SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito: 06.05.2005

Assinatura:

Ana Paula Jampais Jr/en

Estratégias para apoiar a detecção de
estruturas em visualizações
multidimensionais perceptualmente
sobrecarregadas

Almir Olivette Artero

Orientadora: Profa. Dra. Maria Cristina Ferreira de Oliveira

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

USP – São Carlos
Maio de 2005

Aluna: Almir Olivette Artero

A Comissão Julgadora:

Profa. Dra. Maria Cristina Ferreira de Oliveira

Maria Cristina Ferreira de Oliveira

Profa. Dra. Rosane Minghim

Rosane Minghim

Prof. Dr. Haim Levkowitz

Haim Levkowitz

Profa. Dra. Carla Maria Dal Sasso Freitas

Carla Maria Dal Sasso Freitas

Prof. Dr. João Batista Souza de Oliveira

João Batista Souza de Oliveira

Dedico este trabalho a meus pais
Francisco e Helena e a minha irmã
Angélica, que sempre me encorajavam a
ir, mesmo querendo que eu ficasse.

Agradecimentos

Agradeço a proteção de Deus e minha mãe poderosa, por estarem sempre comigo em todos os momentos desta jornada.

A Professora Cristina, minha orientadora, por me introduzir na área da Visualização e me guiar em todos os passos desse trabalho, minha eterna gratidão.

Aos meus professores nas disciplinas cursadas no ICMC, Luis Gustavo Nonato, Maria Carolina Monard e Maria das Graças Campos Pimentel; e no IFSC Luciano da Fontoura Costa, por contribuírem com a minha formação.

Aos professores Haim Levkowitz e Rosane Minghim, pelas suas importantes contribuições.

Aos professores Agma Juci Machado Traina, Carla Maria Dal Sasso Freitas e Marcelo Knörich Zuffo, pelas sugestões apresentadas no exame de qualificação.

A Neri Alves, por sua determinação, sempre me incentivando a continuar.

A Josmar Mazucheli, Messias Meneguette e Ricardo Barbosa, pelo apoio e amizade constante.

A Silvio Tadao Fujisaki, pelo companheirismo em nossas viagens memoráveis.

A Rogério Eduardo Garcia e Mario Pazzoti pelas nossas várias discussões.

As secretárias do ICMC Beth, Laura e Ana Paula.

Aos amigos e amigas de São Carlos.

Aos funcionários e funcionárias do ICMC.

Sumário

Lista de Caixas..... vi

Lista de Tabelas..... vii

Lista de Figuras..... viii

Resumo..... xv

Abstract..... xvi

Capítulo 1 Introdução..... 1

Capítulo 2 Visualização de Informação..... 6

2.1 Considerações Iniciais..... 6

2.2 Terminologia..... 6

2.3 Tipos de Dados Básicos..... 10

2.4 Conjuntos de Dados..... i

2.4.1 Conjuntos de Dados Reais..... 1

2.4.2 Conjuntos de Dados Sintéticos..... 13

2.4.3 Processo de Geração de Dados..... 15

2.5 Classificação das Técnicas..... 16

2.6 Técnicas de Visualização de Informação..... 18

2.6.1 Técnicas 2D/3D Tradicionais..... 18

2.6.2 Técnicas Orientadas a Pixel..... 19

2.6.3 Técnicas de Projeção Geométrica..... 20

2.6.4 Técnicas Iconográficas..... 26

2.6.5 Técnicas Hierárquicas..... 28

2.7 Técnicas de Interação e Distorção..... 29

2.7.1 Técnicas de Projeção Interativa..... 30

2.7.2 Filtragem..... 31

2.7.3 Zoom..... 33

2.7.4 Distorção..... 33

2.7.5 Seleção e Ligação..... 34

2.8 Considerações Finais..... 38

Capítulo 3 Mineração de Dados..... 40

3.1 Considerações Iniciais..... 40

3.2 Processo de Descoberta do Conhecimento..... 41

3.3 Técnicas de Mineração de Dados..... 46

3.3.1 Regras de Associação..... 46

3.3.2 Sumarização.....	47
3.3.3 Classificação.....	47
3.3.4 Agrupamento.....	48
3.4 Mineração Visual de Dados.....	50
3.4.1 Visualização Direta.....	51
3.4.2 Visualização de Modelos de Mineração.....	52
3.4.3 Visualização Acoplada à Mineração.....	55
3.5 Comentários Finais.....	59
Capítulo 4 Viz3D: Visualização Multidimensional no Espaço 3D.....	61
4.1 Considerações Iniciais.....	61
4.2 Trabalhos Relacionados.....	61
4.2.1 Análise de Componentes Principais.....	62
4.2.2 <i>Projection Pursuit</i>	62
4.2.3 Decomposição em Valores Singulares.....	63
4.2.4 Escalonamento Multidimensional.....	63
4.2.5 Mapeamento não Linear.....	64
4.2.6 <i>FastMap</i>	64
4.2.7 <i>VHDR</i>	65
4.3 <i>Viz3D</i> : Visualização Multidimensional no Espaço 3D.....	65
4.3.1 Projecção no <i>Viz3D</i>	65
4.3.2 Desempenho.....	67
4.4 Resultados.....	68
4.5 Comentários Finais.....	71
Capítulo 5 Visualizações Realçadas com Informação de Frequência e Densidade.....	73
5.1 Considerações Iniciais.....	73
5.2 Definições.....	73
5.3 Trabalhos Relacionados.....	75
5.4 <i>IPC Frequency</i> e <i>Density Plots</i>	80
5.4.1 Desempenho dos <i>IPC Frequency</i> e <i>Density Plots</i>	84
5.4.2 Interação com os <i>IPC Frequency</i> e <i>Density Plots</i>	85
5.5 Frequência e Densidade no <i>Radviz</i> e <i>Viz3D</i>	87
5.6 Comentários Finais.....	90
Capítulo 6 Exploração Visual de Dados usando as Técnicas Propostas.....	92
6.1 Comentários Iniciais.....	92
6.2 Agrupamento Visual no <i>Viz3D</i> e <i>Viz3D Density Plot</i>	93
6.2.1 Agrupamento apoiado pelo <i>Viz3D</i>	93
6.2.2 Identificação de Pontos Incomuns.....	96
6.2.3 Agrupamento com o <i>Viz3D Density Plot</i>	96
6.3 Agrupamento Visual nos <i>IPC Frequency</i> e <i>Density Plots</i>	98
6.4 Análise Visual Integrada.....	106

Sumário	
6.5 Análise dos Agrupamentos.....	110
6.6 Considerações Finais.....	113
Capítulo 7 Arranjo e Redução de Atributos.....	115
7.1 Considerações Iniciais.....	115
7.2 Trabalhos Relacionados.....	116
7.3 Arranjo de Atributos Baseado em Similaridade.....	117
7.4 Redução da Dimensionalidade.....	120
7.5 Desempenho.....	122
7.6 Experimentos.....	123
7.7 Comentários Finais.....	131
Capítulo 8 Conclusões.....	133
Referências Bibliográficas.....	138

Lista de Caixas

5.1 – Algoritmo para a construção do <i>IPC Frequency Plot</i>	82
5.2 – Algoritmo para a construção de gráficos de densidades no <i>Viz3D</i>	88
6.1 – Algoritmo proposto para obter a distância intra-cluster.....	111
6.2 – Algoritmo proposto para obter a distância inter-cluster.....	112
7.1 – Algoritmo <i>SBAA (Similarity Based Attribute Arrangement)</i>	120
7.2 – Algoritmo <i>SBAR (Similarity Based Attribute Reduction)</i>	120
7.3 – Algoritmo <i>SBAR_m (Similarity Based Attribute Reduction-modified)</i>	121

Lista de Tabelas

2.1 – Caracterização do tamanho dos conjuntos de dados, segundo Wegman.....	9
2.2 – Caracterização do tamanho dos conjuntos de dados, segundo Oliveira e Levkowitz.....	9
2.3 – Grupos do conjunto de dados <i>Synth3.data</i>	14
2.4 – Grupos do conjunto de dados <i>Synth4.data</i>	14
5.1 - Valores originais.....	81
5.2 - Valores mapeados no intervalo $[1,L]$	81
7.1 – Distâncias intra-e inter-grupo relativas aos grupos do conjunto de dados <i>Synth3.data</i> obtidas a partir dos 20 atributos na sequência original dada no arquivo.....	126
7.2 – Distâncias intra e inter-grupos relativas aos grupos do conjunto de dados <i>Synth3.data</i> obtidas a partir das coordenadas tridimensionais geradas no <i>Viz3D</i> , considerando a configuração original dos atributos.....	126
7.3 – Distâncias intra- e inter-grupos relativas aos grupos do conjunto de dados <i>Synth3.data</i> obtidas a partir das coordenadas tridimensionais geradas no <i>Viz3D</i> , considerando os atributos arranjados com o <i>SBAA</i> , usando a Equação 7.2.....	126

Lista de Figuras

1.1 – Visualização Científica de dados de climatologia..... 2

2.1 – Modelo de referência para o processo de visualização..... 8

2.2 – Nove espécies de folhas descritas no conjunto de dados *Folhas.data*..... 12

2.3 – Classificação das Técnicas de Visualização..... 16

2.4 – Gráficos bidimensionais e tridimensionais tradicionais..... 18

2.5 – Visualização orientada a pixels de um conjunto de dados aleatório em que cada item de dado tem nove atributos, usando dois arranjos distintos para a disposição dos itens a) espiral e b) por eixos..... 19

2.6 – a) Arranjo adotado na técnica Segmentos Circulares e b) Visualização de um conjunto de dados com 50 atributos..... 20

2.7 – Matriz de gráficos de dispersão do conjunto de dados *Pollen.data* (o gráfico de dispersão dos atributos 2 e 3 é ampliado à direita)..... 21

2.8 – Visualização por *Coordenadas Paralelas* do conjunto de dados *Cars.data*. Cada eixo é rotulado com o nome, o menor e o maior valor de cada atributo..... 21

2.9 – a) Visualização do conjunto de dados *Pollen.data* com apenas 3.848 registros (cinco atributos) gera intensa sobreposição das linhas e; b) Visualização de um conjunto de dados contendo algumas centenas de registros com 40 atributos..... 22

2.10 – a) Obtenção da técnica Gráfico Estrela a partir da técnica Coordenadas Paralelas; e b) Visualização de dois registros de dimensão oito utilizando o Gráfico Estrela..... 23

2.11 – Visualização do conjunto de dados *Cars.data* no *RadViz*..... 24

2.12 – a) Mapeamento de registros *n*-dimensionais no *RadViz*; b) Conjunto de dados contendo quatro registros (r_0, r_1, r_2 e r_3) com quatro atributos (a_0, a_1, a_2 e a_3); c) Visualização dos registros usando a disposição de eixos a_0, a_1, a_2 e a_3 ; d) Visualização usando a disposição a_0, a_2, a_1 e a_3 24

2.13 – a) Visualização do conjunto de dados *Iris.data* no *Star Coordinates*; b) Visualização obtida após interação do usuário com os eixos..... 25

2.14 – a) Disposição dos dados no Tubo de Dados; b) Visualização de alguns registros de um conjunto de dados com seis atributos..... 26

2.15 – a) Faces de Chernoff (*Chernoff Faces*); b) Ícones Estrela (*Star Glyphs*)..... 27

2.16 – a) *Stick Icon*; b) Visualização de dados 5-dimensionais relativos à região dos Grandes Lagos..... 27

2.17 – Dados de dimensionalidade seis mapeados para o espaço tridimensional utilizando a técnica Mundo dentro de Mundos. No caso, os valores dos atributos x_3, x_4 e x_5 são mantidos constantes..... 28

2.18 – Técnica Empilhamento de Dimensões usando duas configurações diferentes. A cor também pode codificar informação..... 29

2.19 – a) Árvore a ser visualizada usando b) a técnica Mapa de Árvore..... 29

2.20 – Visualização dos registros pertencentes aos grupos do conjunto de dados *Pollen.data* a) antes e b) após a aplicação do *Grand Tour*..... 30

2.21 – Exibição de dados do censo inglês de 1991. No caso, a projeção não revela a presença de grupos, sugerindo que esta característica não ocorre ou não é evidenciada em duas dimensões..... 31

2.22 – Utilização de lentes para auxiliar a visualização de um gráfico de dispersão. O filtro em (a) amplia a exibição, revelando quatro elementos (cidades americanas), enquanto que em (b) o filtro exhibe os dados associados aos elementos gráficos selecionados..... 32

2.23 – a) *Slider* comum; b) *Slider* com dois cursores e a indicação gráfica da densidade dos dados; c) Botões de rádio e d) Caixas de verificação..... 32

2.24 – a) Zoom interativo; b) Visualização utilizando *TableLens*, com *Zoom* Interativo aplicado sobre três registros..... 33

2.25 – Distorção obtida por meio de perspectiva..... 34

2.26 – Seleção de marcadores em: um gráfico de dispersão a) individualmente apontando o mouse sobre o marcador e b) definindo uma região; Coordenadas Paralelas c) individualmente apontando o mouse sobre o marcador e d) definindo uma região sobre um dos eixos.....	35
2.27 – a) Conjunto de objetos e b) Seleção especificada no espaço da hierarquia dos dados; c) Representação da árvore em (b), usada na ferramenta <i>XmDyTool</i>	36
2.28 – a) Exibição do conjunto de dados <i>Cars.data</i> ; b) Exibição dos valores médios dos atributos entre os registros de mesma origem.....	36
2.29 – Ligação (<i>linking</i>) entre as visualizações a) Coordenadas Paralelas e b) Matriz de Gráficos de Dispersão. A cor escura indica registros selecionados, no caso, os carros americanos.....	37
3.1 – Etapas do processo <i>KDD</i>	41
3.2 – a) Tabela original com um atributo a_i categórico; b) Tabela expandida.....	43
3.3 – a) Distâncias entre treze cidades da Nova Zelândia; b) Posicionamento obtido usando o <i>MDS</i> ; c) Mapa da Nova Zelândia.....	44
3.4 – a) Registros contendo um atributo categórico; b) determinação de índices para cada valor do atributo a_i ; c) Mapeamento aplicado ao atributo em (a).....	45
3.5 – a) Conjunto de dados e b) seu respectivo dendograma. Os comprimentos dos segmentos verticais do dendograma são proporcionais às distâncias entre os agrupamentos que interligam, conforme indicam d_{ij} e d_{hi}	49
3.6 – Proposta para auxiliar o estudo da correlação entre os atributos.....	51
3.7 – a) Visualização em coordenadas paralelas do conjunto <i>Cars.data</i> ; b) Sumarização (valor médio) das poligonais selecionadas pelo usuário em (a).....	51
3.8 – a) Visualização do conjunto de dados <i>Iris.data</i> por Coordenadas Paralelas; b) Visualização das regras de associação usando o <i>RuleViz</i>	52
3.9 – a) Conjunto de dados (bidimensional); b) Informação de densidade; c) Visualização da densidade de um conjunto de dados com 30.000 registros com 17 atributos usando Segmentos Circulares.....	53
3.10 – Sequência de visualizações de um volume de dados com 230.000 itens de dados, em diferentes níveis de detalhe (acima à esquerda, menos detalhe; abaixo à direita, mais detalhe; na última todos os registros do volume são exibidos).....	54
3.11 – Ferramenta para <i>brushing</i> hierárquico; (a) Representação da árvore de hierarquia; (b) Região selecionada; (c) Raios dos grupos selecionados; (d) Raios dos grupos não selecionados; (e) Contorno da árvore; (f) Delimitadores horizontais da região selecionada; e (g) Escala de cores utilizada no mapeamento.....	54
3.12 – a) Utilização do <i>brushing</i> hierárquico; b) Definindo os raios dos grupos selecionados iguais a 0.0 obriga a sua exibição individual; definindo os raios dos grupos não selecionados iguais a 1.0 anexa todos os demais registros em um único grupo (região transparente).....	55
3.13 – a) Visualização orientada a pixel de um conjunto de dados com nove atributos, utilizando o arranjo em (b); c) Árvore de decisão; d) Visualização da árvore.....	56
3.14 – Identificação dos grupos usando projeções. Em a) o grupo (3) é adequadamente identificado nas projeções paralelas aos eixos, enquanto que os grupos (1) e (2) precisam de uma projeção diagonal. Em b) é exibido um particionamento (<i>Grid</i>) obtido no primeiro passo do <i>OptiGrid</i>	56
3.15 – a) Ícones usados para visualizar informação de densidade; b) Ícone mapeando informação de densidade; c) Histograma da densidade obtida em uma projeção; d) Gráfico unidimensional de densidade; e) Gráfico bidimensional de densidade.....	57
3.16 – a) Principais passos da exploração usando o <i>Eureka!</i> ; b) Visualização do conjunto <i>Image Segmentation.data</i> ; c) Usuário define uma região de interesse.....	58
3.17 – Agrupamento de dados com o <i>HC-Enhanced</i>	59
4.1 – Visualização de um conjunto de dados com dimensão nove utilizando o método proposto por Sammon.....	64
4.2 – Projeção 3D no <i>Viz3D</i> ; b) Mapeamento dos registros r_0 , r_1 , r_2 e r_3 (dimensionalidade quatro, ver Figura 2.16) no <i>Viz3D</i> , adotando a sequência de eixos a_0 , a_1 , a_2 e a_3 ; c) Mapeamento com a sequência de eixos a_0 , a_1 , a_2 e a_3	67

4.3 – a) Tempos de execução (em segundos) para obter projeções de um conjunto de dados de dimensionalidade 11 com <i>PCA</i> , <i>FastMap</i> e <i>Viz3D</i> ; b) respectivo gráfico exibindo o tempo de processamento em função do número de registros.....	68
4.4 – Projeção de um hiper-cubo (4D) - a) no plano, com o <i>RadViz</i> ; no espaço 3D com: b) <i>PCA</i> , c) <i>FastMap</i> e d) <i>Viz3D</i>	68
4.5 – a) Projeção do conjunto de dados <i>Iris.data</i> usando: a) <i>Fastmap</i> , b) <i>PCA</i> , c) <i>NLM</i> , d) <i>RadViz</i> e e) <i>Viz3D</i>	69
4.6 – a) Visualização inicial do conjunto de dados <i>Pollen.data</i> no <i>Viz3D</i> ; b) Visualização após ampliação da área central; c) Identificação visual dos grupos.....	70
4.7 – Visualização <i>Viz3D</i> do conjunto de dados <i>Out5d.okc</i>	70
4.8 – Visualização do conjunto <i>Synth3.data</i> no <i>Viz3D</i>	71
5.1 – a) Conjunto de dados bidimensional; b) Visualização em que os níveis de cinza apresentam a informação de densidade.....	74
5.2 – a) <i>Density graph</i> , e b) <i>Contour graph</i> . Em ambos os casos, a densidade de pontos é indicada pelo matiz do pixel (tons laranja codificam valores maiores e tons púrpura codificam valores menores). No <i>contour plot</i> a cor de uma região também é influenciada pela cor das regiões vizinhas.....	76
5.3 – a) Visualização do conjunto de dados <i>Synth1.data</i> (7.500 registros de cinco atributos).....	76
5.4 – a) Múltiplos registros, idênticos ou não, mapeados no mesmo marcador, causam a sobreposição das poligonais; b) Registros distintos ocasionam cruzamentos das poligonais; c) Registros com alguns atributos consecutivos iguais causam cruzamentos e sobreposição parcial das poligonais.....	77
5.5 – a) Visualização do conjunto de dados <i>Pollen.data</i> em Coordenadas Paralelas convencional; b) Visualização da informação de densidade dos mesmos dados, com a intensidade do pixel proporcional à sobreposição no local.....	78
5.6 – a) Visualização do conjunto de dados <i>Cancer.data</i> em Coordenadas Paralelas; b) <i>Frequency Plot</i> ; c) <i>Frequency Plot</i> dos registros da Classe 1; d) <i>Frequency Plot</i> dos registros da Classe 2.....	79
5.7 – <i>Frequency Plot</i> do conjunto de dados <i>Pollen.data</i>	79
5.8 – Visualizações do conjunto de dados <i>Synth1.data</i> a) <i>Frequency Plot</i> e b) <i>Relevance Plot</i>	80
5.9 – a) Gráfico de dispersão dos dados exibidos na Tabela 5.2; b) Visualização por Coordenadas Paralelas dos mesmos dados; c) Matriz de frequências <i>F</i> deste par de atributos; d) Visualização por Coordenadas Paralelas em que a intensidade das poligonais é proporcional ao valores de frequência em <i>F</i>	82
5.10 – a) Filtro de suavização onda quadrada; b) matriz de densidades obtida aplicando o filtro sobre a matriz de frequência <i>F</i> (Figura 5.9 (c)); c) <i>IPC Density Plot</i> resultante.....	84
5.11 – a) Tempos de execução (em segundos) do algoritmo proposto, aplicado em conjuntos de dados com diferentes valores de <i>m</i> (número de registros) e <i>n</i> (número de atributos); e b) correspondente gráfico de tempo, escala logarítmica.....	85
5.12 – a) Visualização do conjunto de dados <i>Synth1.data</i> com <i>s</i> = 1.0; b) com diferentes valores de <i>s</i>	85
5.13 – Visualização do conjunto de dados <i>Pollen data</i> usando a) <i>IPC Frequency Plot</i> ; b) <i>IPC Density Plot</i>	86
5.14 – a) <i>IPC Frequency Plot</i> do conjunto de dados <i>Synth1.data</i> (7.500 registros com cinco atributos); b) <i>IPC Density Plot</i> ; c) visualização dos 2.920 registros efetivamente pertencentes aos cinco grupos do conjunto de dados.....	86
5.15 – Visualização do conjunto de dados <i>Synth4.data</i> com o <i>RadViz</i> a) marcadores traçados com intensidade constante, e b) marcadores traçados com intensidade proporcional à frequência.....	88
5.16 – Núcleo onda quadrada tridimensional ($K_w = 5$).....	89
5.17 – <i>Viz3D Density Plot</i> do conjunto de dados <i>Synth3.data</i>	90
5.18 – Visualização do conjunto de dados <i>Out5d.okc</i> a) no <i>Viz3D</i> e b) no <i>Viz3D Density Plot</i> ..	90
6.1 – Processo de agrupamento visual usando o <i>Viz3D</i>	94

6.2 – a) Registros identificados no processo de agrupamento visual no <i>Viz3D</i> ; b) Visualização dos mesmos registros por Coordenadas Paralelas indica a presença de registros que não devem ser mantidos no grupo.....	95
6.3 – a) Visualização <i>Viz3D</i> do conjunto de dados <i>Folhas.data</i> , com os marcadores coloridos pelo atributo classe; b) Exibição parcial dos atributos dos elementos da Classe 8, em particular do registro 998; c) Comprovação do ponto incomum na visualização por Coordenadas Paralelas.....	96
6.4 – a) Visualização <i>Viz3D</i> do conjunto de dados <i>Out5d.okc</i> ; b) Visualização dos registros que possuem densidade maior que um limiar definido pelo usuário.....	97
6.5 – Seleção dos registros obtida com a operação de Limiarização a partir da informação de densidade no <i>Viz3D</i>	98
6.6 – Agrupamento visual interativo com o <i>IPC Density Plot</i>	99
6.7 – a) Visualização do conjunto de dados <i>Pollen.data</i> usando o <i>IPC Density Plot</i> com as seleções do usuário sobre os quatro primeiros eixos; b) Exibição dos registros selecionados em (a)	100
6.8 – Identificação de um grupo no conjunto de dados <i>Synth4.data</i>	102
6.9 – a) Visualização do conjunto de dados <i>Out5d.okc</i> por Coordenadas Paralelas. b) Visualização no <i>XmdvTool</i> usando a técnica Coordenadas Paralelas Hierárquicas (raios selecionados = 0.58 e <i>showband</i> = true).....	103
6.10 – Identificação de um grupo no conjunto de dados <i>Out5d.okc</i> usando o <i>IPC Density Plot</i>	104
6.11 – a) 2.773 registros obtidos usando uma limiarização $T(AND) = 2$; b) 1.802 registros obtidos usando o limiar $T(AND) = 3$	105
6.12 – a) 117 registros obtidos com limiar $T(OR) = 3$; b) respectiva projeção usando os três primeiros atributos dos registros; c) 87 registros obtidos com $T(OR) = 4$; d) respectiva projeção dos registros.....	105
6.13 – a) Visualização por Coordenadas Paralelas dos registros do conjunto de dados <i>Pollen.data</i> , pertencentes ao grupo identificado na Figura 6.7; b) exibição dos mesmos registros no <i>Viz3D</i> ; c) Visualização do conjunto de dados <i>Hypercube.data</i> por Coordenadas Paralelas; d) e no <i>Viz3D</i> ; e) Visualização de 10 registros pertencentes a uma linha reta com dimensão 9 em Coordenadas Paralelas; f) Visualização dos mesmos registros no <i>Viz3D</i>	107
6.14 – Visualizações do conjunto de dados <i>Image Segmentation.data</i> a) <i>Viz3D</i> , b) Coordenadas Paralelas.....	108
6.15 – Exploração do conjunto de dados <i>Houses.data</i> : a) visualização <i>Viz3D</i> usando apenas dos atributos Latitude e Longitude, marcadores coloridos pelo atributo classe; b) visualização por Coordenadas Paralelas; c) <i>IPC Density Plot</i> indica áreas de maior concentração de residências; d) visualização <i>Viz3D</i> de alguns grupos identificados pelo procedimento de agrupamento visual apoiado no <i>IPC Density Plot</i> exibido em (c).....	109
6.16 – Visualização do conjunto de dados <i>Cancer.data</i> a) em Coordenadas Paralelas; b) no <i>Viz3D</i>	110
6.17 – Distâncias usadas no cálculo da distância intra-grupo para a) um agrupamento circular e b) um agrupamento alongado.....	111
6.18 – Distâncias usadas na métrica proposta para o cálculo da distância intra-grupo a) agrupamento circular; e b) agrupamento alongado.....	111
6.19 – Análise dos grupos obtidos na com o processo de agrupamento visual ilustrado na Figura 6.6; a) Matriz de Confusão; b) Descritivo dos grupos (classes indicam a formação real e obtidos).....	112
6.20 – a) Visualização do conjunto de dados <i>Synth1.data</i> no <i>Viz3D</i> ; b) Valores de distância intra-grupo e inter-grupo, usando a métrica Euclidiana.....	113
7.1 – Configurações dos atributos: a) Linear 1D, b) Circular 1D e c) 2D.....	121
7.2 – a) Conjunto de dados contendo dois registros de seis atributos; b) Visualização <i>RadViz</i> com o arranjo dos atributos segundo a sequência dada nos registros; c) Visualização por Coordenadas Paralelas, com o mesmo arranjo; d) Visualização <i>RadViz</i> com o arranjo baseado em similaridade; e) Visualização por Coordenadas Paralelas com o mesmo arranjo.....	122

7.3 – a) Tempos de execução, em segundos, do algoritmo <i>SBAA</i> ; b) respectivo gráfico em escala logarítmica.....	123
7.4 – a) Visualização do conjunto de dados <i>Mammals2.data</i> no <i>RadViz</i> , adotando o arranjo original dos atributos; b) Visualização que adota o arranjo baseado em similaridade.....	124
7.5 – Visualizações <i>Viz3D</i> do conjunto de dados <i>Mammals2.dat</i> : a) com um arranjo arbitrário dos atributos; b) com o arranjo baseado na similaridade dos atributos.....	125
7.6 – Visualizações <i>Viz3D</i> do conjunto de dados <i>Synth3.data</i> : a) arranjo original; b) arranjo obtido com o <i>SBAA</i>	125
7.7 – a) Visualização <i>Viz3D</i> do conjunto de dados <i>Synthetic Data Set 2</i> , adotando a sequência original dos atributos; b) Visualização <i>Viz3D</i> adotando a disposição dos eixos obtida com o <i>SBAA</i>	127
7.8 – a) Visualizações por Coordenadas Paralelas do conjunto de dados <i>Mammals2.data</i> (72 atributos): a) adotando o mesmo arranjo dos atributos fornecido nos dados; b) adotando o arranjo por similaridade dos 72 atributos; c) com apenas 40 atributos obtidos com o <i>SBAR</i> (usando a Equação 7.2); d) idem, para 30 atributos.....	128
7.9 – Visualizações do conjunto de dados <i>Synth_Sierp.data</i> a) Coordenadas Paralelas; b) <i>IPC Density Plot</i> ; c) <i>IPC Density Plot</i> após arranjo dos atributos usando o <i>SBAA</i> ; d) <i>IPC Density Plot</i> após redução de cinco para três atributos.....	129
7.10 – a) Visualização do conjunto de dados <i>Synth_Sierp.data</i> com o <i>IPC Density Plot</i> , adotando a sequência dos atributos obtida com o <i>SBAA</i> usando o coeficiente de correlação de Pearson como medida de similaridade (Equação 7.4; b) Visualização considerando apenas os três atributos obtidos com o <i>SBAR</i>	130
7.11 – Visualizações <i>Viz3D</i> do conjunto de dados <i>Synth_Sierp.data</i> : a) posicionamento inicial considerando todos os atributos; b) após operação de rotação; c) usando a sequência de atributos <i>Corr_Rand</i> , <i>Y_Sierp</i> e <i>Corr_XY</i> ; d) usando a sequência de atributos <i>Rand</i> , <i>X_Sierp</i> e <i>Y_Sierp</i>	131

Resumo

A disponibilização de técnicas efetivas para a Visualização Exploratória e Mineração Visual de conjuntos de dados volumosos e de alta dimensionalidade ainda representa um grande desafio para os pesquisadores. Apesar das muitas contribuições e dos avanços recentes, técnicas convencionais de visualização direta – em que cada elemento do conjunto de dados é mapeado em um elemento visual na tela – ainda são bastante limitadas nesse contexto. Neste trabalho técnicas de visualização exploratória foram revisitadas, e são enriquecidas com algumas estratégias que permitem gerar representações visuais mais efetivas para apoiar processos interativos de exploração de grandes conjuntos de dados de alta dimensionalidade. As técnicas introduzidas apresentam baixa complexidade computacional, de modo a não prejudicar ou inviabilizar a interação do usuário com as visualizações, na busca de padrões como agrupamentos e regiões de alta densidade nos dados, e podem ser caracterizadas como quatro contribuições distintas: (i) primeiramente, é apresentada uma nova técnica de visualização multi-dimensional, denominada *Viz3D*, que projeta os dados em três dimensões espaciais e assim ameniza o efeito indesejado da excessiva sobreposição dos marcadores, típica de estratégias bidimensionais, como o *RadViz* e *Star Coordinates*, oferecendo uma alternativa de baixo custo para a projeção de dados multidimensionais no espaço 3D; (ii) técnicas de visualização direta, como *Coordenadas Paralelas*, *RadViz* e o próprio *Viz3D*, são enriquecidas com informações de frequência e densidade computadas a partir dos dados, as quais são mapeadas visualmente buscando reduzir o efeito prejudicial de uma excessiva sobreposição dos marcadores, atenuar os efeitos indesejados da presença de ruído nos dados e ainda realçar a presença de padrões nos dados; (iii) é introduzido um procedimento interativo para apoiar a identificação de agrupamentos nos dados, o qual é apoiado nas visualizações enriquecidas com informação de frequência e densidade introduzidas anteriormente; (iv) finalmente, é apresentada uma estratégia simples para configurar a disposição dos atributos de dados mapeados na visualização, a qual resulta em visualizações de melhor qualidade e mais informativas no caso de dados de alta dimensionalidade. As propostas apresentadas são comparadas a trabalhos relacionados existentes na literatura, e avaliadas por meio de sua aplicação a vários conjuntos de dados reais e sintéticos, de vários tamanhos e dimensionalidades. Os resultados indicam que, apesar das estratégias propostas não eliminarem os problemas associados à oclusão e congestionamento visual na visualização de grandes conjuntos de dados, elas oferecem soluções efetivas e de baixo custo que ampliam a escalabilidade visual de técnicas tradicionais, viabilizando tarefas de exploração e mineração que seriam impossíveis, ou muito dificultadas, sem esses recursos.

