.**, ACM, New York, NY, USA, v. 24, n. 2, p. 163–174, maio 1995. ISSN 0163-5808. Disponível em: <<http://doi.acm.org/10.1145/568271.223812>>. Citado na página 32.

FUA, Y.-H.; WARD, M. O.; RUNDENSTEINER, E. A. Hierarchical parallel coordinates for exploration of large datasets. In: **Proceedings of the IEEE Visualization 1999 (VIS'99)**. Los Alamitos, CA, USA: IEEE Computer Society Press, 1999. p. 43–50. ISBN 0-7803-5897. Citado nas páginas 46 e 47.

HALKIDI, M.; VAZIRGIANNIS, M. Clustering validity assessment using multi representatives. **Proceedings of SETN Conference**, ..., n. April, p. 237–248, 2002. Disponível em: <http://lps.csd.auth.gr/setn02/poster{_}papers/237>. Citado na página 65.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning: data mining, inference and prediction**. 2. ed. Springer, 2009. Disponível em: <<http://www-stat.stanford.edu/~tibs/ElemStatLearn/>>. Citado na página 26.

_____. **The elements of statistical learning: data mining, inference and prediction**. 2. ed. Springer, 2009. Disponível em: <<http://www-stat.stanford.edu/~tibs/ElemStatLearn/>>. Citado nas páginas 36, 37, 38 e 76.

HOFFMAN, P.; GRINSTEIN, G.; MARX, K.; GROSSE, I.; STANLEY, E. DNA visual and analytic data mining. **Proceedings Visualization '97 (Cat. No. 97CB36155)**, p. 437–441, 1997. ISSN 1. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=663916>>. Citado nas páginas 25 e 31.

HOFFMAN, P. E. **Table Visualizations: A Formal Model and Its Applications**. Tese (Doutorado), 2000. AAI9950455. Citado na página 34.

INGRAM, S.; MUNZNER, T.; OLANO, M. Glimmer: Multilevel mds on the gpu. **IEEE Transactions on Visualization and Computer Graphics**, v. 15, n. 2, p. 249–261, March 2009. ISSN 1077-2626. Citado na página 33.

INSELBERG, A.; DIMSDALE, B. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: . [S.l.: s.n.], 1990. p. 361–378. Citado nas páginas 25 e 33.

_____. Parallel coordinates: a tool for visualizing multi-dimensional geometry. In: . [S.l.: s.n.], 1990. p. 361–378. Citado na página 66.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, Elsevier B.V., v. 31, n. 8, p. 651–666, 2010. ISSN 01678655. Disponível em: <<http://dx.doi.org/10.1016/j.patrec.2009.09.011>>. Citado na página 80.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: A review. **ACM Comput. Surv.**, ACM, New York, NY, USA, v. 31, n. 3, p. 264–323, set. 1999. ISSN 0360-0300. Disponível em: <<http://doi.acm.org.ez67.periodicos.capes.gov.br/10.1145/331499.331504>>. Citado na página 37.

JERDING, D. F.; STASKO, J. T. The information mural: A technique for displaying and navigating large information spaces. **IEEE Transactions on Visualization and Computer Graphics**, v. 4, n. 3, p. 257–271, 1998. ISSN 10772626. Citado na página 61.

JOHNSON, B. S. **Treemaps: Visualizing Hierarchical and Categorical Data**. Tese (Doutorado), College Park, MD, USA, 1993. UMI Order No. GAX94-25057. Citado na página 45.

JOIA, P.; COIMBRA, D.; CUMINATO, J. A.; PAULOVICH, F. V.; NONATO, L. G. Local affine multidimensional projection. **IEEE Transactions on Visualization and Computer Graphics**, v. 17, n. 12, p. 2563–2571, Dec 2011. ISSN 1077-2626. Citado na página 33.

JOLLIFFE, I. T. **Principal Component Analysis**. [S.l.]: Springer-Verlag, 1986. Citado na página 33.