Abstract

The lack of effective techniques for Exploratory Visualization and Visual Mining of very large high-dimensional data sets still represents a major challenge to be addressed by researchers in the field. Despite the many contributions and recent advances, conventional direct visualization techniques – in which each item in the data set is individually mapped into a visual element in the screen – are still highly limited in this context. In this work we revisit direct exploratory visualization techniques, and propose several enhancements to support the creation of more effective visual representations to assist interactive exploration of large high-dimensional data sets. Four contributions are introduced: (i) first, a new multi-dimensional visualization technique that projects data in three spatial dimensions, called *Viz3D*, is presented; this technique reduces the excessive marker overlapping and concentration that is typical of two-dimensional visualizations such as *RadViz* and *Star Coordinates* and offers a low cost 3D projection technique; (ii) direct visualization techniques, such as *Parallel Coordinates*, *RadViz* and *Viz3D* itself, are enhanced with frequency and density information, computed from the data set and mapped into the data visual representations; the goals of this mapping are to reduce the harmful effects of excessive marker overlapping, to reduce the artifacts due to the presence of noise in the data, and to highlight the presence of patterns in data; (iii) the visualizations enhanced with frequency and density information provide the basis for a new interactive approach to assist users in identifying and extracting clusters in data sets; (iv) finally, a new and simple strategy is introduced to configure the layout of multiple data attributes mapped into a visualization that results in better and more informative visualizations of high dimensional data. The proposed techniques are compared with related work from the literature, and illustrated with their application to several real and synthetic data sets of varying size and dimensionality. Results indicate that, although the new techniques do not completely eliminate the problems associated with occlusion and overlapping, they offer effective low cost solutions that increase the visual scalability of the techniques, enabling users to conduct visual exploration tasks that would not be possible otherwise.

Capítulo 1

Introdução

A capacidade de acumular grandes quantidades de dados de alta dimensionalidade hoje excede em muito a capacidade humana de processá-los adequadamente. Os avanços na área da informática, que por um lado contribuíram para este acúmulo de informações, por outro lado também são extremamente necessários para viabilizar a análise desses dados. Keim [2001] menciona uma pesquisa, realizada na Universidade da Califórnia, que estimou a quantidade de dados gerada anualmente no planeta em torno de um hexabyte (10^{18} bytes), dos quais 99,97% na forma digital. Este enorme volume de informações é consequência do grande número de transações registradas diariamente, e também da grande quantidade de atributos adotados para descrever detalhadamente cada operação. O interesse pelo conhecimento embutido nos dados tem incrementado a busca por ferramentas mais eficientes de extração automática de conhecimento. Porém, as técnicas de mineração disponíveis não conseguem apresentar resultados satisfatórios, de forma totalmente automática, em todas as situações, e a presença de um analista humano continua indispensável. De fato, a própria escolha da técnica de mineração mais adequada constitui uma importante atribuição do analista.

A visualização como alternativa para a exploração dos dados tem merecido bastante atenção, apoiando de várias formas processos de análise e descoberta de conhecimento. Uma das vantagens mais interessantes da utilização de técnicas de visualização de informação na exploração de dados é que elas conseguem inserir o potencial de análise humano no processo de descoberta do conhecimento.

No caso da Visualização Científica, que trabalha com dados originários de algum tipo de fenômeno ou modelo do mundo físico, geralmente existe uma representação espacial natural dos dados. A Figura 1.1 apresenta um exemplo de visualização de dados de

Climatologia [Rhyne e Treinish, 2000], sendo que o modelo visual é construído a partir do conhecimento sobre a organização espacial das entidades envolvidas, no caso, obedecendo ao posicionamento geográfico do terreno. O terreno é representado utilizando as coordenadas espaciais do local, enquanto a cor codifica a elevação do terreno e as partículas (os glifos) representam o campo da velocidade do vento (intensidade e direção). Essa visualização poderia ser animada para mostrar a variação nas direções do vento ao longo do tempo. Em outras palavras, atributos de natureza espacial e temporal têm papel dominante na representação visual gerada, o que é característico em aplicações de Visualização Científica.

Figura 1.1 – Visualização Científica de dados de climatologia [Rhyne e Treinish, 2000].

Por outro lado, a Visualização de Informação manipula dados abstratos, para os quais, geralmente, não existe uma representação visual óbvia, ou um mapeamento imediato em uma entidade geométrica. Isso caracteriza o maior desafio na área, que é criar novas metáforas visuais para apresentar as informações e desenvolver maneiras de manipular estas metáforas de forma a transmitir o sentido da informação [Eick, 1994]. Uma consequência imediata é que representações visuais geradas por técnicas de visualização de informação tendem a ser menos intuitivas, demandando treino do usuário para que sejam interpretáveis. Além disso, o grande volume e a alta dimensionalidade dos conjuntos de dados tornam impraticáveis muitas abordagens clássicas de visualização de informação. Com o objetivo de resolver este problema, alguns autores sugerem a utilização integrada de técnicas de mineração e de visualização, de forma cooperativa, em uma abordagem referenciada como Mineração Visual de Dados (*Visual Data Mining – VDM*).

Ankerst [2001] identifica três abordagens de mineração visual de dados. Na primeira, técnicas de visualização direta, o mapeamento visual é aplicado diretamente sobre cada item de dado, o que caracteriza o contexto convencional de visualização de informação. Nesse caso, a tarefa de identificar padrões e estruturas de interesse fica totalmente por conta do usuário. O maior obstáculo nesta abordagem é que a exibição de todos os itens pode comprometer a visualização, devido à sobrecarga visual provocada pela excessiva sobreposição de marcadores gráficos no caso de conjuntos de dados volumosos.

Na segunda abordagem, a visualização sucede a aplicação de um algoritmo de mineração e, tipicamente, exibe os resultados desse algoritmo, i.e., o modelo extraído. Esta estratégia é interessante porque a mineração normalmente oferece uma descrição mais condensada do conjunto de dados, que pode ser adequadamente visualizada. Entretanto, é preciso considerar que determinar a técnica de mineração mais adequada, sem um conhecimento prévio a respeito dos dados, não é uma tarefa trivial. Assim, a própria escolha do processo de mineração pelo analista pode se mostrar uma tarefa difícil. Mesmo que esta tarefa seja realizada satisfatoriamente, os modelos obtidos com a mineração podem ser de difícil interpretação ou pouco compreensíveis para o usuário.

Uma terceira abordagem consiste em integrar técnicas de mineração e de visualização, de modo a viabilizar um processo de mineração guiado pelo usuário. Com isso, espera-se que o analista ganhe uma melhor compreensão e maior controle do processo de mineração, o que inclui o espaço de busca e as características dos dados manipulados em cada etapa, o que deverá proporcionar um maior direcionamento do processo como um todo, e melhores resultados no final.

Neste trabalho são introduzidas diversas estratégias para apoiar processos mais efetivos de exploração visual de grandes conjuntos de dados de alta dimensionalidade, as quais podem ser agrupadas em quatro categorias: (i) é apresentada uma nova técnica de visualização tridimensional que, em comparação a estratégias bidimensionais similares, ameniza o efeito indesejado da sobreposição dos marcadores na visualização interativa de grandes conjuntos de dados; (ii) informações de frequência e densidade dos dados são introduzidas em técnicas de visualização direta, também com o objetivo de reduzir o efeito prejudicial da sobreposição e de modo a realçar a presença de padrões, facilitando a tarefa de interpretação pelo usuário; (iii) é apresentado um procedimento interativo para apoiar a identificação de agrupamentos em

dados, o qual é apoiado nas visualizações introduzidas anteriormente; (iv) finalmente, é introduzida uma estratégia para configurar eixos associados aos atributos dos dados que resulta em visualizações de melhor qualidade e mais informativas no caso de dados de alta dimensionalidade;

A primeira contribuição consiste em uma nova técnica de visualização direta, inspirada na técnica *RadViz*, que mapeia dados definidos em espaços n -dimensionais projetando-os no espaço tridimensional. A nova técnica, denominada *Viz3D*, procura reduzir o problema da sobreposição de marcadores, crítico em visualizações bidimensionais, preservando a estrutura original dos dados ao projetá-los no espaço tridimensional de saída. Como tal, o *Viz3D* oferece uma alternativa a técnicas de redução da dimensionalidade, ainda que limitada especificamente à redução para o espaço 3D. Esta técnica é apresentada no Capítulo 4, e ilustrada por meio de sua aplicação na visualização de diferentes conjuntos de dados.

A segunda contribuição, introduzida no Capítulo 5, busca melhorar a qualidade de visualizações que projetam dados multidimensionais no espaço bidimensional, como Coordenadas Paralelas e *RadViz*, ou mesmo no espaço tridimensional, como o *Viz3D*, tratando os efeitos da excessiva sobreposição dos marcadores decorrente do grande volume de dados. Para isso, um dos atributos visuais dos marcadores gráficos – no caso, o brilho, ou intensidade – mapeia informação de frequência ou densidade de modo a realçar a concentração dos dados nas diferentes regiões do espaço de dados n -dimensional. Com isso, é possível eliminar da visualização o ruído presente nos dados, pois os marcadores associados a registros de baixa frequência/densidade são exibidos com baixos valores de brilho, e são sobrepostos pelos marcadores de maior brilho. Da mesma forma, regiões de alta intensidade são realçadas. A informação de densidade é computada no espaço discreto da visualização, o que reduz a complexidade do problema. A informação resultante, apesar de aproximada, mostra-se bastante eficaz no contexto da visualização exploratória, como discutido no Capítulo 5. As visualizações enriquecidas com informação de frequência ou densidade estabelecem a base para a terceira contribuição deste trabalho. São definidos procedimentos visuais interativos para a identificação de agrupamentos, descritos e ilustrados no Capítulo 6, em que o usuário interage diretamente com visualizações por Coordenadas Paralelas, *RadViz* ou *Viz3D* para identificar e extrair agrupamentos em conjuntos de dados volumosos. Esta solução pode ser

caracterizada como um exemplo da terceira abordagem identificada por Ankerst et al. [2001], pois dá ao analista a possibilidade de direcionar e acompanhar o processo de agrupamento.

A quarta contribuição deste trabalho aborda outro problema que afeta a qualidade das visualizações, que é a definição de um *layout* adequado para a exibição dos eixos em técnicas de visualização de informação. Em várias dessas técnicas, a sequência em que os atributos são processados e exibidos altera fortemente a visualização resultante, sendo que certas configurações são mais efetivas do que outras para revelar a presença de estruturas nos dados. Este problema é abordado no Capítulo 7, em que é introduzida uma estratégia de otimização para técnicas de visualização baseadas em projeções, como Coordenadas Paralelas, *RadViz* e *Viz3D*. A estratégia proposta é baseada na aplicação de uma heurística simples, baseada em medidas de similaridade de atributos, para organizar e/ou reduzir os atributos a serem exibidos, tornando as visualizações mais eficazes para apoiar tarefas de exploração visual. A estratégia foi desenvolvida com o propósito de tornar viável a sua aplicação a grandes volumes de dados de alta dimensionalidade.

Os demais capítulos deste trabalho estão organizados da seguinte forma: O Capítulo 2 apresenta a terminologia utilizada na área de Visualização de Informação e uma breve descrição dos tipos de dados que ocorrem com frequência nas bases atuais. Os conjuntos de dados usados nos demais capítulos para exemplificar técnicas e soluções também são apresentados. Em seguida, são apresentadas e discutidas técnicas de visualização de dados multidimensionais e algumas técnicas de interação e distorção. No Capítulo 3 são apresentados conceitos da área de Descoberta de Conhecimento em Bases de Dados (*KDD – Knowledge Discovery in DataBases*), sendo descritos os passos do processo de descoberta, que inclui a etapa de Mineração de Dados. Algumas das técnicas de mineração mais utilizadas são apresentadas sucintamente, bem como alguns exemplos de abordagens de Mineração Visual de Dados. Os Capítulos 4 a 7 apresentam as contribuições deste trabalho, como mencionado anteriormente. Por fim, no Capítulo 8 são apresentadas as conclusões e sugestões para trabalhos futuros.

Capítulo 2

Visualização de Informação

2.1 Considerações Iniciais

O objetivo deste capítulo é apresentar um modelo do processo de Visualização de Informação, definir a terminologia básica adotada na área, e introduzir técnicas visuais clássicas, capazes de apoiar processos de análise de dados. Também são introduzidas algumas técnicas de interação, indispensáveis em processos de visualização exploratória. O capítulo está organizado da seguinte maneira: Na Seção 2.2 é apresentada uma caracterização do processo de visualização, bem como a terminologia associada. A Seção 2.3 relaciona os tipos de dados que tipicamente precisam ser manipulados por técnicas de visualização. A Seção 2.4 introduz os conjuntos de dados utilizados com as técnicas visuais abordadas neste trabalho, além da geração de conjuntos de dados sintéticos. A seguir, é apresentada na Seção 2.5, a classificação proposta por Keim [2002] para as técnicas de visualização. Essa classificação considera a natureza do dado a ser visualizado, a abordagem adotada pela técnica em si, e os mecanismos de interação e distorção que podem ser usados para manipular a representação visual criada. Algumas técnicas de visualização relevantes, segundo essa categorização são introduzidas. Na Seção 2.6 são enfocadas algumas técnicas de interação e distorção. Por fim, na Seção 2.7, são discutidas as principais limitações das técnicas de visualização e apresentadas as considerações finais do capítulo.

2.2 Terminologia

Os termos Visualização Exploratória, Visualização de Informação e Visualização Multidimensional denotam o uso de metáforas visuais para representar dados abstratos e apoiar um processo dinâmico e interativo de exploração dos mesmos [Wong e Bergeron, 1997; Card et al., 1999]. Entretanto, técnicas de visualização permitem não apenas criar

representações visuais diretamente dos dados, mas também representar visualmente os resultados de algoritmos de mineração, ou de outras etapas do processo de descoberta de conhecimento, tornando-os mais acessíveis, compreensíveis e eficazes. Alguns autores denominam esta última aplicação de Mineração Visual de Dados, enquanto outros usam o termo de forma mais ampla para denotar quaisquer processos de exploração e análise de dados baseados em metáforas visuais [Hoffman, 1999].

Em Visualização de Informação os modelos gráficos representam conceitos abstratos que não têm, necessariamente, atributos espaciais ou temporais, tipicamente existentes no mundo físico. Por exemplo, registros descrevendo dados demográficos, itens comprados pelos clientes de um supermercado, registros descrevendo acessos de usuários às páginas de um portal na *Internet*, ou ainda registros descrevendo a evolução dos índices de Bolsas de Valores ao longo de um certo período. O número de atributos de cada registro define a dimensionalidade dos dados, assim, dados descritos por registros com um, dois e três atributos são ditos unidimensionais, bidimensionais e tridimensionais, respectivamente, enquanto que dados com mais de três atributos são ditos multidimensionais ou n -dimensionais (em que n é o número de atributos). Entretanto, essa terminologia não é uniforme nas diversas disciplinas que manipulam dados, como a Matemática, Engenharia, Estatística, Mineração e Visualização. Wong e Bergeron [1997] usam os termos ‘multidimensional’ e ‘multivariável’ para referenciar a dimensionalidade das variáveis independentes e dependentes, respectivamente. Entretanto, em tarefas típicas de visualização, nem sempre é possível, ou necessário, determinar quais são as variáveis dependentes e independentes. Assim, neste trabalho, os termos multidimensional e n -dimensional são utilizados, indistintamente, para referenciar a dimensionalidade dos conjuntos de dados com mais de três atributos.

Card et al. [1999] definem Visualização de Informação como a área que estuda o uso interativo, apoiado por computador, de representações visuais de dados abstratos, com o objetivo de ampliar a cognição humana. A meta é oferecer recursos de exploração (visual) dos dados, capazes de aliar a flexibilidade, criatividade e o conhecimento humano à enorme capacidade de armazenamento e poder de processamento dos computadores [Keim, 2001]. A Figura 2.1 mostra o modelo de referência para o processo de visualização proposto por esses autores. O processo é descrito como um mapeamento dos dados brutos para uma representação visual, controlado por meio de interações do usuário nas suas diversas etapas.

Figura 2.1 – Modelo de referência para o processo de visualização [Card et al., 1999].

Em sua forma mais simples, os dados podem ser organizados em tabelas (*data-table* ou *flat-file*), em que cada linha descreve um registro e cada coluna descreve um atributo. Essa organização é adequada para descrever instâncias não relacionadas, em que todos os registros possuem o mesmo conjunto de atributos, representando diferentes instâncias de observação de um mesmo fenômeno. Por exemplo, os registros podem descrever sessões de acesso de diferentes usuários a uma página da *Web*, ou transações de vendas de um supermercado. Representações mais complexas podem ser necessárias quando os dados descrevem relações, tipicamente representadas como Árvores, Redes ou Grafos. Na etapa de transformação devem ser tratados os valores incorretos e não informados (*missing values*), aspectos de normalização, limpeza, etc.

A partir dos dados organizados, o mapeamento visual cria modelos que combinam informações espaciais e propriedades gráficas. Nesta etapa são determinados os atributos a serem codificados em posição, cor, transparência, tipo de marcador (ponto, linha, área, volume), etc. Modelos visuais também podem codificar temporalidade. No caso de dados que descrevem relações, essas são essenciais para a sua compreensão, o que requer técnicas adequadas, como as diversas estratégias específicas para a visualização de dados hierárquicos e de grafos. As técnicas de visualização focadas neste trabalho, entretanto, são orientadas primariamente a dados organizados em tabelas.

Finalmente, as transformações visuais estabelecem parâmetros gráficos para criar as visualizações a partir dos modelos. As transformações mais comuns são sondagem local (*location probe*), controles do ponto de observação (*viewpoint controls*) e distorções (*distortions*) [Card et al., 1999]. A Sondagem Local utiliza a localização para revelar informações adicionais sobre um elemento ou conjunto de elementos. Os Controles Visuais atuam, principalmente, na definição do ponto de observação adotado para visualizar o modelo, incluindo operações como ampliação, translação e recorte. A Distorção permite

destacar áreas de interesse na visualização. Enfim, as transformações visuais consistem em interações que permitem ao usuário observar de diferentes maneiras um mesmo modelo, mudando o foco de interesse e, com isso, ampliar o potencial de uma técnica em revelar mais informação.

Quanto ao tamanho dos conjuntos de dados, a terminologia passa por constantes modificações, em vista do rápido avanço tecnológico que disponibiliza novos dispositivos de armazenamento e, conseqüentemente, permite a criação de volumes de dados cada vez maiores. Uma caracterização dos conjuntos de dados segundo o tamanho é apresentada na Tabela 2.1 [Wegman, 1995].

Tabela 2.1 – Caracterização do tamanho dos conjuntos de dados, segundo Wegman [1995].

Descrição	Tamanho do conjunto de dados (em bytes)	Tipo de armazenamento
Muito Pequeno	10^2	Folha de papel
Pequeno	10^4	Algumas folhas de papel
Médio	10^6	Disquete
Grande	10^8	Disco rígido
Muito Grande	10^{10}	Vários discos rígidos
Massivo	10^{12}	Sistemas robóticos de fitas magnéticas
Super Massivo	10^{15}	Arquivos distribuídos

Considerando o avanço tecnológico desde então, observa-se, por exemplo, que um conjunto de dados “Muito Grande” (com 10^{10} bytes ou 10 *Gigabytes*) hoje pode ser facilmente armazenado em um único disco rígido. A própria memória RAM utilizada atualmente (em torno de 10^9 bytes – 1 Giga byte) comporta, simultaneamente, até dez conjuntos de dados classificados como “Grandes”. Outra tentativa de caracterizar os conjuntos de dados segundo o tamanho é apresentada na Tabela 2.2 [Oliveira e Levkowitz, 2003].

Tabela 2.2 – Caracterização do tamanho dos conjuntos de dados, segundo Oliveira e Levkowitz [2003].

Descrição	Tamanho do conjunto de dados (em registros)
Grande	> 100.000
Muito Grande	> 1.000.000
Massivo	> 100.000.000

Neste trabalho são propostas técnicas capazes de tratar conjuntos de dados grandes e muito grandes, segundo a classificação apresentada na Tabela 2.2.

2.3 Tipos de Dados Básicos

As informações em bancos de dados tipicamente possuem atributos de diferentes naturezas, sendo representados por diferentes tipos de dados, alguns destes descritos em seguida.

Nominal - Também chamado categórico, discreto ou enumerado, é utilizado para representar objetos definidos em um espaço discreto, pertencentes a um grupo restrito de elementos. Dados dessa natureza podem ou não apresentar uma ordem natural entre seus elementos. Alguns autores adotam o termo *enumerado* ou *ordinal* para os dados nominais que apresentam uma ordem, como os dias da semana (segunda, terça, etc.). Em sistemas de visualização, dados do tipo cadeia de caracteres são, geralmente, tratados como nominais;

Numérico - Pode ser definido em um espaço discreto ou contínuo (sem necessariamente a implicação de continuidade matemática [Witten e Frank, 2000]). Valores numéricos podem ser representados como inteiros ou reais (precisão simples ou dupla), sendo que a utilização de um ou outro pode causar sensível diferença no espaço de armazenamento e tempo de processamento;

Data/Horário - Um tipo particular do dado numérico que registra informação temporal. O formato “*aaaa/mm/dd*” (ano, mês e dia) é tipicamente utilizado em bancos de dados relacionais para registrar datas. Entretanto, devido à existência de muitas variações na representação, a possibilidade de manipular formatos diferentes é muito desejável. Para descrever horário o formato “*hh:mm:ss*” é o mais usual, podendo ser necessário eventualmente acomodar submúltiplos do segundo (décimos, centésimos e milésimos).

Além desses tipos de dados básicos, muitos outros são comuns, como imagens, documentos, filmes e outros. Entretanto, é preciso observar que o sistema de referência para a exibição de uma visualização é o sistema de coordenadas da tela, que adota valores inteiros e positivos. Assim, o processo de mapeamento visual demanda, obrigatoriamente, antes do estágio de exibição, o mapeamento do dado em um valor numérico inteiro. Estratégias para mapear valores de tipos diversos em valores numéricos inteiros são abordadas no Capítulo 3, que apresenta as etapas do processo de mineração de dados, incluindo a etapa de transformação, que também realiza este mapeamento.

2.4 Conjuntos de Dados

Dois repositórios de dados muito conhecidos, que incluem dados reais e sintéticos, são o *UCI Machine Learning* (<http://www.ics.uci.edu/~mllearn/MLSummary.html>) e o *StatLib* (<http://lib.stat.cmu.edu/datasets>). Ambos contêm dados provenientes de diferentes áreas e mantêm, para cada conjunto de dados, uma breve descrição que informa a origem dos dados, o número de registros, o número de atributos e informação sobre as classes, caso existam. Alguns dos conjuntos de dados reais utilizados para ilustrar as técnicas propostas neste trabalho vieram desses repositórios, e são introduzidos em seguida. Também foram usados dados sintéticos, cujo processo de geração também é descrito nesta Seção.

2.4.1 Conjuntos de Dados Reais

Conjuntos de dados reais permitem avaliar o comportamento dos algoritmos em situações reais, para as quais eles foram de fato idealizados. Entretanto, dados reais em geral não possuem estruturas e padrões conhecidos de antemão, e apresentam problemas como valores ausentes e ruído, tornando difícil a sua utilização para avaliar novas técnicas. A seguir são descritos os conjuntos de dados reais usados neste trabalho.

Cancer.data – Contém 457 registros com onze atributos cada, sendo que os dez primeiros atributos referem-se a informações de exames para diagnóstico de câncer. O primeiro atributo informa o número da amostra, os outros 9 são: *ClumpThickness*, *UniforSize*, *UniforShape*, *MargAdhes*, *SingleEpithSize*, *BareNuclei*, *BlandChromatin*, *NormalNucleoli*, *Mitoses*. O último atributo identifica a classe do tumor (maligno ou benigno) [Bennett e Mangasarian, 1992]. Pode ser obtido no endereço www.discover.nci.nih.gov/nature2000/.

Cars.data – Contém 406 registros com sete atributos contendo informações sobre automóveis de origem americana, européia e japonesa. Os atributos são: consumo, em milhas por galão (*MPG*), número de cilindros do motor (*Cyl*), potência do motor em cavalos (*HP*), peso do veículo (*Weight*), aceleração (*Accel*), ano de fabricação (*Year*) e origem do carro (*USA*, *Europa* ou *Japão*) (disponível em <http://www.cs.uta.fi/~hs/pcef/>).

Iris.data – Possui 150 registros com quatro atributos, referentes às medidas sobre três espécies de flores (*iris virginica*, *iris setosa* e *iris versicolor*), possuindo assim três classes. Esse conjunto de dados é muito utilizado na avaliação de algoritmos de classificação, cuja tarefa, em geral, consiste em identificar a classe da flor com base nas quatro medidas. Uma

das classes é linearmente separável das demais, o que não ocorre com as outras duas. Estes dados podem ser obtidos no *UCI Machine Learning*.

Folhas.data – Contém 1.197 registros com 20 atributos (*classe, área, perímetro, df1, df2, df3, df4, df5, df6, df7, df8, df9, m1, m2, m3, m4, m5, m6, m7, circularidade*). O primeiro atributo indica as nove espécies de folhas (classes), os demais são: a área e perímetro, nove descritores de Fourier, sete momentos e a circularidade. Os dados foram obtidos extraindo as medidas de 133 folhas de cada espécie, as quais foram digitalizadas em um *scanner*. Um exemplar de cada uma das classes de folhas é apresentado na Figura 2.2.

Figura 2.2 – Nove espécies de folhas descritas no conjunto de dados *Folhas.data*.

Houses.data – Contém 20.640 registros com nove atributos, sendo que os dados foram extraídos do Censo Americano de 1990. Os atributos são: *Median House price, Median income of individual living in block, Housing median age, Total rooms in block, Total*

bedrooms in block, Population in block, Number of households in block, Latitude e Longitude. Tem sido utilizado para testar classificadores, sendo que o objetivo pode ser obter um modelo de classificação usando apenas os últimos oito atributos, e avaliando os resultados por meio do primeiro atributo, o preço médio. Por exemplo, pode-se considerar que, se o preço médio é superior a US\$180.000 a casa é de alto valor, caso contrário é de baixo valor (impondo duas classes). Disponível no *StatLib*.

Out5d.oke – Contém 16.384 registros com cinco atributos (*Spot, Magnetics, Potassium, Thorium e Uranium*) obtidos a partir de imagens de satélite (sensoriamento remoto) da região oeste da Austrália. Pode ser obtido no endereço <http://davis.wpi.edu/~xmdv/datasets.html>.

Image Segmentation.data – Criado pelo *Vision Group*, da Universidade de Massachusetts, contém 2.310 registros com 20 atributos. O último atributo corresponde às classes, enquanto que os primeiros dezoito descrevem características de regiões 3x3 em imagens. Os atributos são: *region-centroid-col, region-centroid-row, region-pixel-count, short line density-5, short-line-density-2, vedge-mean, vegde-sd, hedge-mean, hedge-sd, intensity-mean, rawred-mean, rawblue-mean, rawgreen-mean, exred-mean, exblue-mean, exgreen-mean, value-mean, saturatoin-mean, hue-mean*). As classes existentes são: *1-brickface, 2-sky, 3-foliage, 4-cement, 5-window, 6-path e 7-grass*. Disponível no *UCI Machine Learning*.

2.4.2 Conjuntos de Dados Sintéticos

Conjuntos de dados sintéticos são úteis para demonstrar características particulares de técnicas de visualização, pois podem ser construídos sob medida para avaliar aspectos específicos das técnicas. Características como número de elementos, dimensionalidade, quantidade e formato de grupos, e nível de ruído, podem ser controladas. Neste trabalho vários conjuntos de dados sintéticos são usados para testar as propostas apresentadas e verificar o seu comportamento em diferentes situações.

Synth1.data – Contém 7.500 registros com cinco atributos cada, sendo que 2.920 deles estão distribuídos em cinco grupos da seguinte maneira: grupo 1 com 848 registros, grupo 2 com 728, grupo 3 com 8 registros, grupo 4 com 608 registros e grupo 5 com 728 registros. Os demais 4.580 registros foram gerados aleatoriamente, constituindo aproximadamente 61% de ruído uniforme. Este conjunto foi gerado conforme recomendações da Seção 2.4.3.

Synthetic Data Set 2 – Gerado por Dembélé e Kastner [2003], possui 480 registros de 12 atributos, sendo que o primeiro é um identificador do registro e o segundo do grupo ao qual o registro pertence. Os outros dez atributos são as coordenadas dos registros, que formam 14 grupos. Pode ser obtido em <http://www.igbmc.u-strasbg.fr/projects/fcm>.

Synth3.data - Possui 38.850 registros de 20 atributos, sendo que 14.831 formam 8 grupos conforme descrito na Tabela 2.3. Os demais 24.019 registros constituem aproximadamente 61% de ruído uniforme. Este conjunto foi gerado conforme recomendações da Seção 2.4.3.

Tabela 2.3 – Grupos do conjunto de dados *Synth3.data*.

#Grupo	#Registros	#Grupo	#Registros
1	2272	5	2.271
2	1.521	6	1.352
3	2.272	7	2.272
4	1.518	8	1.353

Synth4.data – Possui 256.250 registros de 40 atributos, sendo que 120.874 formam 30 grupos, constituídos conforme a Tabela 2.4. Os demais 135.376 registros constituem quase 53% de ruído uniforme. Este conjunto foi gerado conforme recomendações da Seção 2.4.3.

Tabela 2.4 – Grupos do conjunto de dados *Synth4.data*.

#grupo	#registros	#grupo	#registros	#grupo	#registros	#grupo	#registros
1	5.168	9	5.173	17	4.435	25	4.435
2	2.959	10	4.429	18	2.960	26	4.435
3	4.430	11	4.437	19	5.174	27	5.167
4	5.174	12	4.433	20	2.631	28	2.960
5	5.172	13	3.696	21	3.692	29	2.961
6	2.958	14	3.693	22	3.695	30	3.698
7	3.693	15	4.435	23	3.698		
8	3.691	16	2.961	24	4.431		

Pollen.data – Disponível no *StatLib*, possui 3.848 registros com cinco atributos cada, dos quais 99 registros definem 6 grupos que formam as letras da palavra EUREKA. Os demais 3.749 registros contém valores aleatórios, constituindo aproximadamente 97% de ruído gaussiano.

Quadruped Mammals1.data – Conjunto sintético [Gennari, Langley e Fisher, 1989] com 96.311 registros com 72 atributos, referentes a aproximações de oito membros (cabeça, cauda, quatro pernas, torso e pescoço) de quatro animais (1-gato, 2-cachorro, 3-cavalo e 4-girafa). Cada membro é aproximado por um cilindro, e representado por nove atributos

(posições x,y,z ; eixos x,y,z ; altura, raio e textura). Bastante utilizado para treinamento de classificadores, este conjunto de dados não contém ruído. A rotina para gerar este conjunto está disponível no *UCI Machine Learning*.

Quadruped Mammals2.data – Variação do anterior, possui 158 registros com os mesmos 72 atributos.

Hypercube.data – Contém 1.281 registros de quatro atributos, que descrevem as coordenadas de quarenta pontos distribuídos uniformemente sobre as 28 arestas de um hipercubo (4D). Este conjunto foi gerado para testes neste trabalho.