KEIM, D.; ANDRIENKO, G.; FEKETE, J. D.; GÖRG, C.; KOHLHAMMER, J.; MELANÇON, G. Visual analytics: Definition, process, and challenges. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 4950 LNCS, p. 154–175, 2008. ISSN 03029743. Citado na página 30.

KEIM, D. a. Pixel-oriented visualization techniques for exploring very large databases. **Journal of Computational and Graphical Statistics**, v. 5, n. 1, p. 58–77, 1996. ISSN 10618600. Disponível em: <[http://links.jstor.org/sici?sici=1061-8600\(199603\)5:1<58:PVTFEV>2.0.CO;2-P](http://links.jstor.org/sici?sici=1061-8600(199603)5:1<58:PVTFEV>2.0.CO;2-P)>. Citado na página 25.

_____. Designing pixel-oriented visualization techniques: theory and applications. **IEEE Transactions on Visualization and Computer Graphics**, v. 6, n. 1, p. 59–78, 2000. ISSN 10772626. Citado na página 25.

KEIM, D. A.; MANSMANN, F.; SCHNEIDEWIND, J.; ZIEGLER, H. Challenges in visual data analysis. In: **Information Visualization, 2006. IV 2006. Tenth International Conference on**. [S.l.: s.n.], 2006. p. 9–16. ISSN 1550-6037. Citado na página 25.

Keim Daniel, K. J.; MANSMANN, G. r. E.; FLORIAN. **Mastering the Information Age Solving Problems with Visual Analytics**. [s.n.], 2010. 57 – 86 p. ISSN 1877-0509. ISBN 978-3-905673-77-7. Disponível em: <<http://diglib.org>>. Citado nas páginas 25, 29 e 50.

KEITH, M. J.; JAMESON, A.; STRATEN, W. van; BAILES, M.; JOHNSTON, S.; KRAMER, M.; POSSENTI, A.; BATES, S. D.; BHAT, N. D. R.; BURGAY, M.; BURKE-SPOLAOR, S.; D’AMICO, N.; LEVIN, L.; MCMAHON, P. L.; MILIA, S.; STAPPERS, B. W. The high time resolution universe pulsar survey – i. system configuration and initial discoveries. **Monthly Notices of the Royal Astronomical Society**, v. 409, n. 2, p. 619–627, 2010. Disponível em: <<http://dx.doi.org/10.1111/j.1365-2966.2010.17325.x>>. Citado na página 69.

LEX, A.; STREIT, M.; PARTL, C.; KASHOFER, K.; SCHMALSTIEG, D. Comparative Analysis of Multidimensional, Quantitative Data. **Visualization and Computer Graphics, IEEE Transactions on**, v. 16, n. 6, p. 1027–1035, 2010. ISSN 1077-2626. Citado nas páginas 56 e 57.

LICHMAN, M. **UCI Machine Learning Repository**. 2013. Disponível em: <<http://archive.ics.uci.edu/ml>>. Citado na página 69.

LOHNINGER, H. **INSPECT: A program system to visualize and interpret chemical data**. 1994. 147–153 p. Citado nas páginas 25 e 34.

LU, L. F.; HUANG, M. L.; CHEN, Y. W.; LIANG, J.; NGUYEN, Q. V. Clutter reduction in multi-dimensional visualization of incomplete data using sugiyama algorithm. In: **2012 16th International Conference on Information Visualisation**. [S.l.: s.n.], 2012. p. 93–99. ISSN 1550-6037. Citado nas páginas 34 e 45.

L’YI, S.; KO, B.; SHIN, D.; CHO, Y.-J.; LEE, J.; KIM, B.; SEO, J. Xclusim: a visual analytics tool for interactively comparing multiple clustering results of bioinformatics data. **BMC Bioinformatics**, v. 16, n. 11, p. S5, 2015. ISSN 1471-2105. Disponível em: <<http://dx.doi.org/10.1186/1471-2105-16-S11-S5>>. Citado nas páginas 51 e 52.