Synth_Sierp.data – Contém 19.683 registros de cinco atributos (X_SIERP , Y_SIERP , $RAND$, $CORR_XY$, $CORR_RAND$). Os dois primeiros atributos correspondem às coordenadas x,y do Triângulo de Sierpinski. $RAND$ é um valor aleatório entre 0.0 e 1.0, $CORR_XY = X_SIERP^2 + Y_SIERP$, portanto, correlacionado com os dois primeiros atributos. $CORR_RAND = \sqrt{5 * RAND / 3}$, portanto, correlacionado com $RAND$. O conjunto é introduzido por Razente [2004].

2.4.3 Processo de Geração de Dados

Conforme mencionado, a geração de dados sintéticos permite obter conjuntos com características especiais para cada análise, como o formato dos agrupamentos (circular ou alongado), o tipo de ruído (gaussiano ou uniforme) e a estrutura espacial, como cubos ou esferas n -dimensionais. Possíveis parâmetros de entrada para um procedimento de geração de dados são: dimensionalidade, número de grupos, número de registros em cada grupo, faixa de valores de cada atributo, distribuição dos dados em cada grupo, raio dos grupos em cada dimensão, nível e tipo de distribuição do ruído.

A geração de grupos pode ser feita, por exemplo, criando um registro, denominado semente, cujos atributos recebem valores aleatórios determinados pela posição do grupo em cada dimensão. Em seguida, a partir da especificação pelo usuário do número de registros em cada grupo, são gerados os demais registros, acrescentando um valor aleatório – determinado pelo raio e distribuição do grupo em cada dimensão – aos valores dos atributos do registro semente. Assim, todos os registros de um grupo são gerados a partir do registro semente. Além de grupos, outras estruturas podem ser geradas, como no conjunto de dados *Hypercube.data*, cujos registros descrevem coordenadas de um hipercubo 4D. Outras estruturas espaciais comuns são hiper-esferas, hiper-planos e hélices. Os

atributos recebem valores numéricos aleatórios, gerados por uma função do tipo *random()*, mas de maneira a atender as entradas do usuário.

2.5 Classificação das Técnicas

As técnicas de visualização de dados multidimensionais podem ser classificadas de diferentes maneiras, segundo diferentes critérios. Keim [2002] sugere uma classificação segundo três critérios, ilustrada na Figura 2.3: a natureza do dado a ser visualizado, a abordagem de mapeamento adotada pela técnica, e os métodos de interação e distorção usados para manipular a representação visual.

Figura 2.3 – Classificação das Técnicas de Visualização [Keim, 2002].

Conforme ilustra a figura, dados podem ser:

Unidimensionais – tais como dados temporais;

Bidimensionais – por exemplo, dados geográficos representando longitude e latitude;

Multidimensionais – apresentam mais de três atributos, como os registros em tabelas, cada qual com vários campos;

Texto e Hipertexto – tais como artigos e documentos da Web;

Hierárquicos/Grafos – tais como estruturas de diretórios e também *links* da Web;

Representações de Algoritmos e Software – que se tornam dados em sistemas de visualização voltados para apoiar tarefas de desenvolvimento e depuração de software [Bick et al., 1992], oferecendo visualizações da estrutura do código e também do fluxo executado.

As técnicas de visualização podem ser agrupadas em:

Gráficos 2D e 3D tradicionais – como os gráficos de barras, de pizza, de linhas e os tradicionais gráficos de dispersão [Bertin, 1981];

Técnicas de Projeção Geométrica – mapeiam os dados em uma representação visual fazendo algum tipo de projeção. Exemplos de técnicas nesta categoria são as Matrizes de Dispersão, Coordenadas Paralelas [Inselberg, 1985] e *RadViz* [Hoffman, 1999];

Técnicas Iconográficas – mapeiam os valores dos atributos dos dados para características visuais de um ícone, como nas Faces de Chernoff [Chernoff, 1973], Ícone *Stick* [Pickett e Grinstein, 1988] e Ícone Estrela [Ward, 1994];

Técnicas Orientadas a Pixels – mapeiam cada atributo dos dados em um pixel na tela, segundo uma cor determinada pelo valor do atributo. Exemplos expressivos desta categoria são apresentados por Keim e Kriegel [1996], Ankerst et al., [1996] e Keim [2000];

Técnicas Hierárquicas – estabelecem uma hierarquia entre os diferentes atributos para o mapeamento visual. Quando os dados não são inerentemente hierárquicos, é necessário impor uma hierarquia arbitrária para a visualização. Exemplos nessa categoria são Mundos dentro de Mundos (*World-Within-Worlds*) [Feiner e Beshers, 1990], Mapas de Árvores (*TreeMaps*) [Shneiderman, 1992] e Empilhamento de Dimensões (*Dimension Stacking*) [LeBlanc et al., 1990].

O terceiro critério de classificação considera os tipos de ações de interação e distorção suportadas:

Projeção Interativa – procura gerar diferentes projeções dos dados, para que o analista identifique as projeções mais adequadas para a análise, conforme proposto, por exemplo, no *Grand Tour* [Asimov, 1985] e no *Projection Pursuit* [Cook et al., 1995; Nason, 1995];

Filtragem Interativa – oferece ao analista a possibilidade de seleccionar intervalos de interesse para a visualização. Um exemplo desta categoria é a estratégia conhecida como Consultas Dinâmicas (*Dynamic Queries*), proposta por Shneiderman [1994];

Zoom Interativo – possibilita a visualização em diferentes resoluções, de modo que o usuário possa ter uma visão geral de todos os dados e também detalhar regiões de interesse. Um exemplo desta categoria é *TableLens* [Rao e Card, 1994; Pirolli e Rao, 1996];

Distorção Interativa – oferece uma visão geral dos dados, ao mesmo tempo em que exibe regiões de interesse em detalhes. Exemplos nessa categoria são *Perspective Wall*

[Mackinlay et al., 1991] e Visualização Hiperbólica [Lamping et al., 1995; Munzner e Burchard, 1995];

Seleção e Ligação (*Brushing and Linking*) a seleção permite selecionar registros de interesse, destacando-os dos demais, para posterior aplicação de outras operações; já a ligação estabelece uma relação entre diferentes visualizações dos mesmos dados, permitindo que ações realizadas em uma visualização sejam refletidas em outras. Por exemplo, registros selecionados em uma visualização são destacados também nas demais, o que pode favorecer a compreensão dos dados, pois estruturas não percebidas em uma visualização podem ser evidenciadas em outras.

A seguir são apresentadas técnicas representativas de visualização de tabelas de dados, agrupadas segundo o eixo correspondente na classificação de Keim.

2.6 Técnicas de Visualização de Informação

Além das técnicas mais elementares de apresentação de dados bidimensionais e tridimensionais, como gráficos de linha e de dispersão [Cleveland, 1993], existem várias técnicas para a visualização exploratória de dados multidimensionais, cada qual com vantagens e desvantagens em diferentes contextos de aplicação. Esta seção apresenta algumas técnicas clássicas e uma breve discussão sobre as vantagens e desvantagens de cada uma.

2.6.1 Técnicas 2D/3D Tradicionais

Esta categoria inclui todas aquelas técnicas amplamente utilizadas para a exibição de dados bidimensionais e tridimensionais, como gráficos de dispersão (*Scatterplots*), gráficos de linhas, de barras e outros, alguns dos quais ilustrados na Figura 2.4.

Figura 2.4 Gráficos bidimensionais e tridimensionais tradicionais.

Embora de grande simplicidade, algumas dessas técnicas, como os gráficos de dispersão, oferecem apoio eficaz para a análise visual, permitindo detectar distribuição, correlação entre os atributos, agrupamentos e outras informações. A principal desvantagem é a sua limitação quanto ao número de atributos que podem ser apresentados simultaneamente. Outras técnicas são voltadas especificamente para a apresentação dos dados e sumarização dos mesmos, e não para a exploração interativa.

2.6.2 Técnicas Orientadas a Pixel

Na categoria das técnicas Orientadas a Pixel os atributos dos dados são mapeados em pixels, os quais são coloridos segundo o valor do atributo representado. O conjunto de valores de cada atributo é apresentado em uma janela individual, de modo que para exibir m atributos a tela é particionada em m janelas (Figura 2.5). Um aspecto relevante é o arranjo espacial dos *pixels* (ou, mais precisamente, dos itens de dados) na tela, que deve apresentar boas propriedades de agrupamento, mantendo algum significado semântico. Por exemplo, Curvas de *Peano-Hilbert* podem dirigir o posicionamento dos pixels, de modo que dados relacionados sejam apresentados em posições próximas [Keim, 2000]. Correlação e dependência funcional entre atributos podem ser detectadas pela análise de regiões correspondentes em diferentes janelas.

Figura 2.5 – Visualização orientada a pixels de um conjunto de dados aleatório em que cada item de dado tem nove atributos, usando dois arranjos distintos para a disposição dos itens a) espiral e b) por eixos [Keim, 1997].

Uma desvantagem é a dependência direta em relação à resolução da tela, pois, quanto maior a dimensionalidade dos dados, maior o número de janelas e, conseqüentemente, menor o número de itens que podem ser exibidos simultaneamente.

Keim e colaboradores têm tido um papel importante no desenvolvimento e aplicação de técnicas nesta categoria [Keim e Kriegel, 1994; Keim e Kriegel, 1996; Ankerst et al., 1996; Keim, 1997; Keim, 2000].

Uma outra técnica orientada a pixel bastante conhecida é denominada Segmentos Circulares (*Circle Segments*). Proposta por Ankerst et al. [1996], adota a disposição apresentada na Figura 2.6(a). Em (b) tem-se a visualização de um conjunto de dados contendo 30.000 registros de 17 atributos. A técnica apresenta limitações para conjuntos de dados muito grandes, tornando-se necessário fazer uma amostragem dos dados.

(a)

(b)

Figura 2.6 – a) Arranjo adotado na técnica Segmentos Circulares e b) Visualização de um conjunto de dados com 50 atributos [Ankerst et al., 1996].

2.6.3 Técnicas de Projeção Geométrica

As técnicas de projeção geométrica buscam apresentar projeções ‘interessantes’ dos conjuntos de dados multidimensionais no espaço bidimensional. Entre as técnicas desta categoria destacam-se: Matrizes de Dispersão (*Scatterplot Matrices*), Coordenadas Paralelas (*Parallel Coordinates*), Gráfico Estrela (*Star Graph*), Visualização Radial (*Radial Visualization – RadViz*), e o Tubo de Dados (*Data Tube*).

A Matriz de Gráficos de Dispersão, ilustrada na Figura 2.7, exibe todos os relacionamentos entre todos os pares de atributos mostrando, simultaneamente, múltiplas projeções bidimensionais dos diferentes pares de atributos. Entretanto, a apresentação de dados de alta dimensionalidade fica comprometida devido à redução da área de cada *scatterplot*, o que pode ser minimizado com o uso de técnicas de interação como o *zoom*, conforme ilustra a Figura 2.7. A descoberta dos relacionamentos bidimensionais contribui para o entendimento geral dos dados e também pode apoiar procedimentos de redução da dimensionalidade.

Figura 2.7 – Matriz de gráficos de dispersão do conjunto de dados *Pollen.data* (o gráfico de dispersão dos atributos 2 e 3 é ampliado à direita) [Wegman e Luo, 1996].

Outra técnica bastante conhecida é Coordenadas Paralelas [Inselberg, 1985; Inselberg e Dimsdale, 1990; Inselberg, 1997], na qual um espaço de dimensão n é mapeado para um espaço bidimensional usando n eixos equidistantes e paralelos a um dos eixos principais. Cada eixo representa um atributo e, normalmente, o intervalo de valores de cada atributo é mapeado linearmente sobre o eixo correspondente. Cada item de dado é exibido como uma linha poligonal que intercepta cada eixo no ponto correspondente ao valor do atributo associado, conforme mostra a Figura 2.8.

Figura 2.8 – Visualização por *Coordenadas Paralelas* do conjunto de dados *Cars.data*. Cada eixo é rotulado com o nome, o menor e o maior valor de cada atributo.

A interpretação é facilitada pela estimativa imediata dos valores dos atributos ao longo dos eixos, o que não ocorre em muitas outras técnicas. Na Figura 2.8, as setas

indicam os valores dos atributos de um registro específico. Entre outras estruturas, podem ser identificadas características como a distribuição dos dados e dependências funcionais, como correlações entre atributos [Wegman e Luo, 1996; Inselberg, 1997], assim como agrupamentos. Entretanto, devido à intensa sobreposição das linhas, a exibição simultânea de alguns milhares de registros já fica comprometida, conforme se observa na Figura 2.9(a). O número de atributos que podem ser visualizados simultaneamente também depende da resolução da tela [Fua et al., 1999; Fua et al., 1999a] e contribui para o congestionamento visual, pois é muito difícil identificar estruturas e relacionamentos entre os dados quando os eixos ficam muito próximos entre si, como ilustrado na Figura 2.9(b).

Figura 2.9 – a) Visualização do conjunto de dados *Pollen.data* com apenas 3.848 registros (cinco atributos) gera intensa sobreposição das linhas e; b) Visualização de um conjunto de dados contendo algumas centenas de registros com 40 atributos.

Ainda na categoria das técnicas de projeção geométrica, uma estratégia inspirada em Coordenadas Paralelas (ver Figura 2.10(a)) é adotada no Gráfico Estrela [Sobol e Klein, 1989; Parsaye e Chignell, 1993; Hoffman, 1999], (ver Figura 2.10(b)). A visualização é construída com n eixos arranjados em uma disposição radial, em que os itens de dados são representados por poligonais que interceptam cada eixo na posição correspondente aos valores dos atributos associados. Chou et al. [1999] observam que, quando o número de atributos é alto, os eixos ficam muito próximos na parte central, dificultando a análise.

Figura 2.10 – a) Obtenção da técnica Gráfico Estrela a partir da técnica Coordenadas Paralelas [Hoffman, 1999]; e b) Visualização de dois registros de dimensão oito utilizando o Gráfico Estrela.

Outra técnica geométrica que adota um arranjo radial é o *RadViz* (*Radial Coordinates Visualization*) [Hoffman, 1999]. Para uma visualização n -dimensional, n linhas emanam radialmente do centro de um círculo e terminam no seu perímetro, como ilustrado na Figura 2.11. Para cada atributo, constantes de atração (um sistema imaginário de molas) são associadas aos valores, sendo que a posição final do marcador visual será aquela em que ocorre o equilíbrio das forças sobre o marcador. O mapeamento resultante constitui uma transformação não linear do espaço original, que preserva algumas simetrias.

Grinstein et al. [2001] salientam que as principais características desta técnica são: registros cujos atributos têm valores iguais são posicionados no centro do sistema de eixos; registros com valores de atributos iguais associados a eixos em posições opostas no gráfico são posicionados próximos ao centro; registros em que um dos atributos tem valor dominante em relação aos demais são mapeados próximos ao eixo correspondente a este atributo. A principal vantagem do *Radviz* é que registros similares no espaço n -dimensional são mapeados próximos entre si no espaço $2D$, o que favorece a identificação de agrupamentos na visualização. Além disso, apresenta baixa complexidade computacional $O(mn)$, permitindo a sua aplicação a grandes conjuntos de dados de alta dimensionalidade. Entre as desvantagens, pode-se citar a intensa sobreposição dos marcadores e o congestionamento visual excessivo quando aplicado a grandes conjuntos de dados; e que é difícil, senão impossível, deduzir o valor dos atributos a partir da posição dos marcadores na visualização, como ocorre, por exemplo, em Coordenadas Paralelas. No caso da análise de agrupamentos, a principal desvantagem é que registros muito diferentes

entre si podem ser mapeados em posições próximas, o que precisa ser tratado de algum modo em etapas posteriores.

Figura 2.11 – Visualização do conjunto de dados *Cars.data* no *RadViz*.

Na Figura 2.11 é possível verificar que os automóveis de mesma origem ficam próximos entre si, como indicado pelos contornos tracejados. Entretanto, vale observar que diferentes configurações dos eixos produzem diferentes posicionamentos dos marcadores, conforme ilustra a Figura 2.12. Uma proposta para definir uma disposição dos eixos capaz de gerar visualizações por *RadViz* e por Coordenadas Paralelas que favorecem a percepção de agrupamentos é apresentada no Capítulo 7, e constitui uma das contribuições deste trabalho.

Figura 2.12 – a) Mapeamento de registros n -dimensionais no *RadViz*; b) Conjunto de dados contendo quatro registros (r_0 , r_1 , r_2 e r_3) com quatro atributos (a_0 , a_1 , a_2 e a_3); c) Visualização dos registros usando a disposição de eixos a_0 , a_1 , a_2 e a_3 ; d) Visualização usando a disposição a_0 , a_2 , a_1 e a_3 .

A técnica *Star Coordinates* [Kandogan, 2001] estende o *RadViz*, permitindo ao usuário controlar a direção e o comprimento dos eixos radiais, bem como selecionar faixas de interesse sobre os eixos. A possibilidade de manipular interativamente a disposição e o mapeamento dos eixos apóia a busca por uma projeção mais adequada. Entretanto, a interação é dificultada quando o número de atributos é muito alto, o que torna impraticável a tarefa de encontrar uma boa configuração para os eixos. A Figura 2.13 mostra a visualização do conjunto de dados *Íris.data* usando o *Star Coordinates*.

Figura 2.13 – a) Visualização do conjunto de dados *Íris.data* no *Star Coordinates*; b) Visualização obtida após interação do usuário com os eixos.

Os problemas de oclusão, que ocorrem no *RadViz* e também no *Star Coordinates*, são decorrentes de duas situações distintas. Primeiramente, registros distintos podem ser mapeados na mesma posição da tela mesmo se tratando de pequenos volumes de dados. Nesse caso é comum empregar um recurso denominado *jittering*, em que um novo posicionamento dos marcadores é obtido aplicando uma pequena perturbação aleatória. Na segunda situação a sobreposição é decorrente da necessidade de apresentar muitos registros na área limitada da visualização, e nesse caso o *jittering* não ameniza o problema, sendo necessário identificar a sobreposição dos marcadores e tratar o problema de outra forma. Uma das contribuições deste trabalho, apresentada no Capítulo 5, sugere exibir os marcadores com o nível de brilho proporcional ao nível de sobreposição. As estratégias apresentadas são baseadas em abordagens similares propostas para técnicas específicas [Miller e Wegman, 1990; Han e Cercone, 2000].

Ankerst [2000] propõe uma técnica denominada Tubo de Dados (*Data Tube*) para a mineração visual de grandes conjuntos de dados, a qual é inspirada na técnica Segmentos Circulares. No Tubo de Dados os valores dos atributos são projetados no interior de um tubo (3D), conforme disposição apresentada na Figura 2.14(a), sendo que o comprimento

do tubo é determinado pelo número de registros do conjunto de dados. A visualização de um conjunto com seis atributos é ilustrada na Figura 2.14(b).

Figura 2.14 – a) Disposição dos dados no Tubo de Dados; b) Visualização de alguns registros de um conjunto de dados com seis atributos [Ankerst, 2000].

O Tubo permite ao usuário navegar pela representação tridimensional, entrando e saindo do seu interior. A principal vantagem sobre a técnica Segmentos Circulares é a capacidade de tratar conjuntos de dados muito grandes, pois todos os registros podem ser apresentados ao usuário durante a navegação pelo interior do tubo.

2.6.4 Técnicas Iconográficas

A proposta das técnicas iconográficas é mapear os atributos de cada item de dado multidimensional em atributos visuais de um ícone. Na abordagem clássica denominada Faces de Chernoff (*Chernoff faces*) [Chernoff, 1973; Tufte, 1983], dois atributos são mapeados para as duas dimensões espaciais da tela, e os demais são mapeados para propriedades visuais de um ícone na forma de uma face estilizada, como o formato da boca, nariz, olhos etc. Valores diferentes nos atributos resultam em diferentes formas e posições dos componentes (ver Figura 2.15(a)). Essa abordagem explora a capacidade humana de reconhecer e analisar faces, mas vários autores [Keim e Kriegel 1996; Chou et al., 1999] observam que, devido à dificuldade de distinguir diferenças muito pequenas nas imagens resultantes, ela não é adequada para a identificação de agrupamentos. Outros ícones podem ser usados, como as ‘estrelas’ no *Star Glyphs* [Ward, 1994] (Figura 2.15(b)).

Figura 2.15 – a) Faces de Chernoff (*Chernoff Faces*) [Chernoff, 1973]; b) Ícones Estrela (*Star Glyphs*).

Outra representação icônica, denominada Configuração de Arestas (*Stick Figures*), foi apresentada por Pickett e Grinstein [1988] (Figura 2.16(a)). Duas dimensões dos dados são mapeadas para as coordenadas da tela, e as demais são mapeadas em ângulos e comprimentos das arestas do ícone. Feito um mapeamento adequado, variações qualitativas nos dados podem ser detectadas facilmente na textura que se forma (Figura 2.16(b)).

Figura 2.16 – a) *Stick Icon*; b) Visualização de dados 5-dimensionais relativos à região dos Grandes Lagos²

O número de atributos que pode ser apresentado por técnicas iconográficas é determinado pelos atributos visuais perceptíveis do ícone adotado, e a quantidade de itens que pode ser visualizada simultaneamente é muito limitada. Essas técnicas são adequadas para a detecção de padrões e variações bruscas nos dados, mas são menos efetivas para realizar comparações.

² <http://ivpr.cs.uml.edu/IVPR/gallery/lineicons.html>

2.6.5 Técnicas Hierárquicas

Técnicas hierárquicas operam subdividindo o espaço n -dimensional em subespaços de menor dimensão, exibindo os subespaços de forma hierárquica. É interessante observar que essas técnicas estabelecem uma hierarquia para a exibição, independentemente dos dados apresentarem uma estrutura hierárquica inerente. Exemplos representativos dessa estratégia são a técnica Mundo dentro de Mundos (*Worlds-within-Worlds*) [Feiner e Beshers, 1990], Empilhamento de Dimensões (*Dimension Stacking*) [LeBlanc et al., 1990] e Mapas de Árvores (*Treemaps*) [Shneiderman, 1992].

A Figura 2.17 ilustra a técnica Mundo dentro de Mundos, exibindo uma função de cinco variáveis $f(x_1, x_2, x_3, x_4, x_5)$ (seis atributos) no espaço tridimensional. No caso, valores de $f(x_1, x_2)$ definem uma superfície (mundo interno) a partir de um posicionamento obtido com os atributos x_3, x_4 e x_5 (mundo externo).

Figura 2.17 – Dados de dimensionalidade seis mapeados para o espaço tridimensional utilizando a técnica Mundo dentro de Mundos. No caso, os valores dos atributos x_3, x_4 e x_5 são mantidos constantes [Beshers e Feiner, 1993].

A Figura 2.18 ilustra a técnica Empilhamento de Dimensões [LeBlanc et al., 1990], que opera de modo semelhante às técnicas iconográficas, no sentido que a visualização explora os padrões gerados com a apresentação dos marcadores. A técnica mapeia os registros para posições de uma grade bidimensional que dependem dos valores dos atributos. Diferentes configurações da grade conduzem a diferentes resultados, conforme ilustra a Figura 2.18(a) e (b).

Figura 2.18 – Técnica Empilhamento de Dimensões usando duas configurações diferentes. A cor também pode codificar informação [LeBlanc et al., 1990].

A Figura 2.19 ilustra a técnica Mapas de Árvore [Shneiderman, 1992], projetada para exibir estruturas de árvores. Quando utilizada para representar diretórios de discos, esta técnica permite, por exemplo, identificar rapidamente o espaço ocupado por cada arquivo ou subdiretório [Shneiderman, 1992].

Figura 2.19 – a) Árvore a ser visualizada usando b) a técnica Mapa de Árvore [Shneiderman, 1992].

2.7 Técnicas de Interação e Distorção

Em processos de exploração visual, recursos de interação constituem ferramentas essenciais, permitindo ao analista selecionar diferentes aspectos dos dados, ou da representação visual, alterando dinamicamente a visualização. A seguir são abordadas algumas técnicas de Interação e Distorção, como Projeção Interativa, Filtragem, *Zoom*, Distorção, Seleção e Ligação, segundo as categorias identificadas na Figura 2.3.

2.7.1 Técnicas de Projeção Interativa

Essa categoria inclui técnicas que procuram exibir diferentes projeções dos dados multidimensionais em uma tela bidimensional. Técnicas de projeção auxiliam o mapeamento de dados de alta dimensionalidade em espaços de dimensionalidade menor, sendo que nem todas as abordagens são interativas (ver Seções 3.2 e 4.2). Entre as abordagens interativas mais conhecidas estão o *Grand Tour* [Asimov, 1985; Wegman e Luo, 1996] e *Projection Pursuit* [Cook et al., 1995; Nason, 1995]. A primeira procura identificar, automaticamente, as projeções bidimensionais mais interessantes, usando uma animação para exibir a sequência das projeções em uma determinada ordem. Trata-se de uma solução lenta e computacionalmente cara no caso de dados de alta dimensionalidade.

A Figura 2.20(a) mostra a visualização inicial dos registros pertencentes aos grupos do conjunto de dados *Pollen.data* por Coordenadas Paralelas, enquanto em 2.20(b) tem-se a disposição obtida em uma das projeções geradas pelo *Grand Tour* [Wegman e Luo, 1996], que apresenta uma separação melhor dos seis agrupamentos presentes nos dados.

Figura 2.20 – Visualização dos registros pertencentes aos grupos do conjunto de dados *Pollen.data* a) antes e b) após a aplicação do *Grand Tour* [Wegman e Luo, 1996].

Projection Pursuit, por sua vez, tenta identificar e exibir apenas as projeções potencialmente mais interessantes, de acordo com uma função de interesse, escolhida de modo a maximizar a eficiência da projeção sob algum aspecto. Um exemplo bastante conhecido desse tipo de operação é a Análise de Componentes Principais (*Principal Component Analysis – PCA*) [Pearson, 1901], que pode ser interpretada como um caso particular de *Projection Pursuit* [Bolton e Krzanowski, 1999], adotando como função de interesse aquela que maximiza a variância dos dados projetados. Uma implementação de *Projection Pursuit* (2D) é oferecida no sistema *XGOBI* [Buja et al., 1996], disponível no endereço www.start.unipg.it/stat/stalib/general/Xgobi/. A Figura 2.21 mostra uma projeção bidimensional de um conjunto de registros de dimensionalidade seis.

Figura 2.21 – Exibição de dados do censo inglês de 1991. No caso, a projeção não revela a presença de grupos, sugerindo que esta característica não ocorre ou não é evidenciada em duas dimensões [Brunsdon et al., 1999].

Todas as técnicas de redução da dimensionalidade abordadas na Seção 4.2 podem ser usadas para apoiar uma projeção interativa. Embora na Figura 2.21 o resultado da projeção tenha sido exibido por Gráficos de Dispersão, outras representações poderiam ser utilizadas, sobretudo se a dimensionalidade final é superior a três. Diversas outras técnicas de projeção de dados em espaços de baixa dimensionalidade, não necessariamente interativas, são abordadas no Capítulo 4, que introduz uma nova abordagem para projetar dados multidimensionais no espaço tridimensional. Outra contribuição deste trabalho (iv) é uma estratégia para gerar uma sequência de apresentação dos atributos, ou de disposição de eixos em visualizações por projeção geométrica, apresentada no Capítulo 7. A estratégia sugerida contribui para reduzir o congestionamento visual nas visualizações e viabiliza um processo automático de redução de dimensionalidade, usando como critério a eliminação de atributos altamente correlacionados.

2.7.2 Filtragem

Técnicas de Filtragem possibilitam ao analista selecionar interativamente regiões de interesse no conjunto de dados. A seleção pode ser feita diretamente nos dados, ou pela especificação de propriedades de interesse. *Magic Lenses* [Fishkin e Stone, 1995] é um exemplo de técnica dessa categoria, atuando como uma lente de aumento sobre um documento. Os registros sob a “lente” são processados por um filtro selecionado e os resultados são exibidos em uma visualização em que os demais registros de dados não são exibidos. Os filtros são operações que podem ser aplicadas aos registros selecionados. A

Figura 2.22 ilustra a operação em um gráfico de dispersão, sendo que em (a) o filtro amplia a região, enquanto que em (b) o filtro exibe os atributos dos dados.

Figura 2.22 – Utilização de lentes para auxiliar a visualização de um gráfico de dispersão. O filtro em (a) amplia a exibição, revelando quatro elementos (cidades americanas), enquanto que em (b) o filtro exibe os dados associados aos elementos gráficos selecionados [Fishkin e Stone, 1995].

Filtragens mais complexas podem ser obtidas sobrepondo-se múltiplas lentes aos dados, provocando a combinação dos filtros utilizados. Ainda nesta classe estão as Consultas Dinâmicas [Shneiderman, 1994], que oferecem um mecanismo visual para o usuário especificar consultas sobre um conjunto de dados relacional. As consultas são especificadas por meio de controles visuais – na forma de barras de rolagem (*sliders*), botões de rádio e caixas de verificação, como ilustrado na Figura 2.23 – que permitem selecionar um subconjunto de valores, ou habilitar/desabilitar valores de interesse. Outros exemplos de filtragem interativa são apresentados em Stolle et al. [2002]. Além dos *sliders* usuais (Figura 2.23(a)), Eick [1994] sugere a inclusão de informações adicionais no próprio *slider*, por exemplo, a distribuição dos dados, exibida em um gráfico de densidade, de modo que o usuário tenha um maior controle ao delimitar os valores, conforme ilustra a Figura 2.23(b). Com dois cursores no mesmo *slider* também pode ser definida uma faixa de valores (ver Figura 2.23(b)).

Figura 2.23 – a) *Slider* comum; b) *Slider* com dois cursores e a indicação gráfica da densidade dos dados [Eick, 1994]; c) Botões de rádio e d) Caixas de verificação.

2.7.3 Zoom

Trata-se de um recurso altamente recomendado em técnicas que manipulam grandes quantidades de dados, por proporcionar uma visão geral dos dados, juntamente com visualizações detalhadas de áreas de interesse. Keim [2002] lembra que a ampliação não significa, necessariamente, apenas aumentar os marcadores gráficos. Por exemplo, em uma resolução baixa os marcadores podem ser constituídos por um único pixel, em uma resolução intermediária por uma figura que representa melhor o objeto e, em uma resolução alta, por figuras e rótulos contendo a descrição dos objetos, conforme mostra a Figura 2.24(a). A técnica *TableLens* [Rao e Card, 1994] utiliza este recurso para permitir que muitos registros sejam apresentados na tela, mapeando cada valor de atributo no comprimento de uma barra horizontal. O número de colunas é determinado pelo número de atributos, sendo que o usuário pode re-arranjar as colunas de modo a favorecer a identificação de correlações. Quando o usuário deseja mais informações sobre algum dado específico, basta clicar sobre o marcador (barra horizontal) para obter os valores numéricos, como ilustrado na Figura 2.24(b).

Figura 2.24 – a) Zoom interativo; b) Visualização utilizando *TableLens*, com Zoom Interativo aplicado sobre três registros [Rao e Card, 1994].

2.7.4 Distorção

Técnicas de distorção são baseadas na premissa de que a informação específica sendo analisada (em foco) precisa receber maior destaque (detalhamento na tela), enquanto

as demais informações podem ser apresentadas com menos detalhe. Duas possibilidades interessantes de implementação consistem em modificar diretamente o tamanho dos objetos, ou então utilizar perspectiva para diminuir o tamanho dos objetos ao redor do objeto em foco, como na técnica *Perspective Wall* [Mackinlay et al., 1991], ilustrada na Figura 2.25.

O que torna esse recurso interessante é que o usuário pode visualizar o objeto de interesse sem perder de vista o seu relacionamento com os demais objetos do conjunto. Isso é particularmente útil quando existem relacionamentos entre os objetos, como ocorre em árvores ou redes. Leung e Apperley [1994] apresentam uma revisão de métodos que podem ser utilizados para realizar a distorção.

Figura 2.25 – Distorção obtida por meio de perspectiva [Mackinlay et al., 1991].

2.7.5 Seleção e Ligação

A seleção (*brushing*), no contexto da visualização multidimensional, se refere ao processo de localizar um subconjunto de registros no conjunto original, apontando os elementos com o mouse ou outro dispositivo. Para selecionar múltiplos elementos gráficos, um retângulo pode ser usado para delimitar a área que contém os marcadores desejados. O mecanismo de seleção depende da técnica de visualização. Em gráficos de dispersão, é comum selecionar um elemento apontando-o com o mouse e clicando sobre ele (Figura 2.26(a)); ou delimitar uma região, definindo um retângulo, para selecionar múltiplos marcadores (Figura 2.26(b)). No caso de Coordenadas Paralelas o procedimento é parecido, pois o usuário pode indicar diretamente um marcador, como na Figura 2.26(c), ou selecionar uma região sobre um dos eixos, conforme mostra a Figura 2.26(d).

Figura 2.26 – Seleção de marcadores em: um gráfico de dispersão a) individualmente apontando o mouse sobre o marcador e b) definindo uma região; Coordenadas Paralelas c) individualmente apontando o mouse sobre o marcador e d) definindo uma região sobre um dos eixos.

Quando os elementos a serem selecionados não podem ser delimitados com um único retângulo, a operação pode ser realizada em múltiplos passos, de forma acumulativa, ou seja, registros selecionados em uma etapa anterior permanecem selecionados na etapa seguinte. As seleções apresentadas nessa figura são realizadas no espaço da tela, mas também é possível especificar a seleção no espaço dos dados, selecionando os registros diretamente em uma tabela.

Uma terceira possibilidade é apresentada por Fua et al. [1999; 1999a], que sugerem um método para realizar a seleção sobre uma representação hierárquica dos dados (*structure-based brushing*), que permite uma visão dos mesmos em múltiplas resoluções, útil para revelar o seu comportamento nos diferentes níveis da hierarquia. Neste caso, a seleção é feita sobre uma representação de uma árvore de hierarquia (dendograma) dos dados, conforme ilustra a Figura 2.27.

Figura 2.27 – a) Conjunto de objetos e b) Seleção especificada no espaço da hierarquia dos dados [Fua et al., 1999; Fua et al., 1999a]; c) Representação da árvore em (b), usada na ferramenta *XmdvTool* [Ward, 1994].

Diversas operações podem ser aplicadas sobre os dados selecionados, como destacar os elementos por meio de uma alteração de cor. Outras operações comuns são eliminação e agregação dos registros. De fato, para facilitar a análise, devem estar disponíveis várias operações (filtros) sobre os elementos selecionados. A determinação dos valores extremos de cada atributo, cálculo dos valores médios e também da variância dos atributos são exemplos de operações interessantes. A Figura 2.28(a) mostra a visualização em Coordenadas Paralelas do conjunto de dados *Cars.data*. Em (b), após selecionar todos os registros pertencentes a cada uma das três classes existentes (origem dos automóveis), são exibidos apenas os valores médios dos valores dos atributos.

Figura 2.28 – a) Exibição do conjunto de dados *Cars.data*; b) Exibição dos valores médios dos atributos entre os registros de mesma origem [Siirtola, 2000].

Na Figura 2.28(b) é fácil constatar que os automóveis japoneses possuem, em média, características parecidas com os europeus, embora sejam mais novos. Várias outras conclusões são aparentes; por exemplo, apesar de os carros americanos possuírem maiores

valores para o número de cilindros e potência, apresentam aceleração inferior aos demais, provavelmente, por serem mais pesados também, como revela o mesmo gráfico.

A ligação (*linking*) consiste em conectar as operações de interação efetuadas sobre múltiplas visualizações de maneira que objetos seleccionados em uma visualização sejam também destacados em outra. Dessa forma, características não evidenciadas por uma determinada técnica de visualização podem ser identificadas em outras. A Figura 2.29 mostra uma seleção em Coordenadas Paralelas refletida na Matriz de Dispersão. Registros cujos marcadores foram seleccionados (contidos no retângulo na parte inferior a direita) são automaticamente seleccionados na Matriz de Dispersão.

Figura 2.29 – Ligação (*linking*) entre as visualizações a) Coordenadas Paralelas e b) Matriz de Gráficos de Dispersão. A cor escura indica registros seleccionados, no caso, os carros americanos.

Neste trabalho foram adotadas algumas técnicas de seleção para auxiliar o processo de exploração dos dados. No caso de visualizações usando Coordenadas Paralelas é adotada a estratégia ilustrada na Figura 2.26(d), enquanto que no *RadViz* e no *Vis3D* [Artero e Oliveira, 2004] (uma nova técnica de visualização proposta no Capítulo 4) é adotada a seleção indicada na Figura 2.26(b). Ambas usam a área de um retângulo definido pelo usuário, no espaço da tela, para indicar a seleção.

2.8 Considerações Finais

Neste capítulo foram apresentados um modelo do processo de visualização e os principais tipos de dados a serem manipulados por técnicas de visualização. Também foram introduzidos os conjuntos de dados utilizados ao longo deste texto para ilustrar as técnicas de visualização. Técnicas clássicas de Visualização de Informação, bem como de interação e distorção foram apresentadas e discutidas.

Todas as técnicas de visualização discutidas neste capítulo apresentam limitações na manipulação de dados de alta dimensionalidade. No caso das Coordenadas Paralelas, o número excessivo de eixos e, conseqüentemente, a proximidade entre eles, compromete a análise visual. No *RadViz*, a dimensionalidade elevada gera uma grande concentração de eixos, principalmente, na parte central do gráfico, contribuindo para um congestionamento que pode inviabilizar a análise. Na Matriz de Dispersão, o tamanho das janelas diminui à medida que aumenta o número de atributos, dificultando a sua apresentação e análise. Quanto às técnicas iconográficas, o número de atributos dos ícones precisa ser, no mínimo, igual ao número de atributos dos dados, o que se torna difícil de tratar, no caso dimensionalidades altas. Uma alternativa para contornar o problema é reduzir o número de atributos, selecionando apenas os que são realmente relevantes para a análise. Abordagens interativas para selecionar atributos visualmente merecem atenção, pois inserem o usuário no processo e podem proporcionar melhor compreensão dos dados. Entretanto, quando a dimensionalidade é muito alta e o volume de dados é grande abordagens automáticas ou semi-automáticas podem ser mais indicadas para apoiar o usuário.

A maior parte das técnicas de visualização tradicionais adota a relação de um marcador para cada registro, o que acarreta uma intensa sobreposição no caso de grandes volumes de dados. Técnicas orientadas a pixels não sofrem esse problema de sobreposição, entretanto, a associação de cada atributo de cada registro a um pixel da tela também torna-se limitada quando o volume e a dimensionalidade dos dados são altos. Uma opção para

tratar a sobreposição consiste em utilizar estratégias capazes de realçar estruturas relevantes nos dados, ou aplicar técnicas de mineração para obter modelos ou abstrações que consigam representar os dados sem congestionar a tela, sendo que ambas foram exploradas nesse trabalho, como discutido nos capítulos subsequentes.

Por fim, Stroe et al. [2000] comentam a necessidade de integração de sistemas de gerenciamento de bancos de dados aos sistemas de visualização, como um dos caminhos para auxiliar a obtenção da desejada escalabilidade.

Capítulo 3

Mineração de Dados

3.1 Considerações Iniciais

Um dos tópicos investigados neste trabalho é a integração entre técnicas de Visualização de Informação e técnicas de Mineração de Dados como forma de apoiar a visualização exploratória de grandes conjuntos de dados de alta dimensionalidade. Embora o termo Mineração de Dados seja, muitas vezes, descrito como o processo de extração de informações (sem conhecimento prévio) de um conjunto de dados, a mineração constitui apenas uma das etapas do processo de Descoberta de Conhecimento em Bancos de Dados (*KDD – Knowledge Discovery in Databases*). Assim, enquanto a etapa de mineração tem por objetivo principal a extração de padrões contidos nos dados, o processo completo de *KDD* abrange todo o processamento (seleção, pré-processamento e transformação dos dados) necessário para que isso ocorra e, posteriormente, a avaliação e a interpretação dos padrões obtidos [Fayyad et al., 1996].

Embora o processo de descoberta de conhecimento como um todo, segundo o modelo sugerido por Fayyad et al. [1996], seja apresentado neste capítulo, a ênfase é na etapa de mineração, cujas técnicas mais usuais são introduzidas. Como a Visualização Exploratória também consiste em um processo de descoberta, observa-se muita semelhança entre os procedimentos adotados nas duas abordagens. São discutidas técnicas apresentadas na literatura que integram, de diferentes maneiras, mineração e visualização, genericamente denominadas estratégias de Mineração Visual de Dados. Algumas técnicas de mineração visual são destacadas com o objetivo de enfatizar o processo de exploração adotado. Em alguns casos, verifica-se que técnicas abordadas no Capítulo 2 foram adaptadas para exibir dados e informações complementares durante o processo de exploração. Finalmente, são feitas algumas considerações sobre o uso integrado de técnicas

de mineração e visualização para auxiliar processos de análise de dados e extração de conhecimento.

3.2 Processo de Descoberta do Conhecimento

O modelo que descreve as etapas do processo *KDD*, segundo Fayyad et al. [1996], é apresentado na Figura 3.1. O detalhamento das etapas que integram esse modelo ilustra que várias decisões precisam ser tomadas em cada passo. Como elas afetam os passos posteriores, é comum a necessidade de re-processamento de etapas anteriores, conforme indicado na figura pelas setas e linhas tracejadas, quando os resultados obtidos em uma etapa não são satisfatórios.

Figura 3.1 – Etapas do processo *KDD* [Fayyad et al., 1996].

Seleção – Nesta etapa são selecionados os dados a serem investigados, além dos atributos de interesse para o estudo em questão. Tais atributos passam a ser denominados atributos ativos, e todo o processamento posterior é baseado unicamente neles. Portanto, a sua determinação é uma operação importante para a obtenção dos resultados finais. Em uma abordagem baseada na aplicação imediata de uma técnica de visualização de informação, também é preciso definir inicialmente quais atributos serão utilizados. Neste trabalho, a seleção dos atributos é realizada imediatamente após o carregamento do conjunto de dados, o qual é armazenado em uma estrutura matricial, cujo número de linhas e de colunas são determinados pelo número de registros e atributos, respectivamente. Todos os demais processamentos são realizados sobre esta estrutura, identificada em vários locais deste documento como a matriz de dados D_{mat} . Cabe ao usuário informar quais atributos participarão do processo posterior. O usuário também pode definir manualmente a sequência de exibição dos atributos, ou usar a estratégia proposta para ordenação e redução de atributos, baseada em medidas de similaridade apresentada no Capítulo 7.

Pré-Processamento – As tarefas realizadas nesta etapa objetivam assegurar a qualidade dos dados. No caso de atributos quantitativos podem ser utilizadas diversas operações, como a verificação do nível de ruído, através de medidas de dispersão que podem identificar valores muito diferentes dos demais. Outra tarefa realizada neste passo é o tratamento de valores inválidos, o que pode ser determinado por uma incompatibilidade entre o valor encontrado e o tipo do atributo, como seria o caso de valores negativos para atributos obrigatoriamente positivos, como idade, distância, velocidade, peso, etc.

Também é importante tratar valores ausentes (*missing values*), o que pode ser feito de três maneiras. A primeira é excluir os registros que apresentam algum atributo não preenchido. Entretanto, é necessário verificar quantos registros se encontram nesta situação para avaliar a perda de informação decorrente. A segunda alternativa consiste em preencher os campos incompletos, usando uma interpolação, realizada a partir de uma análise dos demais registros. A média e mediana dos valores também podem ser utilizadas no caso de atributos quantitativos, enquanto que a moda é indicada para atributos categóricos. Neste último caso, uma outra opção é tratar o campo ausente como uma nova categoria, por exemplo, “desconhecido” [Witten e Frank, 2000]. A terceira maneira de tratar o problema consiste em excluir aqueles atributos que apresentam muitos registros com valores não preenchidos, o que consiste de fato em uma nova seleção de dados. Esta escolha precisa ser feita com bastante critério, considerando a análise a ser realizada. A etapa de pré-processamento tem consequências cruciais nas demais etapas do processo de *KDD*, e os resultados ao final do processo podem sugerir novas tentativas com diferentes configurações. Por fim, as decisões mais indicadas são aquelas que minimizam a perda de informações.

Transformação – Nesta etapa são realizadas operações como a conversão dos dados em uma representação mais adequada para manipulação pelas técnicas de mineração, por exemplo, mapeando atributos categóricos em valores numéricos. Ainda nessa fase podem ser feitas combinações de atributos para reduzir a dimensionalidade dos dados.

Embora dados de entrada de vários tipos possam ser tratados, como valores numéricos inteiros ou reais, campos representando datas ou horários, e cadeias de caracteres, em geral o processamento posterior é feito numericamente. Portanto, esta etapa contempla a conversão do tipo original para o tipo numérico. Para valores do tipo data e horário, que possuem uma ordem natural, uma conversão bastante utilizada, e disponível nas linguagens de programação atuais, consiste em obter, no caso de datas no formato

dd/mm/yyyy, o valor inteiro que corresponde ao número de dias transcorridos desde o dia 01/01/1900 (datas anteriores resultam em números negativos). Uma conversão similar pode ser aplicada a dados que informam horários. Por outro lado, cadeias de caracteres são de difícil tratamento, pois não existe um mecanismo de conversão genérico para obter um valor numérico. Embora existam métricas para medir a similaridade entre duas cadeias de caracteres, elas não garantem um mapeamento capaz de representar adequadamente a estrutura original dos dados. Uma proposta bastante conhecida consiste em gerar tabelas expandidas (*flattened tables*) com valores numéricos obtidos a partir das tabelas de valores não numéricos, como exemplificado na Figura 3.2 [Han e Kamber, 2001].

(a)	a_1	...
	<i>s</i>	...
	<i>s</i>	...
	<i>n</i>	...
	<i>s</i>	...
	<i>n</i>	...

(b)	$a_1(s)$	$a_1(n)$	...
	1	0	...
	1	0	...
	0	1	...
	1	0	...
	0	1	...

Figura 3.2 – a) Tabela original com um atributo a_1 categórico; b) Tabela expandida.

Em (a) tem-se uma tabela cujo atributo a_1 é categórico, e assume os valores *s* ou *n*, significando *sim* ou *não*. Em (b) tem-se a conversão para valores numéricos, sendo que o atributo é expandido em dois ($a_1(s)$ e $a_1(n)$), tal que $\{a_1(s) = 1 \text{ e } a_1(n) = 0 \text{ se } a_1 = s\}$ ou $\{a_1(s) = 0 \text{ e } a_1(n) = 1 \text{ se } a_1 = n\}$. Este método é viável apenas quando a cardinalidade dos atributos (número de valores possíveis) é baixa, como ocorre no exemplo. Em geral, conjuntos de dados grandes possuem atributos com alta cardinalidade, e torna-se necessário usar, por exemplo, técnicas baseadas em medidas de similaridade entre os valores não numéricos, de modo a mapear cadeias de caracteres similares em valores numéricos próximos. Várias métricas para medir a similaridade (distância) entre cadeias de caracteres foram propostas, como a distância de edição (*distance editing*), que define a distância entre duas cadeias como o número de alterações necessárias para converter uma na outra. Muitas medidas de distância podem ser utilizadas nesse contexto, cabendo uma investigação sobre o efeito da escolha nos resultados.

Em alguns casos, a informação (unidimensional) fornecida é a distância entre os elementos, mas sabe-se que estes estão posicionados em um arranjo bidimensional. Neste caso, a transformação pode ser realizada, por exemplo, usando o Escalonamento Multidimensional (*Multidimensional Scaling – MDS*) para mapear dados numéricos de uma dimensão para outra. O *MDS* usa informação de dissimilaridade e procura preservar a estrutura espacial dos dados [Shepard, 1962; Kruskal, 1964], sendo que o caso mais

comum [Kruskal e Wish, 1978] consiste em mapear cada um dos registros n -dimensionais do conjunto de dados D em um espaço k -dimensional, minimizando a função erro (função *stress*), dada por:

$$E = \sqrt{\frac{\sum_{i,j} (\epsilon_{i,j} - \delta_{i,j})^2}{\sum_{i,j} \delta_{i,j}^2}} \tag{3.1}$$

onde $\epsilon_{i,j}$ é a medida de dissimilaridade entre os registros D_i e D_j no espaço de entrada (n -dimensional) e $\delta_{i,j}$ é a distância entre as projeções de D_i e D_j no espaço de saída (k -dimensional). A função retorna o erro relativo das distâncias nos espaços n e k . Entretanto, esta técnica tem complexidade $O(m^2)$ (m é o número de registros), o que a torna pouco indicada para grandes conjuntos de dados. A Figura 3.3 [Manly, 1986] apresenta um exemplo ilustrativo em que as distâncias (1D) entre algumas cidades da Nova Zelândia são usadas para gerar um mapeamento das posições (2D) das cidades.

Figura 3.3 – a) Distâncias entre treze cidades da Nova Zelândia; b) Posicionamento obtido usando o *MDS*; c) Mapa da Nova Zelândia [Manly, 1986].

Observa-se que foi obtido um bom posicionamento relativo considerando apenas as distâncias entre as cidades, ainda que a disposição em (b) não coincida exatamente com a disposição das cidades no mapa em (c).

Nas técnicas introduzidas neste trabalho cadeias de caracteres são mapeadas de maneira similar à adotada em algumas ferramentas, como o *SpotFire Decision Site* © (www.spotfire.com). O mapeamento consiste simplesmente em listar os valores dos atributos em ordem alfabética e, em seguida, associar valores inteiros consecutivos a cada um deles, para efeito de estabelecer uma seqüência de exibição ao longo dos eixos. A Figura 3.4(a) mostra seis elementos de um conjunto de dados contendo um atributo do tipo cadeia de caracteres (*string*), que tem cardinalidade 4. Em (b) os valores do atributo a_1 são ordenados e associados a valores inteiros consecutivos entre 1 e 4. Em (c) tem-se o mapeamento final.

(a)	a_1	...
	aa	...
	ba	...
	ca	...
	baa	...
	aa	...
(b)	mapeamento	
	aa	1
	baa	2
	ba	3
(c)	a_1	...
	1	...
	3	...
	1	...
	4	...
	1	...

Figura 3.4 – a) Registros contendo um atributo categórico; b) determinação de índices para cada valor do atributo a_1 ; c) Mapeamento aplicado ao atributo em (a).

Vale observar que as técnicas de transformação procuram encontrar bons mapeamentos, geralmente baseando-se em uma análise que envolve os demais atributos dos registros. No caso de visualizações por Coordenadas Paralelas, resultados ruins da etapa de transformação não comprometem a visualização. Entretanto, no caso do *RadViz* ou *Vis3D*, que posicionam os marcadores a partir dos valores obtidos na etapa de transformação, as visualizações obtidas podem ser muito prejudicadas.

Mineração – Esta etapa utiliza algoritmos para extrair padrões a partir dos dados, podendo ser utilizados diversos métodos conforme o tipo de análise a ser feita e a natureza dos dados. Alguns algoritmos mais conhecidos são brevemente apresentados nas Seções 3.3 e 3.4.

Análise ou Interpretação dos Resultados – Nesse passo os resultados são analisados por especialistas do domínio para verificação de consistência. Também é definido como as informações obtidas serão apresentadas e utilizadas para auxiliar tomadas de decisão. Essa fase também pode realimentar o processo todo, oferecendo novas informações que podem

ser novamente processadas nas etapas anteriores, em uma tentativa de refinamento dos resultados.

Fayyad et al. [1996] salientam que todas as etapas do processo de descoberta são bastante interativas, principalmente, porque a experiência do analista é fundamental para a obtenção de bons resultados. Na etapa de mineração, normalmente, cabe também ao analista definir, com base em sua experiência, a técnica mais indicada em cada situação.

3.3 Técnicas de Mineração de Dados

A intensa utilização de métodos estatísticos tradicionais em mineração de dados tem ocasionado uma certa dificuldade de distinção entre as duas áreas. Em uma tentativa de definir a mineração de dados, Cabena et al. [1998] afirmam que ela pode ser vista como uma adaptação de técnicas estatísticas e de outras áreas, como o Aprendizado de Máquinas, o Reconhecimento de Padrões e a Visualização de Dados, objetivando a análise de grandes bancos de dados, constituindo-se, portanto, em uma área bastante interdisciplinar.

As várias técnicas de mineração podem ser classificadas de maneiras distintas, de acordo com diferentes critérios, sendo que um deles é o tipo de informação extraída [Chen et al., 1996]. Por esse critério, as técnicas são agrupadas em Regras de Associação, Sumarização, Classificação e Agrupamento.

3.3.1 Regras de Associação

Essa abordagem procura identificar associações entre os registros, de tal modo que a presença de um atributo implica na presença de outro, na tentativa de expor características e tendências [Cabena et al., 1998]. Uma regra de associação exibe em forma textual, os atributos que implicam a presença de outros, possuindo a forma “Se A , então B ” ($A \Rightarrow B$), em que A é denominado o corpo da regra e B a cabeça da regra. Os algoritmos para a obtenção de regras de associação trabalham principalmente com medidas estatísticas dos dados [Witten e Frank, 2000], sendo muito eficientes na obtenção das regras. Além das regras, os algoritmos retornam dois parâmetros importantes: o fator de confiança e o fator de suporte.

O fator de confiança indica a proporção em que a regra é verdadeira entre os registros individuais, sendo dado pela razão entre o número de registros que sustentam a regra e o número de registros que sustentam somente o corpo da regra. O fator de suporte

indica a ocorrência relativa da regra no conjunto completo dos dados, sendo definido pela razão entre o número de registros que sustentam a regra e o número total de registros. Assim, uma regra $A \Rightarrow B$ tem confiança c , se $c\%$ dos registros que contêm A também contêm B , e suporte s se $s\%$ dos registros da base do conjunto dados satisfazem $A \Rightarrow B$. O fator de confiança e o fator de suporte são importantes para auxiliar o analista na escolha das regras mais confiáveis.

Em geral, buscam-se regras capazes de explicar melhor o comportamento dos dados, que apresentam uma configuração do tipo " $A_1 \wedge \dots \wedge A_n \Rightarrow B_1 \wedge \dots \wedge B_m$ ". O processo de busca das regras consiste em dois passos:

- 1 – Descobrir o conjunto de registros que possuem fator de suporte maior do que um valor mínimo estabelecido;
- 2 – Usar o conjunto obtido no passo 1 para construir as regras.

Chen et al. [1996] observam que o desempenho do processo é determinado, principalmente, pelo primeiro passo, que manipula diretamente o conjunto de dados original. Dois algoritmos clássicos para a identificação de regras de associação são o *Apriori* [Agrawal e Srikant, 1994] e o *DHP (Scaling for Direct Hashing and Pruning)* [Park et al., 1995].

3.3.2 Sumarização

Amplamente utilizada em ferramentas de mineração, a sumarização consiste em oferecer descrições mais compactas dos dados. Pode ser utilizada na etapa de pré-processamento, corrigindo valores incorretos e preenchendo campos vazios. Medidas estatísticas como média e moda são exemplos de sumarizações simples [Fayyad et al. 1996]. A Figura 3.7 na Seção 3.4 especifica uma situação em que a exibição dos marcadores correspondentes às médias dos registros em cada classe, resume bem o comportamento dos dados [Siirtola, 2000]. Técnicas de visualização também podem ser consideradas alternativas para a sumarização, pois, em várias situações, conseguem realçar padrões e tendências existentes nos dados.

3.3.3 Classificação

A Classificação divide os dados em classes considerando os valores de seus atributos. Assim, o processo consiste em particionar o espaço dos atributos de tal modo que os registros pertencentes a uma certa classe sejam colocados sempre na partição

correspondente. Entretanto, este objetivo nem sempre pode ser atingido, assim, o procedimento habitual consiste em definir as partições minimizando algum critério. A classificação pode ser supervisionada ou não supervisionada.

Na classificação supervisionada um conjunto de registros (chamado conjunto de treinamento) é utilizado para definir as classes. O usuário informa as classes de cada registro, de modo a caracterizar as classes. Após essa etapa, outros registros poderão ser alocados nas suas devidas classes, por meio de uma análise dos seus atributos.

Na classificação não supervisionada, também conhecida por Agrupamento e discutida na Subseção 3.3.4, nenhuma informação sobre as classes é fornecida. Nesse caso, o classificador deve obter as classes buscando maximizar a similaridade entre seus registros, ao mesmo tempo minimizando a semelhança entre os registros de classes distintas, garantindo assim máxima separação entre as classes. A Classificação é utilizada na mineração de dados para predição, ou seja, para inferir novas informações em situações novas. Em Estatística, a predição com Classificação é chamada Árvore de Decisão. Um algoritmo muito conhecido de Árvore de Decisão é o *C4.5* [Quinlan, 1993]. Outro mais recente é o *SLIQ* (*Supervised Learning in QUEST*), apresentado por Mehta et al. [1996]. Projetado para a mineração de grandes conjuntos de dados, o *SLIQ* opera com atributos numéricos e categóricos.

Muitos sistemas de mineração atuais utilizam Redes Neurais na etapa de classificação. As Redes Neurais apresentam bom desempenho, principalmente, quando a saída não é linear, sendo indicadas, especialmente, quando se dispõe de grandes quantidades de dados para o treinamento da rede.

3.3.4 Agrupamento

A análise de agrupamentos procura encontrar classes para os dados, de modo que registros similares sejam colocados em uma mesma classe, maximizando a similaridade intraclasse, ao mesmo tempo em que é minimizada a similaridade interclasse. Técnicas de agrupamento têm sido utilizadas em várias áreas, com diferentes ênfases, e podem ser divididas, basicamente, em duas categorias: hierárquicas e não-hierárquicas.

Um dos algoritmos não hierárquicos mais conhecidos é o *k*-médias, que define inicialmente *k* registros para serem os centros c_j de grupos $Clust_j$. Em seguida, os demais registros são anexados aos centros mais próximos. No próximo passo os centros dos grupos são recalculados, por exemplo, usando os centróides, e o processo se repete até não ocorrerem grandes modificações nas configurações dos grupos. A principal desvantagem

do método é a necessidade de definir inicialmente o número de agrupamentos e os registros que constituirão os primeiros centros, pois esta escolha pode ser decisiva para os resultados finais.

Os métodos hierárquicos realizam o agrupamento em etapas, definindo os grupos usando árvores, também conhecidas como *dendogramas* (Figura 3.5(b)). Duas estratégias se destacam na implementação dos algoritmos nesta categoria. Na primeira, inicialmente, cada registro do conjunto de dados constitui um grupo (Figura 3.5(a)), e nos passos seguintes registros ou grupos próximos são unidos. Na segunda, todos os registros são colocados inicialmente em um único grupo, que vai sendo progressivamente subdividido nas etapas seguintes.

Figura 3.5 – a) Conjunto de dados e b) seu respectivo dendograma. Os comprimentos dos segmentos verticais do dendograma são proporcionais às distâncias entre os agrupamentos que interligam, conforme indicam $d_{b,c}$ e $d_{d,e}$.

O crescimento dos conjuntos de dados motivou o surgimento de vários algoritmos capazes de manipular grandes volumes. O algoritmo *CLARANS* (*Clustering Large Applications Based Upon RANdomized Search*) proposto por Ng e Han [1994] apresenta complexidade em cada iteração proporcional ao número de registros. Entretanto, assume que todos os registros estão armazenados na memória, o que não ocorre de fato, quando se opera com conjuntos de dados muito grandes. Outros dois algoritmos são o *BIRCH* (*Balanced Iterative Reducing and Clustering using Hierarchies*), apresentado por Zhang et al. [1996], e o *CURE* (*Clustering using Representatives*), apresentado por Guha et al. [1998]. O primeiro consegue encontrar bons agrupamentos com uma única varredura do conjunto de dados, possibilitando ainda que, caso necessário, outras varreduras sejam realizadas para refinar a resposta. O segundo é mais recente, e procura atacar alguns dos pontos fracos do primeiro. As principais melhorias são a identificação mais eficiente de agrupamentos com formas alongadas, além de um melhor desempenho.

Os algoritmos baseados em densidades são os mais indicados para tratar ruído e grupos com formas variadas, como é o caso do *DBSCAN* [Ester et al., 1996], entre outros. O algoritmo *DENCLUE* [Hinneburg e Keim, 1998] se destaca nesta categoria pela maneira de calcular a informação de densidade, baseando-se em vizinhanças. Com o objetivo de tornar o processamento simples e rápido, alguns autores propõem algoritmos que operam sobre uma grade, de modo a reduzir o espaço dos atributos, como é o caso do *OPTICS* (*Ordering Points To Identify the Clustering Structure*) [Ankerst et al., 1999]. Os algoritmos *HC-Cooperative* [Aggarwal, 2001] e *HC-Enhanced* [Gamero e Minghim, 2003] conseguem bons resultados e bom desempenho utilizando informação de densidades sobre uma grade. Na implementação do *HC-Enhanced* o usuário precisa definir oito parâmetros de entrada, e em seguida, definir visualmente o nível de ruído em cada projeção exibida, além de definir se devem ser feitas novas projeções.

3.4 Mineração Visual de Dados

Nesta seção são apresentadas algumas técnicas de mineração visual de dados, enfatizando, principalmente, o processo contínuo e interativo de descoberta de conhecimento auxiliado por representações gráficas, de modo a explorar ao máximo a capacidade de análise visual humana. As técnicas abordadas são agrupadas segundo a classificação apresentada por Ankerst [2000], que sugere três categorias de técnicas de mineração visual de dados: visualização dos dados brutos, visualização dos resultados de um processo de mineração e visualização acoplada ao processo de mineração.

3.4.1 Visualização Direta

Essa categoria engloba abordagens em que as estratégias de visualização são aplicadas diretamente aos dados. Assim, as técnicas de visualização de informação apresentadas no Capítulo 2 podem ser interpretadas como técnicas de mineração visual nesta categoria. Entretanto, no caso de grandes volumes a exibição de todos os itens de dados geralmente compromete a visualização, devido à intensa sobrecarga visual na tela. Em seguida, são apresentadas algumas estratégias que se amparam em técnicas de visualização para apoiar tarefas de exploração de dados.

Siirtola [2000] sugere dois mecanismos para a exploração de dados usando a técnica Coordenadas Paralelas. O primeiro apóia o estudo da correlação entre atributos mapeados em eixos consecutivos da visualização. No caso, o usuário seleciona os

elementos (marcadores) que deseja analisar e, em seguida, é exibido um retângulo verde sobre a linha central quando a correlação entre os eixos adjacentes é positiva, ou um retângulo vermelho abaixo da linha central quando a correlação é negativa, conforme ilustra a Figura 3.6.

Figura 3.6 – Proposta para auxiliar o estudo da correlação entre os atributos [Siirtola, 2000].

O segundo mecanismo, denominado *Polyline Averaging*, permite sumarizar, dinamicamente, um conjunto de poligonais. A Figura 3.7(a) apresenta a visualização por Coordenadas Paralelas do conjunto *Cars.data*. Em (b) são exibidos os valores médios dos atributos dos carros europeus, japoneses e americanos.

Figura 3.7 – a) Visualização em coordenadas paralelas do conjunto *Cars.data*; b) Sumarização (valor médio) das poligonais selecionadas pelo usuário em (a) [Siirtola, 2000].