MAMANI, G. M. H. **Empregando técnicas de projeção multidimensional para transformação interativa de espaços de características**. Dissertação (Mestrado), 2012. Citado na página 32.

MUNZNER, T.; MAGUIRE, E. **Visualization analysis and design**. Boca Raton, FL: CRC Press, 2015. (AK Peters visualization series). Disponível em: <<https://cds.cern.ch/record/2001992>>. Citado nas páginas 26, 35 e 61.

ONO, J. H. P. **Visualização de similaridades em bases de dados de música**. Dissertação (Mestrado) — Instituto de Ciências Matemáticas e de Computação, São Carlos, Universidad de São Paulo, <<http://www.teses.usp.br/teses/disponiveis/55/55134/tde-15092015-104008/>>, 2015. Citado na página 32.

PAGLIOSA, P.; MARTINS, R. M.; CEDRIM, D.; PAIVA, A.; MINGHIM, R.; NONATO, L. G. Mist: Multiscale information and summaries of texts. In: **2013 XXVI Conference on Graphics, Patterns and Images**. [S.l.: s.n.], 2013. p. 91–98. ISSN 1530-1834. Citado nas páginas 49 e 50.

PAIVA, J. G. d. S. **Técnicas computacionais de apoio à classificação visual de imagens e outros dados**. Tese (Doutorado) — Universidade de São Paulo, São Carlos, dec 2012. Disponível em: <<http://www.teses.usp.br/teses/disponiveis/55/55134/tde-02042013-084718/>>. Citado na página 29.

PAULOVICH, F. V. **Mapeamento de dados multi-dimensionais - integrando mineração e visualização**. 1–144 p. Tese (Doutorado) — Universidade de São Paulo, São Carlos, oct 2008. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-04032009-145018/publico/tese{_}paulovich.pdf<http://www.teses.usp.br/teses/disponiveis/55/55134/tde-04032009-1450>>. Citado na página 31.

PAULOVICH, F. V.; ELER, D. M.; POCO, J.; BOTHA, C. P.; MINGHIM, R.; NONATO, L. G. Piecewise Laplacian-based Projection for Interactive Data Exploration and Organization. **Computer Graphics Forum**, The Eurographics Association and Blackwell Publishing Ltd., 2011. ISSN 1467-8659. Citado nas páginas 32 e 33.

PAULOVICH, F. V.; MINGHIM, R. Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 6, p. 1229–1236, Nov 2008. ISSN 1077-2626. Citado nas páginas 48 e 49.

PAULOVICH, F. V.; NONATO, L. G.; MINGHIM, R.; LEVKOWITZ, H. Least square projection: a fast high precision multidimensional projection technique and its application to document mapping. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 3, p. 564–575, May/Jun 2008. Citado na página 33.

PAULOVICH, F. V.; SILVA, C. T.; NONATO, L. G. Two-phase mapping for projecting massive data sets. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 16, n. 6, p. 1281–1290, 2010. Citado na página 33.

QUINLAN, J. R. Combining instance-based and model-based learning. In: . [S.l.]: Morgan Kaufmann, 1993. p. 236–243. Citado na página 69.

RAO, R.; CARD, S. K. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In: **CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems**. New York, NY, USA: ACM, 1994. p. 318–322. ISBN 0-89791-650-6. Citado nas páginas 25, 34 e 61.

_____. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In: **CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems**. New York, NY, USA: ACM, 1994. p. 318–322. ISBN 0-89791-650-6. Citado na página 26.

ROUSSEEUW, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **J. Comput. Appl. Math.**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 20, n. 1, p. 53–65, nov. 1987. ISSN 0377-0427. Disponível em: <[http://dx.doi.org.ez67.periodicos.capes.gov.br/10.1016/0377-0427\(87\)90125-7](http://dx.doi.org.ez67.periodicos.capes.gov.br/10.1016/0377-0427(87)90125-7)>. Citado na página 65.