Conforme mencionado anteriormente, uma visualização pode ser interpretada como uma técnica de sumarização, pois em várias situações ela é suficiente para auxiliar a identificação de estruturas presentes nos dados. Da mesma forma, a média também é uma sumarização relevante. No exemplo, Siirtola [2000] combina ambas de modo a tornar as visualizações ainda mais informativas.

3.4.2 Visualização de Modelos de Mineração

Nas técnicas dessa categoria são visualizados resultados da aplicação de um algoritmo de mineração, ou seja, o modelo extraído dos dados, que pode oferecer uma descrição condensada dos mesmos. Entretanto, nem sempre o modelo resultante da mineração é intuitivo e compreensível para o usuário, ainda que apresentado visualmente. Adicionalmente, a determinação da técnica de mineração mais adequada requer familiaridade prévia com os dados, e novas tentativas com diferentes algoritmos de mineração podem ser necessárias se esse conhecimento não está disponível.

Han e Cercone [2000] introduzem o *RuleViz*, propondo que os resultados da extração de regras de associação sejam visualizados com uma adaptação da técnica Coordenadas Paralelas. A Figura 3.8(a) apresenta as visualizações dos registros do conjunto de dados *Iris.data* por Coordenadas Paralelas. Em (b) tem-se a visualização das regras de associação usando o *RuleViz*.

Figura 3.8 – a) Visualização do conjunto de dados *Iris.data* por Coordenadas Paralelas; b) Visualização das regras de associação usando o *RuleViz* [Han e Cercone, 2000].

Ankerst [2001] introduz a técnica *OPTICS* (*Ordering Points to Identify the Clustering Structure*), que sugere exibir informação de densidade objetivando identificar grupos nos dados, a qual se apoia na técnica Segmentos Circulares (*Circle Segments*) para exibir informação da densidade dos vários atributos. A Figura 3.9(b) mostra o gráfico de densidade correspondente ao conjunto de dados bidimensional apresentado em (a). A

Figura 3.9(c) ilustra como Segmentos Circulares apresenta a informação de densidade de um conjunto com 30.000 registros de 17 atributos.

Figura 3.9 – a) Conjunto de dados (bidimensional); b) Informação de densidade; c) Visualização da densidade de um conjunto de dados com 30.000 registros com 17 atributos usando Segmentos Circulares [Ankerst, 2001].

No caso, valores elevados de densidade são representados por tons claros e valores baixos são representados com tons escuros. Os tons que mapeiam o atributo 17 (*reachability attr.*) são quantizados em apenas três níveis, para destacar a formação dos agrupamentos.

A técnica Coordenadas Paralelas Hierárquicas [Fua et al., 1999; Fua et al., 1999a] procura contornar o problema da sobrecarga visual em Coordenadas Paralelas, aplicando um algoritmo de agrupamento hierárquico aos dados, antes da visualização. É utilizado o algoritmo *BIRCH* [Zhang et al., 1996] para gerar uma hierarquia de agrupamentos, a qual permite exibir visualizações dos dados em diferentes níveis de detalhe. Assim, as linhas poligonais na visualização podem representar itens de dados individuais ou codificar informação relativa aos agrupamentos, como ilustrado na Figura 3.10. O usuário pode seleccionar diferentes níveis de detalhe variando um parâmetro que define o raio dos grupos. A Figura 3.10 mostra uma sequência de visualizações de um volume de dados contendo 230.000 registros de oito atributos, em diferentes níveis de detalhe.

Figura 3.10 - Sequência de visualizações de um volume de dados com 230.000 itens de dados, em diferentes níveis de detalhe (acima à esquerda, menos detalhe; abaixo à direita, mais detalhe: na última todos os registros do volume são exibidos) [Fua et al., 1999; Fua et al., 1999a].

A Figura 3.11 mostra a ferramenta para seleção de registros (*brushing*) implementada no sistema *XmdvTool* [Ward, 1994], como discutido na Seção 2.75, baseada na estrutura da árvore hierárquica gerada pelo algoritmo de agrupamento.

Figura 3.11 – Ferramenta para *brushing* hierárquico; (a) Representação da árvore de hierarquia; (b) Região selecionada; (c) Raios dos grupos selecionados; (d) Raios dos grupos não selecionados; (e) Contorno da árvore; (f) Delimitadores horizontais da região selecionada; e (g) Escala de cores utilizada no mapeamento [Fua et al., 1999; Fua et al., 1999a].

A Figura 3.12 ilustra a sua utilização para selecionar sub-conjuntos dos dados. No caso, foi atribuído o valor mínimo (zero) aos raios dos grupos contidos na região selecionada em (a); assim, cada registro constitui um grupo e o seu respectivo marcador é

exibido. Os raios dos grupos que contém os registros não selecionados receberam o valor máximo (1.0), com isso todos registros não selecionados são colocados em um único grupo (raio máximo). A cor sólida no centro da faixa transparente em (b) indica a posição do valor médio dos registros no grupo dos registros não selecionados, enquanto que a largura da faixa indica a extensão do grupo em cada dimensão.

Figura 3.12 – a) Utilização do *brushing* hierárquico; b) Definindo os raios dos grupos selecionados iguais a 0.0 obriga a sua exibição individual; definindo os raios dos grupos não selecionados iguais a 1.0 anexa todos os demais registros em um único grupo (região transparente) [Fua et al., 1999; Fua et al., 1999a].

3.4.3 Visualização Acoplada à Mineração

Essa categoria engloba abordagens que integram técnicas de mineração e de visualização, de modo que os passos intermediários da mineração e a construção do modelo possam ser visualizados, e o usuário possa usar tal informação para interferir no processo. Tais abordagens podem oferecer uma melhor compreensão do processo de mineração e dos dados que estão sendo manipulados, e gerar melhores resultados que um processo totalmente automático.

Ankerst [2001] apresenta uma estratégia que permite ao usuário participar da construção da árvore de decisão (Figura 3.13) em um processo de classificação. Nessa abordagem, cada atributo do conjunto de dados é exibido usando uma técnica orientada a pixel (Figura 3.13(a)), conforme o esquema ilustrado na Figura 3.13(b). Em seguida, o usuário deve posicionar visualmente os separadores (*splits-points*) que melhor separam as classes. No exemplo apresentado na Figura 3.13, o separador é colocado no atributo 1 na

Figura 3.13(a). Após definir uma partição, são geradas novas visualizações que apoiam a definição das demais partições nos próximos níveis da árvore (Figura 3.14(c) e (d)).

Figura 3.13 – a) Visualização orientada a pixel de um conjunto de dados com nove atributos, utilizando o arranjo em (b); c) Árvore de decisão; d) Visualização da árvore [Ankerst, 2001].

Hinneburg et al. [1999] apresentam a técnica *HD-Eye*, idealizada para auxiliar a identificação visual de grupos em dados de alta dimensionalidade. A técnica é baseada no algoritmo *OptiGrid* (*Optimal Grid-Clustering*) [Hinneburg e Keim, 1999] e também utiliza a análise visual de gráficos de densidade obtidos com a projeção dos dados. A Figura 3.14 ilustra, em (a), como algumas projeções são usadas para detectar os grupos e, em (b), o particionamento obtido no primeiro passo do algoritmo automático *OptiGrid*.

Figura 3.14 – Identificação dos grupos usando projeções. Em a) o grupo (3) é adequadamente identificado nas projeções paralelas aos eixos, enquanto que os grupos (1) e (2) precisam de uma projeção diagonal [Hinneburg et al., 1999]. Em b) é exibido um particionamento (*Grid*) obtido no primeiro passo do *OptiGrid* [Hinneburg e Keim, 1999].

O *HD-Eye* [Hinneburg et al., 1999], ilustrado na Figura 3.15, propõe a determinação de separadores de forma visual, a partir da informação de densidade. Como o número de projeções interessantes é muito grande, são apresentadas inicialmente todas as projeções paralelas, algumas projeções diagonais e algumas combinações aleatórias entre elas. O usuário pode, em seguida, interagir com as projeções, bem como gerar outras, em busca de projeções que separam melhor os grupos. Para auxiliar a análise dos gráficos de densidade são adotadas diferentes alternativas para visualizá-los. A primeira exibe a informação de densidade (Figura 3.15(c)) usando ícones (Figuras 3.15(a) e (b)) que mapeiam a separação, a forma e o tamanho dos grupos para a cor e a forma do ícone. A segunda exibe gráficos de densidade coloridos (Figuras 3.15(d) e (e)). No processo, o usuário deve definir os separadores conforme indicado nas Figuras 3.15(d) e (e). Em seguida, a partir dos separadores, é definida a grade multidimensional que armazena a formação dos grupos. O processo todo pode ser repetido para refinar os grupos.

Figura 3.15 – a) Ícones usados para visualizar informação de densidade; b) Ícone mapeando informação de densidade; c) Histograma da densidade obtida em uma projeção; d) Gráfico unidimensional de densidade; e) Gráfico bidimensional de densidade [Hinneburg et al., 1999].

Para tratar o problema da alta dimensionalidade Manco et al. [2004] apresentam a ferramenta *Eureka!*, que projeta dados multidimensionais no espaço bidimensional usando Decomposição por Valores Singulares (SVD) (discutida no Capítulo 4). Em seguida o usuário deve definir partições sobre as visualizações obtidas. A partir das partições é construída a árvore de decisões. A Figura 3.16 ilustra a exploração do conjunto *Image Segmentation.data* no *Eureka!*.

Figura 3.16 – a) Principais passos da exploração usando o *Eureka!*; b) Visualização do conjunto *Image Segmentation.data*; c) Usuário define uma região de interesse [Manco et al., 2004].

O algoritmo de agrupamento *HC-Enhanced*, apresentado por Gamero e Minghim [2003] também opera visualmente. A Figura 3.17 ilustra o processo de agrupamento do conjunto de dados *Synth1.data*. Inicialmente, é apresentada ao usuário a informação de densidade obtida a partir da projeção dos dados multidimensionais no espaço 3D, utilizando Análise de Componentes Principais. Em seguida, o usuário interage com o modelo tridimensional e define o nível de ruído que separa adequadamente as áreas de alta densidade. Como não se consegue bons resultados com uma única partição, o processo todo deve ser repetido algumas vezes, de modo a garantir boa separação entre os grupos. Todas as etapas calculam e exibem a informação de densidade para a projeção definida. Na Figura 3.17(a) tem-se a visualização inicial e, em (b), o usuário determina visualmente o corte que melhor separa os grupos (nível de ruído), solicitando uma nova interação. Em (c) tem-se a visualização obtida e, em (d), o usuário determina novamente o nível do corte que melhor separa os grupos (nível de ruído), solicitando em seguida novas projeções, exibidas em (e) e (f).

Figura 3.17 – Agrupamento de dados com o *HC-Enhanced* [Gamero e Minghim, 2003].

Assim como *HC-Cooperative* [Aggarwal, 2001], em cada uma das interações (projeções), os registros recebem um identificador correspondente ao número do grupo que foram inseridos. No final do processo, os registros que receberam os mesmos identificadores em todas as projeções são alocados ao mesmo grupo n -dimensional.

3.5 Comentários Finais

Neste capítulo foram apresentadas as etapas do processo de descoberta do conhecimento em bancos de dados, destacando algumas das principais técnicas de mineração de dados. Também foi discutido como técnicas de mineração podem apoiar a exploração visual de grandes conjuntos de dados, e também serem apoiadas por recursos visuais. Percebe-se que os dois problemas críticos em Visualização de Informação também são críticos em processos de Mineração de Dados. O primeiro é a alta dimensionalidade,

tipicamente tratada durante a etapa de Transformação dos Dados do processo *KDD*, por meio de técnicas de projeção (discutidas no Capítulo 2) e de seleção de atributos. O segundo problema é a grande quantidade de registros contidos nas bases de dados atuais, o que demanda o uso de algoritmos cada vez mais eficientes.

Técnicas de agrupamento aplicáveis a dados volumosos e de alta dimensionalidade foram discutidas, com ênfase em técnicas que utilizam informação de densidade para gerar os grupos, porque conseguem realçar as áreas com grandes concentrações de registros. A abordagem de Fua et al. [1999; 1999a], que aplica um algoritmo de agrupamento (*BIRCH*) aos dados e integra os resultados em uma versão hierárquica de Coordenadas Paralelas (Coordenadas Paralelas Hierárquicas), permitindo ao usuário controlar o nível de detalhe da exibição, também mereceu destaque.

Capítulo 4

***Viz3D*: Visualização Multidimensional no Espaço 3D**

4.1 Considerações Iniciais

Técnicas de visualização direta são limitadas para tratar um número excessivo de registros, ou dados de dimensionalidade muito elevada. Os problemas são decorrentes, principalmente, da necessidade de acomodar um número elevado de marcadores no espaço reduzido da tela do computador e, ao mesmo tempo, revelar ao máximo as estruturas presentes no espaço multidimensional dos dados. As técnicas de Visualização de Informação mais conhecidas, apresentadas no Capítulo 2, projetam os dados em um espaço bidimensional, o que restringe as possibilidades de interação com o modelo visual exibido. Neste capítulo é apresentada uma nova técnica de Visualização de Informação, denominada *Viz3D* [Artero e Oliveira, 2004], em que os dados são projetados no espaço tridimensional. O *Viz3D* foi inspirado na abordagem adotada pelo *RadViz*, uma técnica 2D, que foi estendida para o espaço 3D objetivando acomodar um maior número de marcadores, bem como contornar problemas de oclusão por meio de interações do usuário com o modelo 3D.

A Seção 4.2 apresenta alguns trabalhos relacionados, enquanto a Seção 4.3 apresenta o *Viz3D*, além de uma breve análise de desempenho. Alguns experimentos com conjuntos de dados apresentados no Capítulo 2 são discutidos na Seção 4.4. Os comentários finais são apresentados na Seção 4.5.

4.2 Trabalhos Relacionados

Muitas técnicas de visualização buscam projeções ‘interessantes’ dos dados n -dimensionais que possam ser exibidas na tela bidimensional. Uma alternativa bastante utilizada consiste em aplicar técnicas de redução da dimensionalidade aos dados, de modo

a obter projeções dos mesmos que podem ser exibidas nos tradicionais gráficos de dispersão 2D ou 3D. Algumas técnicas amplamente utilizadas para a redução da dimensionalidade são apresentadas a seguir. Os resultados obtidos com a aplicação de algumas delas serão comparados, posteriormente, com resultados obtidos pela projeção implementada na técnica *Viz3D*. Assume-se que um conjunto de dados contendo m registros n -dimensionais é dado na forma matricial $D_{m \times n}$ e considera-se que as técnicas mapeiam os dados para um espaço de saída k , $k < n$.

4.2.1 Análise de Componentes Principais

A abordagem mais tradicional de redução da dimensionalidade consiste na transformação rotacional (linear) utilizando componentes principais (*Principal Component Analysis – PCA*) [Pearson, 1901], obtidos a partir dos auto-vetores. O critério adotado nos cálculos procura maximizar a variância dos dados projetados. A transformação realizada, usando a notação matricial [Anderson, 1958], pode ser resumida nos seguintes passos:

Dado o conjunto de dados $D_{m \times n}$

- 1 – Obtenha a matriz de covariâncias Σ_D
- 2 – Obtenha os auto-valores λ_i fazendo $| \Sigma_D - \lambda I | = 0$
- 3 – Calcule os auto-vetores v_i fazendo $[\Sigma_D - \lambda_i I] v_i = 0$
- 4 – Aplique a rotação para obter a matriz transformada $Y = V^T D$

O tratamento da correlação é similar ao da *transformada de Hotelling* [Hotelling, 1933], que consiste no caso discreto da *transformada de Karhunen-Loeve*, considerada o mapeamento ótimo porque minimiza o erro médio quadrático das distâncias entre cada ponto n -dimensional e o seu correspondente no espaço k -dimensional. Embora produza bons resultados de maneira geral, existem situações específicas que não podem ser tratadas pela *PCA*. A mais previsível é aquela em que o eixo principal leva a uma projeção em que grupos bem definidos no espaço original ficam sobrepostos no espaço projetado. Outra limitação é que, pelo fato da *PCA* utilizar a correlação dos dados, a presença de pontos incomuns (*outliers*) introduz problemas indesejáveis. Finalmente, deve ser considerado o seu alto custo computacional, $O(m^2)$.

4.2.2 Projection Pursuit

Assim como a *PCA*, *Projection Pursuit (PP)* [Friedman e Tukey, 1974; Friedman, 1987] também procura encontrar combinações lineares das variáveis multidimensionais.

Entretanto, propõe a substituição do critério adotado na *PCA* (maximizar a variância no espaço de saída) por outros critérios mais adequados a cada situação, dependendo dos objetivos da análise ou das características dos dados. Assim, *PP* pode ser vista como uma generalização da *PCA*.

4.2.3 Decomposição em Valores Singulares

A decomposição em valores singulares (*Singular Value Decomposition* – *SVD*) foi introduzida por Beltrami [1873], e propõe a fatorização de uma matriz $D_{m \times n}$ na forma:

$$D = UAV^T \quad (4.1)$$

onde U é uma matriz ortogonal $m \times m$; V é uma matriz ortogonal $n \times n$; A é uma matriz diagonal $m \times n$ com $\lambda_{i,j} = 0$ se $i \neq j$. As matrizes U e V devem ser obtidas de modo a ordenar os elementos λ_i da diagonal de A ($\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$). Os elementos λ_i são chamados valores singulares de D e correspondem às raízes quadradas dos auto-valores de $D^T D$. De fato, a decomposição em valores singulares também consiste em uma variação de *PCA*, assim como a transformada de *Karhunen-Loève* e de *Hotelling*.

4.2.4 Escalonamento Multidimensional

O Escalonamento Multidimensional (*Multidimensional Scaling* – *MDS*) utiliza informação de distância entre os objetos no espaço original para gerar uma nova representação dos dados no espaço reduzido [Shepard, 1962; Kruskal, 1964]. A abordagem mais comum [Kruskal e Wish, 1978] consiste em mapear cada um dos registros n -dimensionais do conjunto de dados D no espaço k -dimensional, minimizando uma função erro (*função stress*), do tipo:

$$E = \sqrt{\frac{\sum_{i,j} (\epsilon_{i,j} - \delta_{i,j})^2}{\sum_{i,j} \delta_{i,j}^2}} \quad (4.2)$$

onde $\epsilon_{i,j}$ é a distância entre os registros D_i e D_j no espaço n -dimensional de entrada, e $\delta_{i,j}$ é a distância entre D_i e D_j projetados no espaço k -dimensional de saída. A função retorna o erro relativo entre os registros i e j nos espaços n e k . A maior desvantagem desta técnica é a sua complexidade $O(m^2)$, que a torna proibitiva para grandes conjuntos de dados. O *MDS* também pode ser usado no sentido contrário, conforme ilustrado com o exemplo da Figura 3.3, elevando a dimensionalidade dos dados.

4.2.5 Mapeamento não Linear

Proposto por Sammon [1969], o mapeamento não linear (*Nonlinear Mapping – NLM*) busca preservar a distância e a geometria local do espaço original. O mapeamento busca minimizar uma função erro dada por:

$$E = \frac{1}{\sum_{i,j} \epsilon_{i,j}} \sum_{i,j} \frac{(\delta_{i,j} - \epsilon_{i,j})^2}{\epsilon_{i,j}} \quad (4.3)$$

onde $\delta_{i,j}$ é a distância entre os objetos no espaço de saída, enquanto que $\epsilon_{i,j}$ é a distância entre os objetos no espaço de entrada. Assim como o *MDS*, o *NLM* é um método iterativo, que inicialmente atribui valores aleatórios aos objetos no espaço de saída e, em cada passo, calcula as correções (no caso, utilizando gradientes). O algoritmo também possui alta complexidade, $O(k^3 m^2)$.

A Figura 4.1 mostra a visualização de dez registros cujos atributos descrevem as coordenadas de pontos em uma linha reta definida em um espaço de dimensão 9. A visualização mostra que a projeção 2D proposta por Sammon preserva a estrutura original dos dados.

Figura 4.1 – Visualização de um conjunto de dados com dimensão nove utilizando o método proposto por Sammon [1969].

4.2.6 FastMap

De maneira similar à *PCA*, o *FastMap* [Faloutsos e Lin, 1995] também busca maximizar a variância no espaço de saída, entretanto, utiliza algumas heurísticas para evitar a complexidade $O(m^2)$. O *FastMap* projeta (utilizando a lei dos cossenos) os dados sobre uma reta obtida a partir dos pontos mais distantes entre si no conjunto de dados, no lugar das componentes principais. Os resultados são equivalentes em qualidade aos do *MDS*, porém em um tempo muito menor, pois tem complexidade $O(mn)$.

4.2.7 VHDR

Yang et al. [2003] observam que os espaços de menor dimensão gerados por métodos como *PCA*, *NLM* e *MDS* nem sempre transmitem um significado intuitivo dos dados, pois o processamento é automático. Esses autores sugerem uma estratégia chamada Redução da Dimensionalidade Hierárquica Visual (*Visual Hierarchical Dimension Reduction - VHDR*), em que o usuário participa do processo de redução da dimensionalidade, o que contribui para a melhor compreensão dos dados. A proposta se enquadra na categoria de técnicas de seleção de atributos (*Feature Selection*), em que os atributos são classificados, segundo algum critério, segundo níveis de prioridade que determinarão a sua inclusão ou não no conjunto final de atributos, conforme o número de atributos que usuário deseja eliminar.

4.3 Viz3D: Visualização Multidimensional no Espaço 3D

O *Viz3D* [Artero e Oliveira, 2004] generaliza a proposta da técnica *RadViz* para a terceira dimensão, projetando dados n -dimensionais na superfície e no interior de um cilindro 3D, cuja base consiste em um sistema de eixos radiais. Esta solução permite acomodar um número maior de marcadores gráficos, oferecendo uma opção adequada para a exploração de conjuntos de dados volumosos e de alta dimensionalidade, que causam um excessivo congestionamento no *RadViz* e em visualizações 2D em geral. Grande parte da ambigüidade introduzida pelas sobreposições pode ser resolvida por meio de interações do usuário em operações de rotação, translação e ampliações com o modelo tridimensional. É evidente que a visualização/manipulação de marcadores dispostos no espaço tridimensional, porém projetados na tela bidimensional, fica prejudicada. Entretanto, em geral usuários estão bastante acostumados a explorar projeções 2D de modelos 3D em jogos, ou mesmo sistemas de projeto gráfico.

4.3.1 Projeção no Viz3D

Dada a matriz de dados $D_{m \times n}$, o *Viz3D* mapeia as coordenadas n -dimensionais dos m registros de D em coordenadas 3D (x_i, y_i, z_i) segundo a Equação (4.4):

$$Viz3D(d_{i,0}, d_{i,1}, \dots, d_{i,n-1}) = \begin{cases} x_i = x_c + \frac{1}{n} \sum_{j=0}^{n-1} \frac{d_{i,j} - \min_j}{\max_j - \min_j} \cos\left(\frac{2\pi j}{n}\right) \\ y_i = y_c + \frac{1}{n} \sum_{j=0}^{n-1} \frac{d_{i,j} - \min_j}{\max_j - \min_j} \sin\left(\frac{2\pi j}{n}\right) \\ z_i = z_c + \frac{1}{n} \sum_{j=0}^{n-1} \frac{d_{i,j} - \min_j}{\max_j - \min_j} \quad i = 0, \dots, m-1 \text{ e } j = 0, \dots, n-1 \end{cases} \quad (4.4)$$

onde: x_c , y_c e z_c são as coordenadas do centro de um sistema de eixos radiais; $\max_j = \text{Máximo}(d_{k,j})$ e $\min_j = \text{Mínimo}(d_{k,j})$ para $k = 0, \dots, m-1$;

Esrita na forma matricial a Equação (4.4) fica:

$$\begin{bmatrix} x_0 & y_0 & z_0 \\ x_1 & y_1 & z_1 \\ \dots & \dots & \dots \\ x_{m-1} & y_{m-1} & z_{m-1} \end{bmatrix} = \frac{1}{n} \begin{bmatrix} d_{0,0} - \min_0 & \dots & d_{0,n-1} - \min_0 \\ \max_0 - \min_0 & \dots & \max_0 - \min_0 \\ \dots & \dots & \dots \\ d_{m-1,0} - \min_{m-1} & \dots & d_{m-1,n-1} - \min_{m-1} \\ \max_{m-1} - \min_{m-1} & \dots & \max_{m-1} - \min_{m-1} \end{bmatrix} \begin{bmatrix} \cos(0) & \sin(0) & 1 \\ \cos\left(\frac{2\pi}{n}\right) & \sin\left(\frac{2\pi}{n}\right) & 1 \\ \dots & \dots & \dots \\ \cos\left(\frac{2(n-1)\pi}{n}\right) & \sin\left(\frac{2(n-1)\pi}{n}\right) & 1 \\ nx_c & ny_c & nz_c \end{bmatrix} \quad (4.5)$$

A função $Viz3D : \mathbb{R}^n \rightarrow \mathbb{R}^3$, constituída por uma soma de funções contínuas em qualquer ponto $x \in \mathbb{R}^n$, também é contínua, pois quando duas funções f e g são contínuas em um ponto x , então sua soma $f+g$, seu produto $f \cdot g$ e, se $g(x) \neq 0$, seu quociente $\frac{f}{g}$ são funções contínuas em x . Todas estas afirmações seguem diretamente das propriedades dos limites. Por fim, dizer que uma função f é *contínua* em um ponto x significa que $f(x)$ existe e que f leva pontos próximos no domínio (nD) em pontos próximos no contradomínio ($3D$). Esta característica é importante para apoiar um processo de agrupamento visual, usando o $Viz3D$, conforme discutido no Capítulo 6.

A Figura 4.2(a) ilustra a projeção feita no $Viz3D$, sendo que as coordenadas x_i e y_i dos marcadores gráficos são obtidas de forma análoga ao $RadViz$, enquanto as coordenadas z_i são dadas pelas médias dos valores de todos os atributos do i -ésimo registro. Como no $RadViz$, diferentes configurações dos eixos produzem visualizações bem diferentes, conforme ilustram as Figuras 4.2(b) e (c) que mostram projeções $Viz3D$ dos mesmos dados exibidos na visualização $RadViz$ apresentada na Figura 2.12, para diferentes configurações dos eixos radiais.

Figura 4.2 – Projeção 3D no *Viz3D*; b) Mapeamento dos registros r_0 , r_1 , r_2 e r_3 (dimensionalidade quatro, ver Figura 2.12) no *Viz3D*, adotando a sequência de eixos a_0 , a_1 , a_2 e a_3 ; c) Mapeamento com a sequência de eixos a_0 , a_2 , a_1 e a_3 .

O *Viz3D* minimiza problemas de oclusão ao permitir uma fácil interação com a representação 3D por meio de operações de translação, rotação e escala. A técnica Coordenadas Estelares [Kandogam, 2000] (Figura 2.13) permite uma interação com os eixos da visualização, que podem ser re-posicionados interativamente pelo usuário em diferentes configurações, de modo a gerar projeções mais adequadas para a identificação de grupos, por exemplo. Entretanto, quando a dimensionalidade dos dados é muito alta o número elevado de eixos torna impraticável a tarefa de definir interativamente o posicionamento adequado para todos eles. No *Viz3D* os eixos são mantidos fixos uma vez gerada a visualização. A sobreposição dos marcadores é inevitável quando o número de registros é alto, entretanto, a dimensão adicional possibilita resultados melhores do que técnicas 2D. Estratégias como *jittering* e a exibição de informação de densidade obtida dos dados, esta última discutida no Capítulo 5, também ajudam a tratar a sobreposição no *Viz3D*.

4.3.2 Desempenho

Assim como o *RadViz*, o *Viz3D* tem complexidade $O(mn)$, uma vez que a projeção de todos os registros requer uma única varredura dos dados. A Figura 4.3 mostra os tempos de execução para gerar projeções com as técnicas *PCA*, *FastMap* e *Viz3D*, em conjuntos de dados com 11 atributos contendo 10.000, 20.000 e 30.000 registros, respectivamente. Os testes foram realizados em uma máquina com processador Pentium III e 200 MB de memória RAM. Devido à sua baixa complexidade, o *Viz3D* apresenta os menores tempos

de processamento, o que é importante para viabilizar a sua utilização em grandes conjuntos de dados multidimensionais.

Figura 4.3 – a) Tempos de execução (em segundos) para obter projeções de um conjunto de dados de dimensionalidade 11 com *PCA*, *FastMap* e *Viz3D*; b) respectivo gráfico exibindo o tempo de processamento em função do número de registros.

4.4 Resultados

A Figura 4.4 mostra diferentes visualizações do conjunto de dados *hypercube.data*, que contém as coordenadas de um hiper-cubo de dimensão quatro, com as técnicas *RadViz*, *PCA*, *FastMap* e *Viz3D*. Em (a) tem-se a projeção 2D obtida com o *RadViz* e em (b), (c), e (d) têm-se as projeções 3D obtidas com a *PCA*, o *FastMap* e o *Viz3D*, respectivamente. As projeções obtidas tanto o *Viz3D* como com a *PCA* e o *FastMap* revelam bem a estrutura dos dados, o que não ocorre com o *RadViz*.

Figura 4.4 – Projeção de um hiper-cubo (4D) - a) no plano, com o *RadViz*; no espaço 3D com: b) *PCA*, c) *FastMap* e d) *Viz3D*.

A Figura 4.5 mostra projeções do conjunto de dados *Iris.data*, obtidas com o *FastMap*, *PCA*, *NLM*, *RadViz* e *Viz3D*, respectivamente. Embora não tenha sido observada uma boa separação entre duas das três classes existentes em nenhuma das projeções, os melhores resultados são obtidos com o *FastMap*, o *Viz3D* e *PCA*.

Figura 4.5 – a) Projeção do conjunto de dados *Iris.data* usando: a) *Fastmap*, b) *PCA*, c) *NLM*, d) *RadViz* e e) *Viz3D*.

A Figura 4.6 mostra a visualização do conjunto de dados *Pollen.data* no *Viz3D*. Em (a) tem-se uma visualização inicial; em (b), após uma operação de escala, é possível identificar a presença de um agrupamento no interior do volume de dados. Em (c), após operações de escala e rotação obtém-se um posicionamento que permite identificar claramente a estrutura dos dados.

Figura 4.6 – a) Visualização inicial do conjunto de dados *Pollen.data* no *Viz3D*; b) Visualização após ampliação da área central; c) Identificação visual dos grupos.

Embora a projeção 3D acomode um maior número de marcadores gráficos do que a projeção 2D, o congestionamento visual e a oclusão continuam severos quando se trata de conjuntos de dados muito volumosos. A Figura 4.7 ilustra a visualização do conjunto de dados *Out5d.okc*, em que é possível identificar visualmente duas estruturas relevantes. Nesta figura são exibidos também os eixos da visualização, que podem auxiliar a percepção tridimensional ou transmitir uma noção melhor dos valores dos atributos, pois, analogamente ao *RadViz*, no *Viz3D* os marcadores são atraídos pelos eixos que mapeiam atributos com valores mais elevados. Além disso, os marcadores mapeados próximo da base inferior do cilindro são aqueles que possuem valores baixos em todos os seus atributos, enquanto os marcadores próximos do topo são aqueles cujos registros possuem valores elevados em todos os atributos.

Figura 4.7 – Visualização *Viz3D* do conjunto de dados *Out5d.okc*.

A Figura 4.8 ilustra a visualização do conjunto de dados *Synth3.data* no *Viz3D*. Neste caso, mesmo após interagir extensivamente com a representação só é possível

identificar visualmente sete dos oito grupos existentes nos dados. Alterando a sequência de exibição dos eixos pode-se obter uma projeção mais adequada para revelar os oito grupos, entretanto, pode ser cansativo e ineficiente buscar exaustivamente um arranjo mais adequado. No Capítulo 7 é apresentada uma estratégia simples idealizada para obter uma sequência de eixos capaz de separar melhor os grupos em visualizações geradas com o *RadViz*, o *Viz3D* e Coordenadas Paralelas.

Figura 4.8 – Visualização do conjunto *Synth3.data* no *Viz3D*.

4.5 Comentários Finais

O *Viz3D* é uma estratégia bastante simples de projeção de dados multidimensionais em um espaço 3D, que oferece uma alternativa interessante às técnicas convencionais de visualização e de redução de dimensionalidade devido, principalmente, à sua simplicidade e eficácia. As projeções geradas no *Viz3D* têm qualidade semelhante às obtidas com a *PCA* e o *FastMap*, mas demandam menores tempo de processamento, o que as torna muito conveniente para manipular grandes conjuntos de dados multidimensionais.

Em comparação com o *RadViz*, que inspirou a abordagem, a técnica minimiza algumas das principais deficiências, como o congestionamento na área central da visualização, causado pela alta dimensionalidade, e as oclusões que dificultam a análise, decorrentes do número alto de marcadores. Analogamente ao *RadViz*, o *Viz3D* garante que registros, próximos entre si no espaço original, sejam mapeados, próximos entre si, no

espaço de saída. Entretanto, registros distantes entre si no espaço original podem ser mapeados em posições próximas no espaço 3D. De fato, qualquer que seja a estratégia de mapeamento adotada, sempre haverá uma perda de informação decorrente da projeção, sendo impossível preservar completamente a estrutura dos dados no espaço original. Finalmente, as visualizações de vários conjuntos de dados de alta dimensionalidade, geradas com o *Viz3D* revelaram-se bastante eficazes para identificar as estruturas presentes nos mesmos e são usadas no Capítulo 6 para apoiar um processo de agrupamento visual.

Capítulo 5

Visualizações Realçadas com Informação de Frequência e Densidade

5.1 Considerações Iniciais

Buscando tratar o problema do excessivo congestionamento visual na visualização direta de grandes conjuntos de dados, que prejudica seriamente as análises visuais, este capítulo introduz uma abordagem para realçar a expressividade das visualizações mapeando informação de frequência e densidade dos dados. O objetivo da estratégia proposta é revelar ao usuário estruturas e padrões que não seriam identificados em abordagens clássicas de visualização.

Os conceitos de frequência e densidade são apresentados na Seção 5.2 e, em seguida, na Seção 5.3, são introduzidas algumas técnicas da literatura que sugerem embutir informação de densidade em Gráficos de Dispersão e em Coordenadas Paralelas. Finalmente, na Seção 5.4, são apresentados algoritmos para exibir informações de frequência e densidade em Coordenadas Paralelas, que possuem baixa complexidade computacional, permitindo o processamento de grandes conjuntos de dados de alta dimensionalidade em tempos aceitáveis pelo usuário. A Seção 5.5 trata da apresentação da informação de frequência e densidade no *RadViz* e *Viz3D*. Por fim, na Seção 5.6 são apresentados os comentários finais.

5.2 Definições

Seja a matriz de dados $D_{m \times n}$, em que cada linha representa um registro n -dimensional.

Definição 1 (Frequência): A função frequência de $D_{m \times n}$ para uma variável n -dimensional x pode ser definida pela Equação 5.1, em que h é a largura dos intervalos (*bins*) adotados no

cálculo de frequência. Tais intervalos são definidos a partir de uma origem O , como $[O+jh, O+(j+1)h)$ [Silverman,1990], para valores positivos e negativos de j .

$$f(x) = \frac{\sigma}{mh} \quad (5.1)$$

sendo σ o número de registros d_i contidos no mesmo intervalo que contém x .

Esta abordagem requer a definição da largura h dos intervalos, e da origem O , sendo que diferentes escolhas podem produzir resultados muito diferentes entre si.

Definição 2 (Densidade): A função densidade da matriz $D_{m \times m}$ para uma variável n -dimensional x , baseada em um núcleo estimador K , pode ser definida pela Equação 5.2 [Silverman,1990].

$$f(x) = \frac{1}{mh^n} \sum_{i=1}^m K\left(\frac{x-d_i}{h}\right) \quad (5.2)$$

onde d_i é o i -ésimo registro e K é a função núcleo estimador, definida para a variável x , satisfazendo $\int_{R^n} K(x)dx = 1$. Embora vários núcleos tenham sido propostos, os mais comuns são ondas quadradas e Gaussianas, definidas pelas Equações 5.3 e 5.4, respectivamente.

$$K(x) = \begin{cases} \frac{1}{(2\delta)^n} & \forall x; |x_i| < \delta \quad (i=0, \dots, n) \\ 0 & \text{caso contrario} \end{cases} \quad (5.3)$$

e

$$K(x) = \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{x^T x}{2}} \quad (5.4)$$

A Figura 5.1(a) ilustra um conjunto de dados bidimensional, enquanto que (b) apresenta a informação de densidade associada (obtida usando o núcleo estimador onda quadrada) em tons de cinza.

Figure 5.1 – a) Conjunto de dados bidimensional; b) Visualização em que os níveis de cinza apresentam a informação de densidade.

O parâmetro h na função densidade (Equação 5.2) constitui um fator de suavização, ou largura de banda. Quando $h \rightarrow 0$ tem-se o equivalente a uma soma de funções Delta de Dirac [Silverman, 1990], o que é equivalente a utilizar histogramas de freqüência, a forma mais antiga de estimar densidades. Enquanto estimativas de densidade obtidas com valores pequenos de h tendem a realçar apenas as sobreposições dos marcadores nos gráficos, valores grandes de h destacam áreas de concentração dos elementos, i.é., a presença de grupos.

Considerando que a influência de cada ponto pode ser formalmente modelada por uma função matemática chamada função de influência (equivalente a um filtro de suavização controlado por h), que descreve o impacto de um elemento do conjunto de dados sobre a sua vizinhança, a densidade pode então ser modelada analiticamente como a soma de funções de influência de todos os elementos do conjunto. Assim, os grupos podem ser matematicamente ou visualmente identificados pelos pontos de máximo no gráfico da função densidade. O erro introduzido por esta abordagem – no lugar de usar a Equação 5.2, com o núcleo estimador gaussiano dado pela Equação 5.4 foi estimado por Hinneburg e Keim [1998], sendo considerado desprezível com conjuntos de dados reais.

No caso de dados bidimensionais a exibição de gráficos de densidade em um sistema de coordenadas cartesianas é imediata, sendo comum codificar a informação da densidade em cores ou tons de cinza, conforme ilustra a Figura 5.1(b). Entretanto, a representação da freqüência ou densidade de um conjunto de dados n -dimensional gera informação $(n+1)$ -dimensional, o que justifica a necessidade de usar técnicas de visualização multidimensionais para exibir tal informação associada a conjuntos de dados com muitos atributos.

5.3 Trabalhos Relacionados

No sistema *Cviz* [Han e Cercone, 2000], gráficos de densidade e de contornos (*density graphs* e *contour graphs*) são gerados para apoiar a interpretação de Gráficos de Dispersão visualmente sobrecarregados (Figura 5.2). O matiz codifica a densidade de pontos na janela de exibição.

Figura 5.2 – a) *Density graph*, e b) *Contour graph*. Em ambos os casos, a densidade de pontos é indicada pelo matiz do pixel (tons laranja codificam valores maiores e tons púrpura codificam valores menores). No *contour plot* a cor de uma região também é influenciada pela cor das regiões vizinhas [Han e Cercone, 2000].

Conforme discutido no Capítulo 2 (Figura 2.13(a)), visualizações por Coordenadas Paralelas também apresentam congestionamento quando são exibidos muitos registros, entretanto, a exibição da informação de frequência ou densidade em Coordenadas Paralelas não é tão imediata. A Figura 5.3 ilustra uma visualização por Coordenadas Paralelas, do conjunto de dados *Synth1.data*, que possui apenas 7.500 registros com cinco atributos, completamente prejudicada pela intensa sobreposição dos marcadores.

Figura 5.3 – a) Visualização do conjunto de dados *Synth1.data* (7.500 registros de cinco atributos).

Na figura 5.3 as sobreposições podem ser decorrentes de três situações distintas. Registros duplicados (ou distintos, porém muito semelhantes) causam a sobreposição total das poligonais, pois são mapeados na mesma poligonal, por causa da resolução da usada, conforme ilustra a Figura 5.4(a). Existe ainda a sobreposição decorrente dos cruzamentos das linhas, conforme ilustra a Figura 5.4(b). A terceira situação é uma combinação das

anteriores, ou seja, sobreposição parcial e devido a cruzamentos, conforme mostra a Figura 5.4(c).

Figura 5.4 – a) Múltiplos registros, idênticos ou não, mapeados no mesmo marcador, causam a sobreposição das poligonais; b) Registros distintos ocasionam cruzamentos das poligonais; c) Registros com alguns atributos consecutivos iguais causam cruzamentos e sobreposição parcial das poligonais.

Diversos autores têm investigado mecanismos para realçar informação relevante em visualizações por Coordenadas Paralelas sobrecarregadas. Miller e Wegman [1990] e Wegman [1990] sugerem o uso de Histogramas Médios Deslocados (*Averaged Shifted Histograms – ASH*) [Scott, 1985] para obter a densidade e apresentá-la em Coordenadas Paralelas. Os *ASHs* minimizam problemas introduzidos no cálculo da frequência, computada em intervalos (*bins*) que precisam ser definidos pelo usuário. Eles também sugerem um esquema para evitar a contagem múltipla de poligonais no cálculo da frequência. Wegman e Luo [1996] também utilizam densidade para realçar estruturas dos dados visualizados em Coordenadas Paralelas. Essa abordagem, bem mais simples que a anterior, sugere exibir os pixels das linhas poligonais com um nível de transparência proporcional ao nível de sobreposição no local. A Figura 5.5(a) mostra a visualização do conjunto de dados *Pollen.data* em Coordenadas Paralelas convencional e, em (b) tem-se a visualização baseada em densidades resultante da proposta de Wegman e Luo [1996].

Figura 5.5 – a) Visualização do conjunto de dados *Pollen.data* em Coordenadas Paralelas convencional; b) Visualização da informação de densidade dos mesmos dados, com a intensidade do pixel proporcional à sobreposição no local [Wegman e Luo, 1996].

Esta técnica pode ser implementada eficientemente usando o canal de transparência das placas gráficas mais modernas e, apesar de sua simplicidade, consegue realçar a presença de padrões nos dados. Entretanto, uma intensa sobreposição decorrente dos cruzamentos das muitas poligonais, como ilustrado na Figura 5.4(b), causa ainda um indesejável congestionamento.

Rodrigues Junior et al. [2003] também utilizam informação de frequência para realçar regiões de interesse em visualizações por Coordenadas Paralelas. A técnica é chamada *Frequency Plots*, e calcula a frequência no espaço unidimensional definido por cada um dos atributos. As poligonais são traçadas da seguinte forma: a intensidade em cada segmento de linha varia de maneira a interpolar linearmente os valores de frequência computados para os atributos associados nas extremidades (sobre os eixos). A Figura 5.6(a) mostra a visualização do conjunto de dados *Cancer.data* em Coordenadas Paralelas convencional, enquanto a Figura 5.6(b) ilustra o *Frequency Plot*. Os registros pertencem a uma de duas classes conhecidas (*Classe 1* e *Classe 2*), cujo atributo determinante foi mapeado no eixo mais à direita. Em (c) e (d) o usuário selecionou apenas os registros pertencentes às classes 1 e 2, respectivamente. Os *Frequency Plots* correspondentes conseguem, de certo modo, realçar o comportamento de cada classe. As visualizações foram geradas no sistema *GBDView* (em <http://www.gbdi.icmc.usp.br/download>).

Figura 5.6 – a) Visualização do conjunto de dados *Cancer.data* em Coordenadas Paralelas; b) *Frequency Plot*; c) *Frequency Plot* dos registros da Classe 1; d) *Frequency Plot* dos registros da Classe 2 [Rodrigues Junior et al., 2003].

Conforme revelam as Figuras 5.6(c) e (d), ao selecionar os registros pertencentes a uma mesma classe, um analista tem uma visão de como esta se comporta em relação a seus atributos. Entretanto, o *Frequency Plot* é menos eficiente na identificação de padrões quando aplicado ao conjunto de dados inteiro, como mostram as Figura 5.6(b) e 5.7. Esta última exibe o *Frequency Plot* gerado para o conjunto de dados *Pollen.data*.

Figura 5.7 – *Frequency Plot* do conjunto de dados *Pollen.data*.

Uma outra técnica apresentada por Rodrigues Junior et al. [2003] é chamada *Relevance Plot*, e exige que o usuário defina pontos de relevância sobre um atributo de interesse sendo analisado. As Figuras 5.8(a) e (b) apresentam o *Frequency* e o *Relevance*

Plots, respectivamente, do conjunto de dados *Synth1.data*, em que é possível identificar alguns padrões no interior dos dados.

Figura 5.8 – Visualizações do conjunto de dados *Synth1.data* a) *Frequency Plot* e b) *Relevance Plot*.

Na Seção 5.4 é introduzida uma nova proposta para exibir informação de frequência e densidade em Coordenadas Paralelas, a qual é de implementação mais simples do que a solução de Miller e Wegman [1990] e apresenta resultados mais satisfatórios do que os obtidos com a solução de Wegman e Luo [1996]. Além disso, oferece mais interatividade, apoiando processos exploratórios de análise visual. Os algoritmos geram uma representação matricial do gráfico de Coordenadas Paralelas que é adequada para pós-processamento por técnicas de processamento digital de imagens, tais como manipulação da escala de cinza, com a finalidade de revelar estruturas ocultas; limiarização, para eliminar ou enfatizar registros com baixa ou alta incidência, caso seja necessário identificar grupos ou pontos incomuns, respectivamente; e inversão dos valores de brilho, que pode tornar a visualização perceptualmente mais adequada à análise visual. As estratégias propostas são denominadas *Interactive Parallel Coordinates Frequency Plot* e *Interactive Parallel Coordinates Density Plot* [Artero et al., 2004].

5.4 IPC Frequency e Density Plots

Os algoritmos propostos computam um histograma de frequência bidimensional para cada par de atributos a serem mapeados em eixos consecutivos de Coordenadas Paralelas. Os histogramas provêm informações de frequência e densidade que controlam a apresentação das poligonais. Os parâmetros h e O (largura e origem) que definem os intervalos na construção dos histogramas são determinados pela resolução, em pixels, dos eixos do gráfico de Coordenadas Paralelas (cada pixel corresponde a um intervalo, ou *bin*,

no cálculo do histograma de frequências). O mapeamento é análogo ao adotado em visualizações por Coordenadas Paralelas convencionais, em que a faixa de valores de cada atributo é mapeada sobre os eixos associados na tela.

A estratégia é ilustrada por meio de sua aplicação a um pequeno conjunto de dados [Witten e Frank, 2000] bastante usado para ilustrar algoritmos de classificação em mineração de dados. Os dados consistem em quatorze registros com cinco atributos, sendo que apenas dois atributos são usados no exemplo. Os dados são representados em uma matriz $D_{14 \times 2}$ e o intervalo de valores dos atributos é mapeado para o intervalo $[1, L]$, sendo L a resolução dos eixos. Este mapeamento é ilustrado nas Tabelas 5.1 e 5.2, que mostram, respectivamente, os valores originais dos atributos *outlook* (categórico) e *temperature* (numérico) e o seu mapeamento para o intervalo $[1, 5]$ ($L=5$).

Tabela 5.1 - Valores originais

Outlook	temperature
sunny	85
sunny	80
overcast	83
rainy	70
rainy	68
rainy	65
overcast	64
sunny	72
sunny	69
rainy	75
sunny	75
overcast	72
overcast	81
rainy	71

Tabela 5.2 - Valores mapeados no intervalo $[1, L]$

Outlook	Temperature
5	5
5	4
1	5
3	2
3	2
3	1
1	1
5	3
5	2
3	3
5	3
1	3
1	4
3	2

Os três valores do atributo categórico *outlook* foram mapeados no intervalo $[1, 5]$ considerando a ordem alfabética dos valores, enquanto os valores numéricos do atributo *temperature*, originalmente no intervalo $[64, 85]$, foram linearmente mapeados no intervalo $[1, 5]$.

A Figura 5.9(a) mostra o Gráfico de Dispersão dos dados da Tabela 5.2, tendo sido introduzido um *jittering* para evitar a sobreposição dos marcadores. A Figura 5.9(b) mostra a visualização dos mesmos registros em Coordenadas Paralelas convencionais. A matriz bidimensional $F_{L \times L}$ exibida na Figura 5.9(c) mostra a frequência computada do par de atributos. A Figura 5.9(d) mostra a visualização por Coordenadas Paralelas, denominada *IPC Frequency Plot*. Nela, cada elemento $f_{i,j}$ não nulo de F é usado para traçar um segmento de reta (i, j) na visualização. A intensidade de todos os pixels de cada segmento

de reta (i,j) é definida pelo valor armazenado em $f_{i,j}$. Pixels associados a intensidades menores são sobrepostos por pixels com intensidades maiores. O procedimento pode ser generalizado para o caso n -dimensional, sendo necessário calcular $n-1$ matrizes de frequências (uma para cada par de atributos mapeados em eixos consecutivos).

Figura 5.9 – a) Gráfico de dispersão dos dados exibidos na Tabela 5.2; b) Visualização por Coordenadas Paralelas dos mesmos dados; c) Matriz de frequências F deste par de atributos; d) Visualização por Coordenadas Paralelas em que a intensidade das poligonais é proporcional ao valores de frequência em F .

O algoritmo que implementa este procedimento é apresentado na Caixa 5.1. O mapeamento descrito no Passo 2 é equivalente ao realizado para gerar uma visualização convencional por Coordenadas Paralelas, mapeando os valores de cada atributo nos valores discretos do sistema de coordenadas da tela. Assim, como discutido, a origem e o número de intervalos usados no cálculo da frequência F é determinado pela resolução da visualização.

No final do processo, os elementos $g_{i,j}$ da matriz G criada no Passo 4.1 determinam os valores de intensidade em cada pixel i,j do gráfico de Coordenadas Paralelas (G consiste em uma representação matricial do gráfico de Coordenadas Paralelas). Entretanto, considerando que os elementos $g_{i,j}$ contém valores em uma faixa que vai de zero até números muito altos (de acordo com a frequência obtida na posição correspondente), torna-se necessário definir uma função de mapeamento que viabilize a sua apresentação em tons de cinza, normalmente na faixa $[0,255]$. Uma forma de mapear os valores é apresentada na Equação 5.5, que determina a intensidade I de um pixel p,q fazendo:

$$I_{p,q} = \frac{255g_{p,q}}{\text{Max}(G)} \quad (5.5)$$

onde $\text{Max}(G)$ é o maior valor da matriz G .

```

1 – Inicialize a matriz  $G_{L \times W}$  com zeros. As dimensões  $L$  e  $W$  são definidas pela resolução em pixels do gráfico a ser gerado. No caso de um gráfico com eixos verticais,  $L$  e  $W$  são determinados, respectivamente, pelas resoluções vertical e horizontal.

Faça  $g_{p,q} = 0$  para  $p=1, \dots, L$  e  $q=1, \dots, W$ 

2 – Dado o conjunto de dados original armazenado em uma matriz  $D_{m \times n}$ , onde  $m$  é o número de registros e  $n$  é a dimensionalidade dos dados (o número de atributos dos registros), construa uma matriz auxiliar  $A_{m \times n}$  fazendo:


$$a_{i,j} = \frac{L(d_{i,j} - \min_j)}{\max_j - \min_j}$$
 para  $i=1, \dots, m$  e  $j=1, \dots, n$ 

onde:  $\max_j = \text{Max}(d_{k,j})$  para  $k=1, \dots, m$ , e
 $\min_j = \text{Min}(d_{k,j})$  para  $k=1, \dots, m$ ;

/* matriz A armazena os valores de D, mapeados linearmente no intervalo [0,L] */

3 – Calcule uma matriz  $F_{L \times L}$  para cada par consecutivo de atributos

Para  $i = \{2, \dots, n\}$ 
Faça  $f_{j,k} = 0$  para  $j=1, \dots, L$  e  $k=1, \dots, L$ ; /* Inicialize F com zeros */
3.1 Para  $r = \{1, \dots, m\}$ 
faça  $b = a_{r,i-1}$  e  $c = a_{r,i}$ 
Incremente o valor do elemento  $f_{b,c}$ ;

4 – Para cada  $f_{j,k} \neq 0$ , com  $j=1, \dots, L$  e  $k=1, \dots, L$  calcule: /* Construa o Gráfico */


$$u_x = j \quad u_y = (j-1) \frac{W}{n};$$


$$v_x = k \quad v_y = i \frac{W}{n};$$


4.1 Use o algoritmo de Bresenham (traçado de linhas) para obter as coordenadas  $s, t$  dos pixels do segmento de reta unindo as posições  $u:(u_x, u_y)$  e  $v:(v_x, v_y)$ ; Para cada posição  $(s, t)$  calculada, defina a correspondente intensidade do pixel na matriz  $G$  fazendo: se  $g_{s,t} < f_{j,k}$  faça  $g_{s,t} = f_{j,k}$ ;

/* o brilho de todos os pixels de cada segmento de reta é proporcional ao correspondente valor armazenado na matriz  $F$  */

5 – Exiba o IPC Frequency Plot armazenado na matriz  $G$  usando a Equação 5.5.

```

Caixa 5.1 – Algoritmo para a construção do *IPC Frequency Plot*.

A informação de densidade é obtida a partir da informação de frequência armazenada em F , através da aplicação de um filtro de suavização, conforme ilustra a Figura 5.10. O filtro utilizado (núcleo estimador onda quadrada) é exibido na Figura 5.10(a), enquanto que (b) mostra a matriz de densidades obtida com a sua aplicação sobre a matriz de frequências da Figura 5.9(c). O *IPC Density Plot* é exibido na Figura 5.10(c). Embora a aplicação do filtro de suavização introduza algumas poligonais que não pertencem de fato ao conjunto de dados original, conforme mostra a Figura 5.10(c), a visualização se revela mais eficiente para realçar grupos, caso existam. Entretanto, caso

seja necessário identificar individualmente elementos do conjunto de dados, é mais conveniente usar o *IPC Frequency Plot*. O algoritmo da Caixa 5.1 pode ser facilmente adaptado para gerar o *IPC Density Plot*, bastando incluir a aplicação do filtro de suavização sobre a matriz de frequências F antes de executar o Passo 4.

Figura 5.10 – a) Filtro de suavização onda quadrada; b) matriz de densidades obtida aplicando o filtro sobre a matriz de frequência F (Figura 5.9 (c)); c) *IPC Density Plot* resultante.

5.4.1 Desempenho dos *IPC Frequency* e *Density Plots*

O algoritmo proposto tem complexidade de tempo $O(mn)$, onde m é o número de registros e n é a dimensionalidade. A complexidade é determinada pelos passos necessários para calcular as matrizes de frequência, uma para cada par de atributos no gráfico (Passo 3 do algoritmo), cujo cálculo requer uma passada completa sobre os dados (Passo 3.1). A Figura 5.10(a) mostra os tempos de execução (em segundos) do algoritmo para conjuntos de dados de vários tamanhos (m e n). A máquina utilizada nos testes possui um processador *AMD Athlontm* de 1.8 GHz e 2 Gigabytes de memória RAM, permitindo que mesmo o maior dos conjuntos fosse totalmente alocado em memória. A quantidade de memória necessária para executar o algoritmo é determinada pela locação das matrizes F e G . A matriz F , cujo número de linhas e de colunas é definido pela resolução dos eixos da visualização, é alocada uma única vez, e reutilizada no cálculo da frequência de todos os pares de atributos. A matriz G tem as dimensões definidas pelas resoluções horizontal e vertical da visualização a ser gerada. Os tempos de execução são satisfatórios, mesmo para o maior conjunto, que contém 1.000.000 de registros com 200 atributos cada. Os tempos obtidos também podem ser analisados na Figura 5.11(b), que mostra um gráfico do tempo em função do tamanho do conjunto de dados, em escala logarítmica.

Figura 5.11 – a) Tempos de execução (em segundos) do algoritmo proposto, aplicado em conjuntos de dados com diferentes valores de m (número de registros) e n (número de atributos); e b) correspondente gráfico de tempo, escala logarítmica.

5.4.2 Interação com os *IPC Frequency e Density Plots*

Um usuário pode interagir com os *IPC Frequency e Density Plot* de diversas maneiras. Pode por exemplo, alterar o mapeamento para a escala de cinza dos valores armazenados na matriz G , alterando o parâmetro s na Equação 5.6. O fator de escala s nessa função de mapeamento permite controlar a apresentação de modo a realçar marcadores de menor frequência/densidade ou reduzir a intensidade de marcadores associados a valores de frequência/densidade muito elevados.

$$I_{p,q} = \text{Máximo} \left\{ 255, \frac{255g_{p,q}}{\text{Max}(G)} s \right\} \quad (5.6)$$

O fator de escala s é um número real positivo, controlado interativamente pelo usuário e $\text{Max}(G)$ é o maior valor em G . A Figura 5.12(a) mostra a visualização do conjunto de dados *Synth1.data*, fazendo $s = 1$ em todas regiões da visualização. Em (b) é aplicado um fator de escala $s = 0.5$ entre o segundo e terceiro eixos, e $s = 2.0$ entre o terceiro e o quarto. Para as demais áreas, foi adotado $s = 1.0$.

Figura 5.12 – a) Visualização do conjunto de dados *Synth1.data* com $s=1.0$; b) com diferentes valores de s .

A Figura 5.13 mostra visualizações do conjunto de dados *Pollen.data*, gerados com os algoritmos propostos. Como antecipado, a visualização baseada na densidade (*IPC Density Plot*), ilustrada na Figura 5.13(b), é mais eficaz para realçar o grupo na área central do que o *IPC Frequency Plot*, ilustrado na Figura 5.13(a). Entretanto, as visualizações não apresentam o excessivo congestionamento visual decorrente do cruzamento dos marcadores obtidos com o método de Wegman e Luo [1996] (Figura 5.5(b)). De fato, após observar a Figura 5.13 fica mais fácil identificar a estrutura que Wegman e Luo afirmam existir na Figura 5.5(b).

Figura 5.13 – Visualização do conjunto de dados *Pollen.data* usando a) *IPC Frequency Plot*; b) *IPC Density Plot*.

As Figuras 5.14(a) e (b) mostram, respectivamente, o *IPC Frequency Plot* e o *IPC Density Plot*, que revelam os quatro grupos relevantes do conjunto de dados *Synth1.data* (exibido na Figura 5.3(a) com Coordenadas Paralelas convencional). Os 2.920 registros que realmente pertencem aos cinco grupos contidos no conjunto são exibidos em (c).

Figura 5.14 – a) *IPC Frequency Plot* do conjunto de dados *Synth1.data* (7.500 registros com cinco atributos); b) *IPC Density Plot*; c) visualização dos 2.920 registros efetivamente pertencentes aos cinco grupos do conjunto de dados.

Duas operações de limiarização (*thresholding*) podem ser aplicadas sobre a matriz G para remover da visualização as poligonais cujos registros apresentam frequência inferior a um valor limiar T . A eliminação dos marcadores associados a registros de baixa frequência tem por objetivo reduzir o congestionamento visual. Nas duas operações o usuário define um valor T (um número inteiro), sendo que na primeira abordagem, chamada *Limiarização AND (AND Thresholding)*, apenas os registros com frequência superior ao limiar T em todas as matrizes F são exibidos. Na segunda abordagem, *Limiarização OR (OR Thresholding)*, os registros exibidos são aqueles que possuem frequência superior ao limiar T em pelo menos uma das matrizes F . Ambas as operações oferecem uma maneira bastante simples, e totalmente automática, de eliminar marcadores associados a valores baixos de frequência/densidade. As operações também podem ser modificadas, de modo que os registros com frequência ou densidade superior a um certo limiar T sejam eliminados, permanecendo apenas os registros incomuns (*outliers*). Também é possível interagir com os *IPC Frequency* e *Density Plots*, para determinar e extrair agrupamentos presentes nos dados. Esse processo iterativo de agrupamento visual é descrito no Capítulo 6.

5.5 Frequência e Densidade no *Radviz* e *Viz3D*

A mesma abordagem de embutir informação de frequência ou densidade pode ser aplicada a outras técnicas de visualização, como o *RadViz* e o *Viz3D*, para tratar o problema do congestionamento visual. A informação de densidade poderia ser calculada no espaço original dos dados (n -dimensional), entretanto, isso exigiria uma estrutura de dados muito complexa, com a necessidade de alocar uma matriz de alta dimensionalidade muito esparsa, o que poderia inviabilizar o processamento. Considerando que as projeções realizadas pelo *RadViz* e *Viz3D* conseguem preservar satisfatoriamente a estrutura dos dados multidimensionais no espaço de exibição 2D ou 3D, uma alternativa conveniente consiste em calcular a densidade no espaço de saída (2D ou 3D).

A Figura 5.15 exibe uma visualização *RadViz* do conjunto de dados *Synth4.data*. Em (a) todos os marcadores são desenhados com a mesma cor, ou tom de cinza. Em (b), as intensidades dos marcadores é proporcional à frequência, ou sobreposição, no local.

Figura 5.15 – Visualização do conjunto de dados *Synth4.data* com o *RadViz*: a) marcadores traçados com intensidade constante, e b) marcadores traçados com intensidade proporcional à frequência.

A projeção 2D do *RadViz* tipicamente resulta em intensa sobreposição dos marcadores. No *Viz3D* a sobreposição é reduzida, mas ainda assim pode ser severa. Assim, a apresentação de informação de densidade, obtida com o *Viz3D*, chamada *Viz3D Density Plot* [Artero e Oliveira, 2004], minimiza o problema e permite observar padrões e estruturas existentes nos dados. A Caixa 5.2 mostra o algoritmo para calcular a densidade no *Viz3D*, sendo utilizada uma matriz de densidades 3D cujas dimensões coincidem com a resolução do modelo 3D (volume) a ser gerado.

```

1 - Seja o conjunto de dados na forma matricial  $D_{m \times n}$ , onde  $m$  é o número de registros e  $n$  é a
dimensão (número de atributos) dos registros;
2 - Construa uma matriz tridimensional  $densidade[w][w][w]$ ;
//  $w$  é o limite em pixels do gráfico (3D) a ser construído ( $w \times w \times w$ )
3 - Faça  $densidade_{p,q,r} = 0 \ \forall \ p,q,r / \ p \in [0,w], \ q \in [0,w] \text{ e } r \in [0,w]$ ;
4 - Para  $i = \{1,2,\dots,m-1\}$  calcule  $x, y$  e  $z$  usando a Equação 4 // calcula a frequência dos
4.1 incremente o valor de  $f[x,y,z]$ ; // pontos no espaço 3D
5 - Para cada  $x,y,z = \{0,1,\dots,w\}$  faça // calcula a densidade usando o núcleo
Para cada  $p,q,r = \{-K_h,\dots,K_h\}$  faça  $densidade[x,y,z] = f[x,y,z] + f[x+p,y+q,z+r]$ ;
6 - Exiba a matriz tridimensional  $densidade$  como voxels atribuindo o valor da matriz ao canal
 $\alpha$  (transparência) da cor ou ao valor de brilho do pixel.

```

Caixa 5.2 – Algoritmo para a construção de gráficos de densidades no *Viz3D*.

Como no *Viz3D*, as coordenadas 3D x,y,z projetadas de cada registro são computadas. As densidades são então calculadas e armazenadas nas posições associadas às coordenadas na matriz de densidades. K_h é dado por $(K_w - 1)/2$, onde K_w é a largura do

núcleo estimador utilizado no cálculo da densidade. A Figura 5.16 mostra o núcleo onda quadrada tridimensional de largura 5.