SAMMON, J. W. A nonlinear mapping for data structure analysis. **IEEE Trans. Comput.**, IEEE Computer Society, Washington, DC, USA, v. 18, n. 5, p. 401–409, maio 1969. ISSN 0018-9340. Disponível em: <<http://dx.doi.org/10.1109/T-C.1969.222678>>. Citado na página 32.

SANTOS, J. A. dos; FERREIRA, C. D.; TORRES, R. d. S.; GONCALVES, M. A.; LAMPA-RELLI, R. A. C. A relevance feedback method based on genetic programming for classification of remote sensing images. **Inf. Sci.**, Elsevier Science Inc., New York, NY, USA, v. 181, n. 13, p. 2671–2684, jul. 2011. ISSN 0020-0255. Disponível em: <<http://dx.doi.org.ez67.periodicos.capes.gov.br/10.1016/j.ins.2010.02.003>>. Citado na página 29.

SEO, J.; SHNEIDERMAN, B. Interactively exploring hierarchical clustering results. **Computer**, v. 35, n. 7, 2002. ISSN 00189162. Citado nas páginas 58 e 67.

SHARKO, J.; GRINSTEIN, G.; MARX, K. A. Vectorized radviz and its application to multiple cluster datasets. **IEEE Transactions on Visualization and Computer Graphics**, v. 14, n. 6, p. 1444–1451, 2008. ISSN 10772626. Citado na página 59.

SHARKO, J.; GRINSTEIN, G. G.; MARX, K. A.; ZHOU, J.; CHENG, C. H.; ODELBURG, S.; SIMON, H. G. Heat map visualizations allow comparison of multiple clustering results and evaluation of dataset quality: Application to microarray data. In: **Information Visualization, 2007. IV '07. 11th International Conference**. [S.l.: s.n.], 2007. p. 521–526. ISSN 1550-6037. Citado na página 57.

SHNEIDERMAN, B. The eyes have it: A task by data type taxonomy for information visualizations. In: **IEEE Symposium on Visual Languages**. Los Alamitos, CA, USA: IEEE Computer Society, 1996. p. 336–343. ISSN 1049-2615. Citado nas páginas 36 e 51.

SILVA, A. T. da; XAVIER, A.; MAGALHÃES, L. P. A new cbir approach based on relevance feedback and optimum-path forest classification. Václav Skala-UNION Agency, 2010. Citado na página 29.

SILVA, V. D.; TENENBAUM, J. B. **Sparse multidimensional scaling using landmark points**. [S.l.], 2004. Citado na página 32.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining, (First Edition)**. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2005. ISBN 0321321367. Citado nas páginas 26, 36, 37, 38, 39, 41, 66 e 76.

TEJADA, E.; MINGHIM, R. Improved visual clustering of large multidimensional datasets. In: **Information Visualization 2005**. [S.l.]: IEEE CS Press, 2005. p. 818–825. Citado na página 31.