Figura 5.16 – Núcleo onda quadrada tridimensional ($K_w = 5$).

A visualização é gerada a partir da matriz de densidades, sendo que cada elemento $densidade(i,j,k)$ é projetado em um *voxel* de coordenadas i,j,k , fazendo o valor associado ao *voxel* proporcional ao valor armazenado na posição correspondente da matriz. Projetando o volume de voxels dessa forma, e assumindo uma visualização monocromática em tons de cinza, regiões com maior densidade de dados serão pintadas com maiores intensidades, permitindo a fácil identificação visual dos grupos (regiões com maior brilho). Novamente, uma inversão do mapeamento de densidades em intensidades pode auxiliar a tarefa de identificar registros incomuns.

Em um processo exploratório, o usuário pode controlar dois parâmetros (Figura 5.17): a largura do núcleo estimador e o limiar mínimo de densidade necessário para que o *voxel* seja apresentado, o que permite controlar o nível de ruído na visualização sendo exibida. A variação da largura do filtro permite controlar a formação dos grupos (hierarquia), pois com um valor mínimo, cada *voxel* constituirá um grupo, e com um valor máximo, todos os *voxels* ficarão conectados formando um único grupo. A Figura 5.17 mostra o *Viz3D Density Plot* do conjunto de dados *Synth3.data*, gerado com o algoritmo apresentado na Caixa 5.2. A sequência de eixos adotada para a gerar a visualização foi determinada com a abordagem apresentada no Capítulo 7. Em (a) foi selecionado um limiar de densidade pequeno (igual a 4), e em (b), (c) e (d) ele é aumentado gradualmente (para 6, 10, e 16), reduzindo o nível de ruído N_l de modo a exibir apenas as regiões de alta densidade, i.e., os grupos. Nas Figuras 5.17(e), (f) e (g) varia-se a largura do núcleo estimador K_w (igual a 11, 7 e 3), sendo possível observar que o filtro mais largo agrega mais marcadores nos grupos.

Figura 5.17 – *Viz3D Density Plot* do conjunto de dados *Synth3.data*.

A visualização na Figura 5.18(a) mostra o conjunto de dados *Out5d.okc* no *Viz3D*, enquanto que em (b) tem-se o *Viz3D Density Plot*, que confirma as estruturas existentes nos dados.

Figura 5.18 – Visualização do conjunto de dados *Out5d.okc* a) no *Viz3D* e b) no *Viz3D Density Plot*.

5.6 Comentários Finais

Técnicas de exploração visual de dados são importantes porque inserem o potencial humano de perceber padrões no processo de investigação. Entretanto, técnicas de visualização diretas são inadequadas para tratar grandes conjuntos de dados de alta

dimensionalidade. Embora a exibição da informação de frequência e densidade para dados bidimensionais, usando gráficos de dispersão (Figura 5.2) seja trivial, o mesmo não ocorre, para a exibição de dados de alta dimensionalidade. Assim, abordagens baseadas em informação de frequência e densidade, como o *Viz3D Density Plot*, *IPC Frequency* e *Density Plot*, que operam com dados de alta dimensionalidade, são importantes porque conseguem realçar as estruturas existentes nos dados.

Os algoritmos propostos apresentam baixa complexidade computacional, o que viabiliza a sua aplicação a grandes conjuntos de dados multidimensionais. O cálculo da informação de frequência e densidade usando a matriz de frequências (bidimensional) nos *IPC Frequency* e *Density Plots* e o cálculo da informação de frequência e densidade no espaço do *Viz3D* (tridimensional), para obter o *Viz3D Density Plot*, simplifica consideravelmente o processamento, pois o cálculo destas informações em espaços de alta dimensionalidade se torna impraticável.

O cálculo da informação de frequência e densidade em espaços discretos, determinados pelo usuário (resolução dos gráficos) não compromete as visualizações, porque considera a sobreposição dos marcadores para a respectiva visualização. Mecanismos de interação, como por exemplo, o *zoom* usado no *Viz3D* e *Viz3D Density Plot* são importantes para auxiliar o processo de exploração dos dados, e devem ser inseridos no *IPC Frequency* e *Density Plots*, embora a proposta destas técnicas estabeleça que a seleção dos grupos existentes seja satisfatória, mesmo sem a necessidade de grande precisão nas seleções em cada eixo, pois a determinação de intervalos de interesse (ainda com baixa precisão) em vários eixos deve ser capaz de caracterizar os grupos.

As técnicas apresentadas nesse capítulo são usadas para apoiar um processo de agrupamento visual no Capítulo 6, sendo que a informação de densidade é utilizada para realçar as regiões com os agrupamentos relevantes, além de reduzir o efeito do ruído presente nos dados.

Capítulo 6

Exploração Visual de Dados usando as Técnicas Propostas

6.1 Comentários Iniciais

A identificação de agrupamentos de itens semelhantes pode revelar importantes informações sobre os dados. Algoritmos de agrupamento, ou classificação não supervisionada, tem utilidade justificada na análise de grandes conjuntos de dados quando o usuário não possui informações sobre as classes, esperando que o processo de agrupamento possa revelá-las. Algoritmos automáticos de agrupamento sempre apresentam limitações quando aplicados a dados de alta dimensionalidade, não conseguindo produzir resultados satisfatórios em muitos casos. Para estas situações, várias técnicas de agrupamento visual foram propostas, em geral solicitando a intervenção do usuário para definir os grupos a partir de projeções dos dados em espaços de menor dimensionalidade.

Diversas propostas para agrupamento visual foram discutidas no Capítulo 3. Algoritmos como o *Eureka!* [Manco et al., 2004], o *HD-Eye* [Hinneburg et al., 1999], *HC-Cooperative* [Aggarwal, 2001] e *HC-Enhanced* [Gamero e Minghim, 2003] projetam os dados multidimensionais em espaços bi ou tridimensionais, nos quais o usuário delimita visualmente regiões para definir os grupos no espaço multidimensional. De maneira similar, a abordagem de mineração visual introduzida neste capítulo consiste em apresentar os dados no espaço 3D, no caso do *Viz3D*, ou no espaço 2D no caso dos *IPC Frequency* e *Density Plots* e, em seguida, permitir que o usuário interaja com as visualizações para determinar visualmente regiões que delimitam grupos de interesse. O *Viz3D* tem uma vantagem sobre a proposta de Manco et al., [2004], pois permite ao usuário rotacionar os marcadores no espaço 3D, resolvendo problemas de oclusão não tratadas na projeção no espaço 2D realizada no *Eureka!*. Definida uma região de interesse que delimita os marcadores identificados visualmente como pertencentes um grupo, apenas os registros

nela contidos são seleccionados, e o usuário pode refinar a seleção do grupo. Essa abordagem evita que registros sejam incluídos indevidamente no grupo porque atenua o congestionamento visual, uma vez que o usuário interage com um sub-conjunto dos dados.

No *Viz3D* o usuário interage com a visualização até obter uma configuração que permita seleccionar (*brushing*) na tela a região, como ilustrado na Figura 2.26(b), que contém um grupo de interesse. Nos *IPC Frequency* e *Density Plots* a seleção é realizada demarcando intervalos de interesse sobre cada um dos eixos da visualização -- como ilustrado na Figura 2.26(d) -- de maneira a filtrar os elementos até isolar os registros que pertencem a um grupo. Tanto no *Viz3D Density Plot* como nos *IPC Frequency* e *Density Plots* a apresentação de informação de frequência ou densidade realça áreas de alta concentração de elementos e, ao mesmo tempo, reduz o ruído, facilitando a tarefa de identificar os grupos. Na Seção 6.2 são apresentadas as estratégias usadas para realizar um agrupamento visual usando o *Viz3D* e *Viz3D Density Plot*. Enquanto a execução da mesma tarefa usando os *IPC Frequency* e *Density Plots* é descrita na Seção 6.4. Na Seção 6.4 é tratada a exploração dos dados usando as técnicas se apoiando mutuamente, como alternativa para contornar suas deficiências. Na Seção 6.5 é apresentada uma análise dos grupos obtidos com as técnicas propostas. Finalmente, na Seção 6.6 são apresentados os comentários finais do capítulo.

6.2 Agrupamento Visual no *Viz3D* e *Viz3D Density Plot*

Nesta seção é ilustrado como as técnicas *Viz3D* e *Viz3D Density Plot* podem apoiar um procedimento iterativo de identificação de agrupamentos de dados n -dimensionais. Também é discutido como interagir com as visualizações para identificar pontos incomuns (*outliers*) nos dados.

6.2.1 Agrupamento apoiado pelo *Viz3D*

A Figura 6.1 ilustra o processo de identificação de um dos grupos contidos no conjunto de dados *Synth3.data*. Em (a) tem-se uma visualização inicial no *Viz3D*, na qual o usuário delimita por um retângulo a região que contém o grupo desejado. Nesta operação todos os registros contidos no retângulo, na projeção atual, são seleccionados. Em seguida, apenas os registros seleccionados são exibidos (Figura 6.1(b)). O usuário interage com a visualização, aplicando rotações, e identifica marcadores que não pertencem de fato ao grupo. Estes foram indevidamente seleccionados devido à projeção exibida na Figura 6.1(a),

na qual os marcadores indevidos estão na frente ou atrás dos marcadores do grupo. Este cenário é ilustrado na Figura 6.1(c). Novamente o usuário delimita a região de interesse sobre o sub-conjunto dos dados previamente selecionado. Em (d), tem-se o resultado desta operação e, após interagir novamente para verificar a presença de marcadores indesejados, um novo retângulo determina os marcadores que devem permanecer no grupo. Em (e), após interagir novamente com os marcadores remanescentes, o usuário verifica que o grupo está adequadamente delimitado. Uma operação finaliza o processo, de tal forma que todos os registros associados aos marcadores selecionados recebem um rótulo identificador do grupo correspondente, por exemplo 1, supondo que este é o primeiro grupo sendo identificado nesses dados. Na etapa seguinte à identificação do Grupo 1, ilustrada na Figura 6.1(f), todos os registros ainda não designados a um grupo qualquer são exibidos na visualização. Os marcadores correspondentes aos registros já alocados no Grupo 1 não são mais exibidos, e o processo pode continuar com a identificação dos demais grupos.

Figura 6.1 – Processo de agrupamento visual usando o Vi3D.

Como mencionado na Seção 4.3, o *Viz3D* mapeia registros próximos entre si, no espaço original, em posições próximas no espaço tridimensional de saída, o que assegura que os grupos observados na tela contenham, de fato, todos os registros pertencentes a eles no espaço de entrada. Por outro lado, registros muito distantes entre si no espaço de entrada podem ser mapeados, indevidamente, em posições muito próximas no espaço 3D. Consequentemente, a abordagem de agrupamento proposta pode gerar grupos de qualidade ruim, incluindo registros estranhos. Dessa forma, é necessária uma avaliação complementar da qualidade dos grupos gerados. Uma alternativa simples para identificar e eliminar registros indevidos é utilizar, como apoio, visualizações por Coordenadas Paralelas dos elementos do grupo. Tais visualizações podem apoiar um processo de refinamento do grupo, como ilustrado na Figura 6.2. Em 6.2(a) têm-se os registros pertencentes ao Grupo 5 do arquivo *Synth1.data*, selecionados no *Viz3D* utilizando o processo de agrupamento visual proposto. Em (b) tem-se a visualização dos mesmos registros por Coordenadas Paralelas, que mostra claramente a presença de vários registros estranhos ao grupo. Estes podem ser eliminados por meio do processo interativo discutido na Seção 6.3.

Figura 6.2 – a) Registros identificados no processo de agrupamento visual no *Viz3D*; b) Visualização dos mesmos registros por Coordenadas Paralelas indica a presença de registros que não devem ser mantidos no grupo.

A identificação do grupo no *Viz3D* resultou em 766 registros (Figura 6.2(a)) e o refinamento aplicado sobre a visualização por Coordenadas Paralelas em 6.2(b) resultou na atribuição de 730 registros ao grupo. Segundo o processo de geração dos dados, 728 registros pertencem de fato ao Grupo 5 do conjunto de dados *Synth1.data*.

6.2.2 Identificação de Pontos Incomuns

O *Viz3D* também pode apoiar um processo de mineração visual para identificar pontos incomuns em dados. A Figura 6.3 exibe uma visualização *Viz3D* do conjunto de dados *Folhas.data*, colorindo os marcadores pelo atributo *classe*. Observa-se que a disposição dos eixos na visualização não é arbitrária, tendo sido gerada pela estratégia apresentada no Capítulo 7. Nota-se a existência de um marcador (registro 998), pertencente à classe 8 – cuja localização está indicada na Figura pela linha tracejada – bastante afastado dos demais registros da mesma classe.

Figura 6.3 – a) Visualização *Viz3D* do conjunto de dados *Folhas.data*, com os marcadores coloridos pelo atributo *classe*; b) Exibição parcial dos atributos dos elementos da Classe 8, em particular do registro 998; c) Comprovação do ponto incomum na visualização por Coordenadas Paralelas.

Uma análise dos valores na Figura 6.3(b) confirma que os valores dos atributos *df3* e *df4* do registro 998 são bastante discrepantes dos outros registros da mesma classe. A visualização por Coordenadas Paralelas confirma o comportamento incomum, conforme ilustrado na Figura 6.3(c), o que provavelmente indica um erro nos valores destes atributos.

6.2.3 Agrupamento com o *Viz3D Density Plot*

O mesmo processo interativo de agrupamento ilustrado para o *Viz3D* convencional pode ser aplicado ao *Viz3D Density Plot*. Adicionalmente, pode-se definir operações de

interação de maneira a controlar o nível de ruído da visualização com base em informação de frequência ou densidade. Por exemplo, pode-se exibir apenas o subconjunto dos dados associados a uma densidade superior a um certo limiar (uma operação de limiarização – *Thresholding*), e salvar tal subconjunto em um arquivo para processamento posterior. Desta maneira, pode-se eliminar o ruído, preservando apenas os registros mais relevantes sob o ponto de vista da informação de densidade. A Figura 6.4 ilustra este processamento com o conjunto de dados *Out5d.okc*. Em (a) são visualizados todos os 16.384 registros do conjunto de dados. Em (b), são exibidos apenas os registros (7.650) projetados em locais com densidade superior a um limiar definido pelo usuário.

Figura 6.4 – a) Visualização *Viz3D* do conjunto de dados *Out5d.okc*; b) Visualização dos registros que possuem densidade maior que um limiar definido pelo usuário.

Em (a) é possível identificar visualmente dois agrupamentos relevantes e, apesar de sua simplicidade, a limiarização consegue isolar satisfatoriamente os grupos de uma forma bem mais rápida, do que o procedimento apresentado na Seção 6.2, em que o usuário delimita manualmente cada região de interesse.

Como a informação de densidade exibida depende de dois parâmetros (largura do filtro de suavização no cálculo da densidade e o limiar de densidade mínima, como discutido no Capítulo 5), o resultado também depende desses parâmetros. No caso, valores grandes para a largura do filtro e valores pequenos para a densidade mínima, tendem a gerar menos grupos contendo mais registros (no limite, todos os marcadores são alocados em um único grupo), ao passo que larguras de filtro pequenas e limiares de densidade altos tendem a diferenciar mais os elementos, gerando mais grupos (no limite, cada elemento é um grupo) e, ainda, registros com baixa densidade tendem a ser eliminados. Na Figura

6.5(a) o usuário definiu a largura do filtro igual a 9 e a densidade mínima igual a 20, obtendo os 12.847 registros que são exibidos em (d). Em (b), o valor do limiar de densidade foi aumentado para 30, e a limiarização resulta nos 10.777 registros exibidos em (e). Aumentando o valor da densidade mínima para 40 são obtidos os 9.182 registros, que são exibidos em (f). A limiarização permite uma seleção de elementos que pode preceder o procedimento de agrupamento visual descrito anteriormente, reduzindo o congestionamento provocado pela exibição dos registros localizados em regiões menos densas.

Figura 6.5 – Seleção dos registros obtida com a operação de Limiarização a partir da informação de densidade no Viz3D.

6.3 Agrupamento Visual nos *IPC Frequency* e *Density Plots*

Os *Interactive Parallel Coordinates Frequency* e *Density plots* também permitem ao usuário identificar e extrair grupos interativamente. A Figura 6.6 mostra o *IPC Density Plot* do conjunto de dados *Synth1.data*, no qual o usuário delimita um intervalo de interesse sobre o quarto eixo para selecionar os registros. No caso a distribuição dos registros ao longo do quarto eixo indica a presença de três grupos bem definidos – o

intervalo selecionado delimita um desses grupos, mas outras áreas poderiam ter sido escolhidas inicialmente. Em (b), apenas os registros que satisfazem a seleção em (a) são exibidos. Em (c) os registros selecionados são exibidos com intensidade constante (em (b) a intensidade é proporcional à densidade). Este recurso ajuda o usuário a decidir se o grupo já foi devidamente demarcado com as seleções feitas. Neste caso, outras seleções são necessárias, delimitando intervalos sobre os outros eixos, para que o grupo de interesse seja isolado.

Figure 6.6 – Agrupamento visual interativo com o *IPC Density Plot*.

A Figura 6.6(d) mostra o grupo totalmente isolado, após o usuário ter realizado as demais marcações sobre os demais eixos, sendo que neste caso 729 registros foram

isolados. Finalizada a identificação do primeiro grupo, os registros são alocados no Grupo 1 e deixam de ser exibidos nas demais visualizações (Figura 6.6(e)), o que facilita a identificação visual dos demais grupos. O usuário pode repetir o procedimento iterativo até isolar todos os grupos. O processo permite ao usuário identificar com sucesso os quatro grupos relevantes do conjunto de dados, resultando em 729 registros no Grupo 1, 609 registros no Grupo 2, 728 no Grupo 3 e 850 no Grupo 4 (os cinco grupos do conjunto *Synth1.data* possuem 848, 728, 8, 608 e 728 registros). O grupo com apenas 8 registros não foi detectado. Por outro lado, quatro registros, originalmente classificados como ruído pelo algoritmo que gerou os dados, foram alocados nos grupos, e de fato, foram posicionados pelo algoritmo (aleatoriamente) no interior desses grupos.

A delimitação visual dos grupos por meio da seleção interativa demarcando os intervalos sobre os eixos pode ser imprecisa, sendo importante verificar se todos os registros dos grupos foram devidamente selecionados. Isto é feito observando se existem marcadores associados à alta frequência ou densidade muito próximos das extremidades dos intervalos definidos pelo usuário, o que pode indicar que o intervalo foi mal definido, e que registros podem ter sido indevidamente excluídos do grupo. Neste caso, o usuário deve cancelar as marcações e refazê-las definindo intervalos maiores nos locais em que isto ocorreu. A Figura 6.7 mostra como os grupos do conjunto de dados *Pollen.data* podem ser identificados com o processo descrito.

Figura 6.7 – a) Visualização do conjunto de dados *Pollen.data* usando o *IPC Density Plot* com as seleções do usuário sobre os quatro primeiros eixos; b) Exibição dos registros selecionados em (a).

Em (a) tem-se o *IPC Density Plot* com as marcações dos intervalos de interesse sobre os primeiros quatros atributos do gráfico. Em (b) tem-se a visualização dos 99 registros selecionados em (a), sendo que 98 deles pertencem aos seis grupos do conjunto

de dados. Um registro pertencente a um dos grupos não foi recuperado. Por outro lado, foi adicionado indevidamente um outro registro, o qual pode ser eliminado posteriormente. O usuário pode refinar a seleção operando sobre a visualização dos 99 registros selecionados em (b). O registros perdido também poderia ser recuperado, caso os intervalos sobre os eixos fossem expandidos e, posteriormente fosse realizado um refinamento do grupo, como ilustrado na Figura 6.2.

A Figura 6.8 mostra o processo de identificação de um dos grupos do conjunto de dados *Synth4.data* que possui 256.250 registros com 40 atributos. Em (a) tem-se o *IPC Density Plot*. Em (b) o usuário delimita um intervalo sobre o eixo correspondente ao atributo 35. O principal objetivo desta primeira seleção é obter uma visualização menos sobrecarregada, isolando um sub-conjunto de trabalho, pois nos demais passos apenas os registros selecionados participam do processamento. Em (c) é exibido o *IPC Density Plot* aplicado aos 121.393 registros selecionados. Em (d) o usuário delimita um intervalo sobre o atributo 30, refinando a seleção. Em (e) é exibido o *IPC Density Plot* dos 48.318 registros selecionados com a demarcação dos dois intervalos. Em (f) o usuário alterou a escala de tons de cinza ($s = 2.0$), para realçar a informação. Em (g) o usuário seleciona um intervalo sobre o atributo 37. Em (h) é exibido o *IPC Density Plot* dos 8.383 registros selecionados; entretanto, nota-se que a seleção sobre o atributo 35, realizada no início do processo, foi imprecisa na região inferior, sendo necessário refazer as seleções, no caso, ampliando o intervalo na parte inferior. Em (i) o usuário refaz todas as seleções, expandindo o intervalo sobre o eixo 35. Em seguida, o usuário especifica seleções adicionais sobre os eixos 1, 2, 3 e 9 em (j). Em (k) os 3.157 registros selecionados até o momento são exibidos com intensidade constante, evidenciando a necessidade de seleções adicionais para isolar o grupo. Em (l), após definir intervalos sobre os eixos 39, 40, 27, 28 e 29, restam apenas 3.146 registros, que visualmente, pertencem ao primeiro grupo identificado. No caso, considerando a constituição do conjunto de dados *Synth4.data*, descrita no Capítulo 2, deduz-se que os registros identificados no processo correspondem aos registros do grupo 28 (o único com 3.146 registros). Em (m) o usuário inicia o processo de encontrar o próximo grupo, sendo que neste caso, os registros alocados no primeiro grupo não são mais exibidos. A visualização em (n) exibe apenas os 3.146 registros do Grupo 28, confirmando que o grupo identificado no processo corresponde a este grupo.

Figura 6.8 -- Identificação de um grupo no conjunto de dados *Synth4.data*.

A Figura 6.9(a) mostra a visualização por Coordenadas Paralelas do conjunto de dados *Out5d.okc*. Em (b) tem-se a visualização com a técnica Coordenadas Paralelas Hierárquicas [Fua et al., 1999], que permite identificar algumas estruturas de interesse.

Figura 6.9 – a) Visualização do conjunto de dados *Out5d.okc* por Coordenadas Paralelas. b) Visualização no *XmdvTool* [Ward, 1994] usando a técnica Coordenadas Paralelas Hierárquicas (raios selecionados = 0.58 e *showband* = true).

A Figura 6.10 mostra a exploração visual desses mesmos dados no *IPC Density Plot*. Em (a) o usuário define um intervalo sobre o eixo número 3 (atributo *Potassium*). Em (b) são exibidos apenas os registros selecionados e, em seguida, o usuário define intervalos sobre os eixos 2 (*Magnesium*) e 4 (*Thorium*). Em (c) são exibidos com intensidade constante os 2.475 registros que pertencem ao agrupamento identificado.

Figura 6.10 – Identificação de um grupo no conjunto de dados *Out5d.okc* usando o *IPC Density Plot*.

A operação de limiarização de densidade também pode ser aplicada aos *IPC Frequency* e *Density Plot* para apoiar uma análise de agrupamentos. A Figura 6.11 ilustra a aplicação da operação de limiarização *AND* sobre o conjunto de dados *Synth1.data*, com diferentes valores de T . Em (a) tem-se os registros exibidos para o valor $T(AND) = 2$, enquanto o resultado obtido usando o limiar $T(AND) = 3$ é exibido em (b). Na primeira visualização são exibidos 2.773 registros e, na segunda, 1.802 registros (estes números são informados ao usuário pelo algoritmo). Embora vários registros pertencentes aos grupos não tenham sido recuperados (existem 2.920 registros nos grupos – ver Figura 5.14(c)), a limiarização pode ser vista como uma operação bastante simples para permitir uma exploração inicial dos dados, podendo ser útil para revelar a estrutura dos grupos mais relevantes nos dados (caso existam), mesmo na presença de bastante ruído (no caso, aproximadamente 60%).

Figure 6.11 – a) 2.773 registros obtidos usando uma limiarização $T(AND) = 2$; b) 1.802 registros obtidos usando o limiar $T(AND) = 3$.

A Figura 6.12 exibe o resultado da aplicação de uma limiarização OR sobre o conjunto de dados *Pollen.data*. Em (a), com $T(OR) = 3$, um total de 117 registros são recuperados. Os mesmos registros são exibidos em um gráfico de dispersão 3D na Figura 6.12 (b) – o gráfico de dispersão foi obtido projetando apenas os três primeiros atributos dos registros. A Figura 6.12(c) mostra 87 registros obtidos com $T(OR) = 4$, e a Figura 6.12(d) mostra o correspondente Gráfico de Dispersão 3D.

Figura 6.12 – a) 117 registros obtidos com limiar $T(OR) = 3$; b) respectiva projeção usando os três primeiros atributos dos registros; c) 87 registros obtidos com $T(OR) = 4$; d) respectiva projeção dos registros.

Nas visualizações das Figuras 6.12(a) e (b) (obtidas com $T(OR) = 3$) são observados alguns registros que não pertencem de fato ao agrupamento. Por outro lado, a limiarização com $T(OR) = 4$ retorna apenas 87 dos 99 registros que pertencem de fato aos grupos. Uma alternativa para isolar os grupos é usar $T(OR) = 3$ e, em seguida, interagir com a visualização para remover os registros estranhos.

6.4 Análise Visual Integrada

Em muitos casos, a análise apoiada por uma única técnica de visualização não consegue revelar todos os padrões presentes nos dados. A análise do mesmo conjunto de dados apoiada por diferentes técnicas de visualização é interessante para tratar diferentes aspectos dos dados, ou mesmo para confirmar padrões, como por exemplo, os agrupamentos identificados no *Viz3D* (Figura 6.2).

A Seção 6.21 mostrou como a técnica Coordenadas Paralelas pode ser usada para refinar os resultados obtidos com o *Viz3D*, por outro lado, embora visualizações por Coordenadas Paralelas, *IPC Frequency* ou *Density Plots* sejam eficazes para auxiliar a identificação de agrupamentos, nem sempre a interpretação dos resultados é intuitiva. Assim, uma visualização com o *Viz3D*, mais intuitiva, pode ser usada como apoio, para revelar a estrutura dos dados. A Figura 6.13(a) mostra uma visualização por Coordenadas Paralelas, dos registros do conjunto de dados *Pollen.data* pertencentes ao grupo identificado na Figura 6.7. Em (b) os mesmos registros são exibidos usando o *Viz3D*. O mesmo ocorre com a exibição do conjunto de dados *Hypercube.data* por Coordenadas Paralelas na Figura 6.13(c), de difícil interpretação, enquanto a sua visualização no *Viz3D*, apresentada na Figura 6.13(d), é mais reveladora. A Figura 6.13(e) mostra a visualização de 10 registros pertencentes a uma linha reta com dimensão 9, cuja interpretação não é tão imediata como a visualização dos mesmos registros no *Viz3D*, apresentada na Figura 6.13(f).

Figura 6.13 – a) Visualização por Coordenadas Paralelas dos registros do conjunto de dados *Pollen.data*, pertencentes ao grupo identificado na Figura 6.7; b) exibição dos mesmos registros no *Viz3D*; c) Visualização do conjunto de dados *Hypercube.data* por Coordenadas Paralelas; d) e no *Viz3D*; e) Visualização de 10 registros pertencentes a uma linha reta com dimensão 9 em Coordenadas Paralelas; f) Visualização dos mesmos registros no *Viz3D*.

A Figura 6.14 mostra a visualização do conjunto de dados *Image Segmentation.data*. Em (a), tem-se a visualização *Viz3D*, que consegue uma boa separação entre os registros da Classe 2 (*sky*) e os demais. Em (b) tem-se a visualização dos mesmos dados por Coordenadas Paralelas, que permite identificar os atributos que melhor separam os registros pertencentes à Classe 2 das demais classes, no caso, 10, 11, 12 e 13.

Figura 6.14 – Visualizações do conjunto de dados *Image Segmentation.data* a) *Viz3D*, b) *Coordenadas Paralelas*.

A Figura 6.15 ilustra a exploração do conjunto de dados *Houses.data*. Em (a) é exibida uma visualização *Viz3D* gerada usando apenas os atributos *Latitude* e *Longitude*, que revela bem o posicionamento geográfico das residências. Os marcadores são coloridos pelo atributo *classe* (pontos claros indicam residências com valor inferior a US\$180.000, enquanto que os pontos escuros indicam residências com valor superior a esse valor). A análise da visualização permite concluir que, em geral, residências de valores elevados são agrupadas em algumas áreas, embora possam ser identificadas algumas exceções. A visualização dos mesmos atributos por *Coordenadas Paralelas* convencionais não traz informações relevantes (Figura 6.15(b)), entretanto o *IPC Density Plot* (Figura 6.15(c))

revela a presença de grupos que, no caso, indicam as áreas densas, ou seja, com maior concentração de residências. Em (d) tem-se a visualização *Viz3D* dos grupos identificados em (c), interagindo com o *IPC Density Plot* conforme o processo de agrupamento visual descrito na Seção 6.3.

Figura 6.15 – Exploração do conjunto de dados *Houses.data*: a) visualização *Viz3D* usando apenas dos atributos Latitude e Longitude, marcadores coloridos pelo atributo *classe*; b) visualização por Coordenadas Paralelas; c) *IPC Density Plot* indica áreas de maior concentração de residências; d) visualização *Viz3D* de alguns grupos identificados pelo procedimento de agrupamento visual apoiado no *IPC Density Plot* exibido em (c).

A visualização do conjunto de dados *Cancer.data* em Coordenadas Paralelas exibida na Figura 6.16(a) não apresenta nenhum padrão de comportamento evidente. Entretanto, a visualização *Viz3D* permite determinar uma certa separação entre as duas classes, que pode auxiliar a classificação dos registros a partir de seu posicionamento no

espaço 3D, definindo visualmente, um plano de separação, como ilustrado na Figura 6.16(b).

Figura 6.16 – Visualização do conjunto de dados *Cancer.data* a) em Coordenadas Paralelas; b) no *Viz3D*.

6.5 Análise dos Agrupamentos

Uma das principais dificuldades na avaliação dos resultados de um algoritmo de agrupamento é a incapacidade de visualizar propriedades geométricas do espaço multidimensional [Tou e Gonzalez, 1974]. Como discutido no Capítulo 3, a análise de agrupamentos procura encontrar classes, de modo que registros similares sejam colocados em uma mesma classe, maximizando as similaridades intraclasse e minimizando as similaridades interclasses. Assim, uma forma de avaliar a qualidade dos agrupamentos obtidos consiste em gerar métricas da similaridade intra-grupo e inter-grupos, as quais normalmente são baseadas em medidas de distância computadas no espaço n -dimensional dos dados.

Tou e Gonzalez [1974] sugerem uma métrica de distância intra-grupo determinada pela média das distâncias entre todos os elementos do grupo.

$$Distancia\ Intra-Grupo = \frac{1}{k(k-1)} \sum_{j=1}^k \sum_{i=1}^k \sum_{k=1}^n (a_k^j - a_k^i)^2 \quad (6.1)$$

A Figura 6.17 ilustra as distâncias consideradas no cálculo da distância definida pela Equação 6.1, para dois grupos distintos. No caso, são calculadas as dez distâncias indicadas pelos segmentos de reta que ligam os registros em cada grupo, (a) e (b).

Figura 6.17 – Distâncias usadas no cálculo da distância intra-grupo para a) um agrupamento circular e b) um agrupamento alongado.