- TEJADA, E.; MINGHIM, R.; NONATO, L. G. On improved projection techniques to support visual exploration of multidimensional data sets. **Information Visualization**, Palgrave Macmillan, v. 2, n. 4, p. 218–231, 2003. ISSN 1473-8716. Citado na página 33.
- TELEA, a.; ERTL, T.; JOY, K.; SANTOS, B. Combining Extended Table Lens and Treemap Techniques for Visualizing Tabular Data. p. 51–58, 2006. Disponível em: <<http://www.win.tue.nl/~alex/ALEX/PAPERS/EuroVis06/tablevision.p>>. Citado na página 61.
- TENENBAUM, J. B.; SILVA, V. D.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **science**, American Association for the Advancement of Science, v. 290, n. 5500, p. 2319–2323, 2000. Citado na página 32.
- TENEV, T.; RAO, R. Managing multiple focal levels in Table Lens. **Proceedings of VIZ '97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium**, 1997. Citado na página 34.
- THOMAS, J. J.; COOK, K. A. (Ed.). **Illuminating the path: The Research and Development Agenda for Visual Analytics**. Los Alamitos: IEEE Computer Society Press, 2005. ISBN 0-7695-2323-4. Citado na página 29.
- THOMAS, J. J.; COOK, K. A. **Illuminating the Path: The Research and Development Agenda for Visual Analytics**. National Visualization and Analytics Ctr, 2005. Paperback. ISBN 0769523234. Disponível em: <<http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0769523234>>. Citado na página 43.
- TORGERSON, W. S. Multidimensional scaling of similarity. **Psychometrika**, v. 30, n. 4, p. 379–393, 1965. ISSN 1860-0980. Disponível em: <<http://dx.doi.org/10.1007/BF02289530>>. Citado na página 32.
- TURKAY, C.; PARULEK, J.; REUTER, N.; HAUSER, H. Integrating cluster formation and cluster evaluation in interactive visual analysis. In: **Proceedings of the 27th Spring Conference on Computer Graphics - SCCG '11**. New York, New York, USA: ACM Press, 2013. p. 77. ISBN 9781450319782. Disponível em: <<http://www.iu.uib.no/vis/publications/publication/2011/pdfs/turkay11cluster.pdfhttp://dl.acm.org/citation.cfm?doid=2461217.2461234>>. Citado nas páginas 53 e 54.
- UTTS, J. **Seeing Through Statistics**. Cengage Learning, 2004. (Available Titles CengageNOW Series). ISBN 9780534394028. Disponível em: <https://books.google.com.br/books?id=5_ZKqfMXXuIC>. Citado na página 31.
- VIERA, F. P. **Mapeamento de dados multi-dimensionais - integrando mineração e visualização**. Doutorado, 2008. Citado nas páginas 31, 34 e 45.
- WARD, M. O.; GRINSTEIN, G.; KEIM, D. **Interactive Data Visualization: Foundations, Techniques, and Applications**. [S.l.]: CRC Press, 2015. Citado nas páginas 25, 31, 34, 35, 36 e 38.
- WONG, P. C. Visual data mining. In: **IEEE Computer Graphics and Applications**. [S.l.: s.n.], 1999. Citado na página 25.
- YANG, J.; WARD, M. O.; RUNDENSTEINER, E. A. Interactive hierarchical displays: a general framework for visualization and exploration of large multivariate data sets. **Computers and Graphics**, p. 265–283, 2003. Citado nas páginas 47 e 48.

YERA, A.; ARBELAITZ, O.; JODRA, J. L.; GURRUTXAGA, I.; PÉREZ, J. M.; MUGUERZA, J. Analysis of several decision fusion strategies for clustering validation. strategy definition, experiments and validation. **Pattern Recognition Letters**, v. 85, p. 42 – 48, 2017. ISSN 0167-8655. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0167865516303324>>. Citado na página 41.

ZHOU, J.; KONECNI, S.; GRINSTEIN, G. **Visually comparing multiple partitions of data with applications to clustering**. 2009. 72430J-72430J-12 p. Disponível em: <<http://dx.doi.org/10.1117/12.810093>>. Citado na página 55.

ZHOU, J.; KONECNI, S.; MARX, K.; GRINSTEIN, G. A visual approach to improve clustering based on cluster ensembles. In: PARK, J.; HAO, M. C.; WONG, P. C.; CHEN, C. (Ed.). **Analysis**. [s.n.], 2010. v. 7530, p. 753009. ISBN 9780819479235. ISSN 0277786X. Disponível em: <<http://link.aip.org/link/PSISDG/v7530/i1/p753009/s1{&}Agg=doihttp://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.844>>. Citado nas páginas 55 e 56.