Entretanto, com essa medida grupos alongados sempre apresentam valores altos de distância intra-grupo, pois ainda que cada ponto do grupo esteja próximo a algum outro, os pontos nos extremos ficam muito afastados, comprometendo a medida. Neste trabalho, adota-se outra métrica de distância intra-grupo, computada como a média das menores distâncias entre todos os registros do grupo, considerando que todos os registros fiquem conectados pelos caminhos usados para medir a distância intra-grupo. No caso dos registros ilustrados na Figura 6.17, são então computadas em cada grupo apenas as quatro distâncias euclidianas indicadas na Figura 6.18.

Figura 6.18 – Distâncias usadas na métrica proposta para o cálculo da distância intra-grupo a) agrupamento circular; e b) agrupamento alongado.

O algoritmo que gera a métrica proposta é apresentado na Caixa 6.1.

<p>Supondo S_k o conjunto contendo todos os registros do grupo k e S_m um conjunto vazio</p> <p>1 – Remova os dois registros r_i e $r_j \in S_k$ mais próximos entre si e os coloque em S_m</p> <p>2 – Faça distância = distância(r_i, r_j); // distância euclidiana</p> <p>3 – Remova o registro $r_i \in S_k$ mais próximo ao registro $r_j \in S_m$</p> <p>4 – Faça distância = distância + distância(r_i, r_j);</p> <p>5 – Repita os passos 3 e 4 até remover todos os registros de S_k e coloca-los em S_m</p>

Caixa 6.1 – Algoritmo proposto para obter a distância intra-cluster.

Tou e Gonzalez [1974] sugerem obter a métrica de similaridade inter-grupos medindo a distância entre os centróides dos grupos. Entretanto, novamente agrupamentos

alongados apresentam problemas com esta estratégia. Assim, neste trabalho a distância inter-grupo entre dois grupos S_k e S_l é obtida medindo a menor distância entre dois registros $r_i \in S_k$ e $r_j \in S_l$.

Dados n grupos S_i $i = 1, \dots, n$:

1 – Para $i = 1$ até $n-1$

Para $j = i + 1$ até n

Faça $distancia\ inter-grupo(i,j)$ = menor distancia entre os elementos dos grupos S_i e S_j

Caixa 6.2 – Algoritmo proposto para obter a distância inter-cluster.

Quando a classe a qual cada registro deve ser alocado é conhecida, os resultados do agrupamento também podem ser avaliados por meio de uma matriz de confusão, que informa os erros e acertos do processo de agrupamento. Este método, bastante usado para comparar os resultados obtidos com os algoritmos de classificação, é usado, por exemplo, por Zhang et al. [1996], para comparar os resultados obtidos com o *CLARANS* e o *BIRCH*. A diagonal da matriz de confusão informa os registros corretamente alocados em seus respectivos grupos, e os demais elementos da matriz informam os registros de uma determinada classe erroneamente alocados. A Figura 6.19 apresenta a matriz de confusão para o conjunto de dados *Synt1.data*, que contém 5 classes e ruído (classe 0), e os resultados obtidos com o agrupamento visual ilustrado na Figura 6.6.

	0	1	2	3	4
0	4575	2	1	0	2
1	0	848	0	0	0
2	0	0	728	0	0
3	8	0	0	0	0
4	0	0	0	608	0
5	0	0	0	0	728

(a)

	0	1	2	3	4	5
Formação das classes	4580	848	728	8	608	728
Formação dos grupos	4583	850	729	608	730	-

(b)

Figura 6.19 – Análise dos grupos obtidos na com o processo de agrupamento visual ilustrado na Figura 6.6: a) Matriz de Confusão; b) Descritivo dos grupos (classes indicam a formação real e obtidos).

Uma análise dos quadros mostra que a Classe 0, que agrupa todo o ruído embutido nos dados, possuía 4.580 registros e, no processo de agrupamento, recebeu 4.583 registros. Dois registros classificados como ruído no processo de geração foram atribuídos ao Grupo 1, um registro foi para o Grupo 2 e outros dois para o Grupo 4. O Grupo 0 recebeu também os oito registros da Classe 3, que não foi identificada visualmente. Aos 848 registros

efetivamente pertencentes à Classe 1 foram somados dois registros da Classe 2, assim o Grupo 1 ficou com 850 registros. Da mesma forma, o Grupo 2 recebeu os 728 registros da Classe 2, além de um registro da Classe 0, ficando com 729 registros. Os 608 registros da Classe 4 foram identificados sem erros e alocados ao Grupo 3. Por fim o Grupo 4 recebeu os registros da Classe 5, além de dois registros vindos da Classe 0.

A Figura 6.20(a) apresenta uma visualização *Viz3D* do conjunto de dados *Synth1.data*. Em (b) são exibidos os valores de distância intra-grupo (na diagonal) e inter-grupos, obtidas com os algoritmos apresentados nas Caixas 6.1 e 6.2. Uma análise da tabela na Figura 6.20(b) mostra que o Grupo 3, apesar de ter poucos registros, apresenta uma distância intra-grupo relativamente alta, o que indica uma grande separação entre os seus elementos. Isso explica a não identificação do grupo no procedimento de agrupamento visual apoiado pelo *IPC Density Plot* (Figura 6.6). O Grupo 5, apesar de ter 728 registros, apresenta baixa distância intra-grupo, indicando ser o grupo mais denso de todos os grupos do conjunto. A tabela também permite verificar a capacidade do *Viz3D* preservar a estrutura dos dados, pois os Grupos 1 e 3, bastante distantes entre si na visualização, possuem elevada distância inter-grupo; já os Grupos 3 e 5, próximos na visualização, apresentam a menor distância inter-grupo. No caso, a distância entre os grupos 4 e 5, bastante elevada, não é evidenciada na visualização.

Figura 6.20 – a) Visualização do conjunto de dados *Synth1.data* no *Viz3D*; b) Valores de distância intra-grupo e inter-grupo, usando a métrica Euclidiana.

6.6 Considerações Finais

Técnicas de exploração visual de dados são importantes porque dão controle ao usuário e inserem o potencial humano de perceber padrões no processo de investigação. Visualizações baseadas em informação de frequência e densidade, como os *Viz3D Density Plot*, *IPC Frequency* e *Density Plot*, mostram-se bastante eficazes para apoiar processos

interativos de mineração de agrupamentos. Entretanto, é importante dispor de mecanismos para refinar a identificação de grupos, por exemplo, analisando informações complementares como estatísticas ou métricas descritivas dos grupos obtidos. O refinamento também pode ser executado com diferentes técnicas de visualização, apoiando-se mutuamente. Por outro lado, a identificação interativa de grupos apoiada por visualizações exige alguma experiência do usuário e bons recursos de interação, pois nem sempre os grupos estão claramente bem definidos nas representações visuais.

Nos experimentos realizados com o *Viz3D Density Plot* na Seção 6.2, a informação de densidades foi obtida usando uma grade tridimensional 300x300x300, enquanto os experimentos realizados com os *IPC Frequency* e *Density Plots* na Seção 6.3 utilizaram uma grade (Matriz de Frequências - *F*) com resolução 720x550 (a resolução dos gráficos gerados), entretanto, nos dois casos, outras resoluções de grade poderiam ser utilizadas.

O processo de agrupamento proposto, usando o *Viz3D* e *Viz3D Density Plot*, que projetam os registros em um espaço 3D, são importantes alternativas ao *Eureka'* [Manco et al., 2004], que projeta os dados em 2D, usando a decomposição em valores singulares (*SVD*), e ainda não prevê a utilização da informação de densidade para auxiliar a identificação dos grupos. Quanto às técnicas *HD-Eye* [Hinneburg et al., 1999], *HC-Cooperative* [Aggarwal, 2001] e *HC-Enhanced* [Gamero e Minghim, 2003] que exibem informação de densidade em projeções dos dados multidimensionais em espaços bi ou tridimensionais, o maior problema é que o usuário tem dificuldades para deduzir a estrutura dos dados no espaço n -dimensional.

Capítulo 7

Arranjo e Redução de Atributos

7.1 Considerações Iniciais

Técnicas de redução da dimensionalidade [Fodor, 2002] e de seleção de atributos [Molina et al., 2002] são usualmente aplicadas aos dados antes de algoritmos de agrupamento, uma vez que a alta dimensionalidade dificulta a operação desses algoritmos e afeta o seu desempenho [Jain, 1999], [Berkhin, 2002]. Da mesma forma, técnicas de Visualização de Informação, ainda que projetadas para operar sobre dados multidimensionais, também apresentam limitações quando a dimensionalidade é muito alta. Em várias dessas técnicas, a sequência em que os atributos são processados e exibidos afeta bastante a qualidade da visualização resultante [Ankerst et al., 1998], conforme já foi ilustrado na Figura 2.12.

Neste capítulo é apresentada uma técnica, denominada Arranjo de Atributos Baseado em Similaridade (*SBAA – Similarity-Based Attribute Arrangement*), muito simples e aplicável a dados de alta dimensionalidade, a qual gera arranjos de exibição de atributos para as técnicas *RadViz*, *Viz3D* e Coordenadas Paralelas que resultam em visualizações mais eficazes para o usuário. A estratégia considera especificidades destas técnicas, que são beneficiadas quando eixos associados a atributos similares são posicionados próximos entre si. No caso de Coordenadas Paralelas, o arranjo de eixos gerado pela técnica *SBAA* reduz o cruzamento excessivo das poligonais, o que resulta em gráficos menos congestionados. No *RadViz* e *Viz3D*, que adotam uma disposição radial dos eixos, o arranjo gerado evita que atributos similares sejam posicionados em extremos opostos e, conseqüentemente, que os marcadores gráficos sejam deslocados indevidamente para a região central. Com isso, obtém-se visualizações menos congestionadas e que separam melhor os diferentes agrupamentos existentes nos dados. Basicamente, a estratégia proposta busca simplesmente gerar um arranjo em que eixos associados a atributos similares sejam posicionados próximos entre si. Também são introduzidas duas

variantes da técnica *SBAA* que dão ao usuário a opção de reduzir o número de atributos exibidos na visualização, as quais foram denominadas Redução de Atributos Baseada em Similaridade (*SBAR* – *Similarity-Based Attribute Reduction*) e Redução de Atributos Baseada em Similaridade Modificada (*mSBAR* – *modified Similarity-Based Attribute Reduction*). Estas técnicas permitem remover da visualização atributos que possuem alta correlação com algum outro, de acordo com a métrica de similaridade adotada.

Na Seção 7.2 deste capítulo são discutidos sucintamente alguns trabalhos da literatura que tratam do problema do arranjo e da redução de atributos em visualização. Os algoritmos propostos no âmbito deste trabalho para configurar (arranjar e reduzir) os atributos por similaridade são introduzidos na Seção 7.3, e uma breve análise de desempenho do *SBAA* é apresentada na Seção 7.4. Na Seção 7.5 são apresentadas diversas visualizações *RadViz*, *Viz3D* e Coordenadas Paralelas de diferentes conjuntos de dados reais e sintéticos para ilustrar as potenciais vantagens das estratégias propostas. Os comentários finais são apresentados na Seção 7.6.

7.2 Trabalhos Relacionados

O número de possíveis arranjos de eixos para exibir n eixos mapeando n atributos é $n!$ (embora muitos desses arranjos sejam equivalentes entre si), o que caracteriza um problema não polinomial. Quanto à ordem dos atributos nas visualizações, Ankerst et al. [1998] afirmam que, em geral, atributos similares devem ser posicionados próximos, sugerindo heurísticas inspiradas em um sistema biológico (colônia de formigas [Dorigo e Gambardella, 1997]) como alternativa para contornar a complexidade do problema. Peng et al., [2004] abordam o mesmo problema de maneira totalmente diferente, estabelecendo algumas métricas para estimar o congestionamento visual (*clutter*) em diversas técnicas de visualização, como Coordenadas Paralelas, Matriz de Gráficos de Dispersão, Ícones Estrela e Empilhamento de Dimensões – critérios diferentes são sugeridos para diferentes técnicas de visualização. Em seguida, constroem as visualizações com diferentes arranjos dos atributos, sendo que o melhor arranjo será aquele que minimizar as métricas de congestionamento visual. Os trabalhos de Ankerst et al. [1998] e de Peng et al., [2004] são dos poucos encontrados na literatura que abordam o problema de definir uma seqüência de exibição dos atributos de modo a gerar visualizações de melhor qualidade.

Por outro lado, técnicas de redução de dimensionalidade [Fodor, 2002], como as discutidas na Seção 4.2, são abundantes. Entretanto, soluções que adotam abordagens

convencionais, como a Análise de Componentes Principais (*PCA*) [Pearson, 1901] e o *FastMap* [Faloutsos e Lin, 1995], em que os dados n -dimensionais são projetados em um espaço k -dimensional ($k < n$), têm como desvantagem perder os valores originais dos atributos no espaço original de entrada. Algoritmos de seleção de atributos (*Feature Selection*) [Molina et al., 2002] adotam diferentes estratégias para identificar quais atributos devem ser selecionados para uma análise em particular, e representam uma alternativa para a redução de dimensionalidade em que os valores originais dos atributos escolhidos são preservados.

De maneira geral, técnicas de seleção de atributos podem ser classificadas em duas categorias. Na primeira, conhecida como Seleção Sequencial Regressiva (*Sequential Backward Selection – SBS*), atributos são gradualmente removidos do conjunto completo, segundo algum critério, até atingir o número desejado de atributos. Na segunda, conhecida como Seleção Sequencial Progressiva (*Sequential Forward Selection – SFS*), atributos são gradualmente inseridos no conjunto inicialmente vazio, também até atingir o número desejado. Uma solução interessante na primeira categoria é apresentada por Traina et al. [2000] e Eneva et al. [2002], que propõem a dimensão fractal dos dados como critério para definir os atributos a serem removidos. A idéia é remover os atributos cuja exclusão introduz a menor variação no valor da dimensão fractal do conjunto de todos os atributos. Outras propostas, apresentadas por Yang et al. [2003; 2003a], adotam uma hierarquia entre os atributos e, em seguida, os atributos são classificados e exibidos em uma estrutura de árvore, de modo que atributos similares sejam posicionados próximos. O conjunto reduzido de atributos pode, então, ser obtido definindo os níveis de interesse na hierarquia da árvore, o que é feito visualmente, especificando o número final de atributos desejados.

7.3 Arranjo de Atributos Baseado em Similaridade

Sendo D_{mn} o conjunto de dados com m registros de n atributos, a estratégia *SBAA* tem como ponto de partida a construção de uma matriz de similaridade S (diagonal inferior) contendo as medidas de similaridade entre todos os pares de atributos.

$$S = \begin{bmatrix} - & - & - & - & - \\ s_{1,0} & - & - & - & - \\ s_{1,0} & s_{2,0} & - & - & - \\ \dots & \dots & \dots & \dots & - \\ s_{n,0} & s_{n,1} & \dots & s_{p,n} \end{bmatrix} \quad (7.1)$$

sendo $s_{i,j}$ alguma medida de similaridade entre o i -ésimo e o j -ésimo atributos, por exemplo:

$$s_{i,j} = 1 - \frac{1}{m} \left[\sum_{k=1}^m \left| \frac{d_{k,i} - \text{Min}_i}{\text{Max}_i - \text{Min}_i} - \frac{d_{k,j} - \text{Min}_j}{\text{Max}_j - \text{Min}_j} \right| \right] \quad (7.2)$$

onde $\text{Min}_i = \text{mínimo}(d_{k,i})$; $k = 1, \dots, m$.

A medida de similaridade dada pela Equação (7.2) é invariante à translação e escala. Entretanto, apresenta problemas no caso de dados ruidosos, o que pode ser contornado com uma normalização inicial dos dados em termos das médias e variâncias, por exemplo, aplicando a Equação 7.3.

$$s_{i,j} = \frac{1}{m} \left[\sum \left(\frac{d_i - \mu_i}{\sigma_i} + \frac{d_j - \mu_j}{\sigma_j} \right)^2 \right] \quad (7.3)$$

onde μ_i e σ_i são, respectivamente, a média e a variância do i -ésimo atributo. No caso da Equação 7.2, valores $s_{i,j}$ próximos de 1.0 indicam alta similaridade entre os respectivos atributos. Outra medida de similaridade interessante para definir os elementos da matriz S é o coeficiente de correlação de Pearson, dado por:

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \quad (7.4)$$

em que $\text{cov}(x, y)$ é a covariância entre os atributos x e y , dada por:

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)(y_i - \mu_y) \quad (7.5)$$

e σ_x e σ_y são os desvios padrão de x e y , dados por:

$$\sigma_x = \sqrt{\frac{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}{n(n-1)}} \quad \text{e} \quad \sigma_y = \sqrt{\frac{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2}{n(n-1)}} \quad (7.6)$$

O coeficiente pode ser calculado de maneira mais eficiente com a Equação 7.7.

$$r = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{\sqrt{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \sqrt{n \sum_{i=1}^n y_i^2 - \left(\sum_{i=1}^n y_i \right)^2}} \quad (7.7)$$

A proposta de Ankerst et al. [1998] também utiliza a matriz de similaridade (7.1), entretanto, busca encontrar o arranjo que minimiza as similaridades entre os atributos adjacentes, fazendo uma analogia com o problema do caixeiro viajante. O processo adotado pela *SBAA* é bem mais simples, pois, após calcular todos os elementos de S , o algoritmo busca o maior elemento s_{ij} de S , que determinará os dois atributos i e j que constituirão o arranjo inicial “ ij ”. No caso, os atributos i e j constituirão, respectivamente, os eixos nos extremos esquerdo e direito do arranjo inicial. Em seguida, o algoritmo busca nas linhas e colunas de S o próximo atributo k a ser inserido no arranjo, que deve ser o atributo mais similar a i ou a j . O atributo k é anexado na extremidade direita no arranjo (resultando em “ ijk ”) se $s_{j,k} > s_{i,k}$, e na extremidade esquerda (resultando em “ kij ”), caso contrário. No primeiro caso, os elementos da linha e da coluna j da matriz S deixam de participar das próximas buscas, e no segundo caso a linha e coluna i serão eliminadas das buscas seguintes. O processo é repetido até que todos os demais atributos do conjunto de dados estejam inseridos no arranjo, o qual é denominado arranjo A . Uma medida de similaridade global $\sum_A$ associada ao arranjo A é obtida pela somatória dos valores de similaridade dos pares de atributos consecutivos na sequência, por exemplo, supondo $A = “ijk”$, então $\sum_A(“ijk”) = s_{i,j} + s_{j,k}$. A Caixa 7.1 mostra o algoritmo *SBAA*, que tem como entrada a matriz de dados D_{nm} .

```

1 – Faça arranjo  $A = ""$  //  $A = \emptyset$ 
2 – Calcule a matriz  $S$ , por exemplo, usando a medida de similaridade da Equação 7.2 ou 7.4:
3 – A partir do maior de valor  $s_{i,j}$ , obtenha os atributos  $i$  e  $j$  ( $i \neq j$ ) mais similares:
4 – Faça arranjo  $A = "ij"$ 
    $Extremo\ Esquerdo = i;$ 
    $Extremo\ Direito = j;$ 
5 – Obtenha o maior valor entre  $s_{i,k}$  e  $s_{j,k}$  tal que  $k \notin A$ 
6 – Se  $s_{i,k} > s_{j,k}$  faça:
   Arranjo  $A = "kij"$  //  $k$  é mais similar ao Extremo Esquerdo
    $Extremo\ Esquerdo = k$ 
   caso contrário faça:
   Arranjo  $A = "ijk"$  //  $k$  é mais similar ao Extremo Direito
    $Extremo\ Direito = k$ 
7 – Repita 5 e 6 até inserir todos os atributos no arranjo  $A$ .

```

Caixa 7.1 – Algoritmo SBAA (*Similarity Based Attribute Arrangement*).

7.4 Redução da Dimensionalidade

Além de definir um arranjo dos atributos mais conveniente para a visualização do que o arranjo arbitrário usualmente adotado, o arranjo A e sua medida de similaridade associada Σ_A , juntamente com a matriz S , permitem implementar um processo automático de remoção de atributos da visualização, o qual foi denominado *SBAR* (*Similarity Based Attribute Reduction*), que adota o critério de eliminar os atributos que são mais similares ao seu vizinho no arranjo definido. Pode-se identificar os dois atributos consecutivos i e j mais similares (maior valor $s_{i,j}$) e eliminar aquele (i ou j) que resulta no menor valor global de similaridade para o arranjo Σ_A . Esse processo pode se repetir até atingir um número de atributos pré-determinado pelo usuário, ou até a visualização ser considerada ‘adequada’ pelo usuário (muitas vezes o número excessivo de atributos dificulta a interpretação), por serem exibidos apenas os atributos ‘pouco’ similares. O algoritmo *SBAR* é apresentado na Caixa 7.2 e tem como entrada apenas a matriz de dados D_{nm} .

```

1 – Obtenha o arranjo  $A$  usando o algoritmo SBAA
2 – Obtenha o maior  $s_{i,j}$  tal que  $i$  e  $j$  são consecutivos em  $A$ 
3 – Remova do arranjo  $A$  o atributo  $i$  ou  $j$ , selecionando aquele que resulta no menor valor de  $\Sigma_A$ 
4 – Repita o passo 3 até obter o número de atributos desejado em  $A$ 

```

Caixa 7.2 – Algoritmo SBAR (*Similarity Based Attribute Reduction*).

Alternativamente, pode-se re-arranjar por similaridade os atributos restantes após cada remoção, antes de eliminar o próximo, conforme apresentado na Caixa 7.3. Esta

variação, denominada *mSBAR* (*modified Similarity Based Attribute Reduction*), apresenta resultados diferentes do *SBAR* convencional, como ilustrado na Seção 7.4.

- 1 – Calcule a matriz S como no algoritmo *SBAA*
- 2 – Obtenha o arranjo A usando a matriz S
- 3 – Obtenha o maior s_{ij} tal que i e j são consecutivos em A
- 4 – Remova o atributo i ou j do arranjo A , que resultar no menor valor de $\sum A$
- 5 – Repita os passos 2 à 4, ignorando os atributos removidos anteriormente até obter o número de atributos desejado em A

Caixa 7.3 – Algoritmo *mSBAR* (*modified Similarity Based Attribute Reduction*).

Não há diferença significativa de tempos de execução entre as variantes *SBAR* (com e sem reordenação dos atributos remanescentes após cada eliminação de atributo), pois a informação de similaridade não precisa ser recalculada a cada passo. Na Seção 7.4 são apresentados alguns experimentos em que estas estratégias são adotadas para arranjar e reduzir atributos em visualizações por Coordenadas Paralelas, *RadViz* e *Viz3D*. Em seu trabalho, Ankerst et al. [1998] identificam diversas configurações de atributos (eixos) possíveis, em particular a Linear 1D, Circular 1D e 2D, ilustradas na Figura 7.1. A configuração Linear 1D é adotada em visualizações por Coordenadas Paralelas, enquanto que a configuração Circular 1D é usada em técnicas similares ao *RadViz* e *Viz3D*. A configuração 2D é adotada, por exemplo, em técnicas orientadas a pixel.

Figura 7.1 – Configurações dos atributos: a) Linear 1D, b) Circular 1D e c) 2D [Ankerst et al., 1998].

No caso da configuração Linear 1D, em que os eixos são posicionados lado a lado aos pares, o arranjo gerado usando *SBAA* define uma sequência de pares de modo que eixos vizinhos correspondem a pares mais similares do que os não vizinhos. Por outro lado, no caso da configuração circular, o *SBAA* ignora a similaridade entre o primeiro e o último atributos, que são posicionados como vizinhos na configuração circular – para todos os demais pares de atributos o posicionamento considera a similaridade entre o par. A Figura

7.2 ilustra visualizações por Coordenadas Paralelas e *RadViz* de dois registros de seis atributos, comparando o resultado obtido com o posicionamento arbitrário dos eixos e com o posicionamento gerado pelo algoritmo *SBAA*. A visualização *RadViz* com o arranjo arbitrário fica comprometida, pois os dois registros, apesar de bastante distintos, são projetados na mesma posição; por outro lado os registros ficam bem separados com o arranjo de eixos baseado na similaridade dos atributos. A visualização *Viz3D* desses dados gera uma situação semelhante. No caso da visualização por Coordenadas Paralelas a visualização inicial apresenta muitos cruzamentos entre as poligonais, o que não ocorre adotando o arranjo por similaridade.

Figura 7.2 - a) Conjunto de dados contendo dois registros de seis atributos; b) Visualização *RadViz* com o arranjo dos atributos segundo a sequência dada nos registros; c) Visualização por Coordenadas Paralelas, com o mesmo arranjo; d) Visualização *RadViz* com o arranjo baseado em similaridade; e) Visualização por Coordenadas Paralelas com o mesmo arranjo.

7.5 Desempenho

A maior parte do processamento realizado pelo algoritmo *SBAA* consiste no cálculo da matriz S , que contém $n^2/2$ elementos para um conjunto de registros com n atributos. Os tempos de processamento obtidos para conjuntos de dados com diferentes números de registros (m) e de atributos (n) são ilustrados na Figura 7.3. A máquina utilizada nos experimentos possui microprocessador AMD Athlon^(tm), operando em 1.91 GHZ (1Gbyte de memória RAM).

Figura 7.3 – a) Tempos de execução, em segundos, do algoritmo *SBAA*. b) respectivo gráfico em escala logarítmica.

7.6 Experimentos

Nesta Seção são apresentados alguns resultados obtidos com os algoritmos *SBAA*, *SBAR* e *mSBAR*, aplicados a visualizações por Coordenadas Paralelas, *RadViz* e *Viz3D*. A Figura 7.4(a) mostra uma visualização *RadViz* do conjunto de dados *Mammals2.data*, em que os eixos adotam a mesma sequência de atributos fornecida no arquivo de dados. Em 7.4(b) tem-se a visualização obtida após configurar os atributos com o algoritmo *SBAA*, usando a medida de similaridade da Equação 7.2. Enquanto em (a) os quatro grupos se sobrepõem, em (b) tem-se uma boa separação entre eles.

Figura 7.4 – a) Visualização do conjunto de dados *Mammals2.data* no *RadViz*, adotando o arranjo original dos atributos; b) Visualização que adota o arranjo baseado em similaridade.

A Figura 7.5(a) apresenta a visualizações do conjunto de dados *Mammals2.data* no *Viz3D*, usando a sequência original dos atributos no arquivo de dados. Em 7.5(b) tem-se a visualização com o arranjo obtido com o *SBAA* usando a medida de similaridade dada pela Equação 7.2, que resulta novamente em uma boa separação entre os quatro grupos existentes nos dados.

Figura 7.5 – Visualizações *Viz3D* do conjunto de dados *Mammals2.dat*: a) com um arranjo arbitrário dos atributos; b) com o arranjo baseado na similaridade dos atributos.

A Figura 7.6 permite comparar a diferença entre as visualizações *Viz3D* do conjunto de dados *Synth3.data*, obtidas adotando o arranjo original (a) e adotando o arranjo obtido com o *SBAA* (b). Em ambos os casos, todos os atributos foram usados como entrada para gerar as visualizações. Observe que na Figura 7.6(a) é possível identificar apenas sete dos oito grupos contidos nos dados, enquanto que em 7.6(b) é possível identificar claramente os oito grupos.

Figura 7.6 – Visualizações *Viz3D* do conjunto de dados *Synth3.data*: a) arranjo original; b) arranjo obtido com o *SBAA*.

A Tabela 7.1 exibe as distâncias (Euclidianas) intra e inter-grupo relativas aos oito grupos do conjunto *Synth3.data*, computadas considerando os 20 atributos na sequência original, usando os algoritmos descritos no Capítulo 6 (Caixas 6.1 e 6.2). As distâncias intra-grupos são posicionadas na diagonal da matriz, pois indicam as distâncias entre os elementos do mesmo grupo – $dt(i,i)$. A Tabela 7.2 apresenta as medidas equivalentes obtidas a partir das coordenadas tridimensionais geradas pela projeção efetuada no *Viz3D*, considerando o arranjo original dos atributos. Por fim, na Tabela 7.3 têm-se as medidas de distância obtidas com as coordenadas geradas pelo *Viz3D* considerando o arranjo dos atributos obtido com o *SBAA*, usando a medida de similaridade da Equação 7.2.

Tabela 7.1 – Distâncias intra-e inter-grupo relativas aos grupos do conjunto de dados *Synth3.data* obtidas a partir dos 20 atributos na sequência original dada no arquivo.

	1	2	3	4	5	6	7	8
1	1026.7292							
2	783.1000	1001.9883						
3	1358.9000	1420.0500	984.9929					
4	1034.7500	881.6500	1517.7500	1156.6227				
5	1158.2500	572.1500	1710.1000	1156.2000	332.3319			
6	552.2000	969.7500	1679.8000	1066.5000	1416.3000	84.5920		
7	928.7500	1165.1500	963.1500	949.9500	1478.7500	1122.6000	511.6462	
8	1587.7000	905.8000	910.0500	956.6000	1319.6000	2186.2000	1166.4000	1573.4642

Tabela 7.2 – Distâncias intra e inter-grupos relativas aos grupos do conjunto de dados *Synth3.data* obtidas a partir das coordenadas tridimensionais geradas no *Viz3D*, considerando a configuração original dos atributos.

	1	2	3	4	5	6	7	8
1	1639.4155							
2	1.9883	10365.7158						
3	116.8893	138.0638	4809.0103					
4	41.9563	13.7620	280.6305	7168.6284				
5	0.0003	0.0681	141.6113	48.5348	571.8056			
6	5.7871	24.7505	299.9836	62.7032	1.2664	400.3593		
7	77.4332	24.9501	139.3563	17.6148	90.4747	192.2090	952.4450	
8	55.7037	73.8445	82.8098	57.6761	76.6096	140.4263	22.6419	8740.3594

Tabela 7.3 – Distâncias intra- e inter-grupos relativas aos grupos do conjunto de dados *Synth3.data* obtidas a partir das coordenadas tridimensionais geradas no *Viz3D*, considerando os atributos arranjados com o *SBAA*, usando a Equação 7.2.

	1	2	3	4	5	6	7	8
1	666.7539							
2	168.0311	4166.7959						
3	593.6727	287.2235	9053.2197					
4	71.2081	35.0773	206.4657	9602.1797				
5	419.2385	52.7767	621.4876	288.1851	7322.0249			
6	8.7779	104.0157	480.3478	77.4482	278.5583	4459.3037		
7	208.3000	52.4537	190.9741	0.9937	363.6246	235.3483	6053.6943	
8	247.8936	11.9576	119.0481	22.3802	177.7615	183.0798	21.4153	2753.6697

Uma análise dos valores nas tabelas mostra que, sem arranjar os atributos por similaridade, a visualização *Viz3D* apresenta alguns grupos (1 e 5 e 2 e 5, por exemplo) muito próximos entre si, ou seja, com pequena distância inter-grupos (valores 0.0003 e 0.0681 na Tabela 7.2). Após arranjar os atributos por similaridade os grupos ficam mais separados na visualização, o que pode ser observado visualmente na Figura 7.6(a) e na Tabela 7.3, na qual a menor distância inter-grupo é 0.9937, entre os grupos 4 e 7.

A Figura 7.7(a) ilustra a visualização *Viz3D* do conjunto de dados *Synthetic Data Set 2*, obtida adotando a configuração original dos atributos, na qual se observa uma separação ruim entre os grupos. As regiões delimitadas pelos círculos tracejados na Figura 7.7(a) indicam locais em que ocorrem sobreposições de grupos diferentes na projeção. Em 7.7(b) tem-se a visualização usando a sequência de atributos (7-11-3-9-5-8-2-10-4-6), obtida com *SBAA* adotando a medida de similaridade da Equação 7.2, a qual consegue uma separação adequada dos quatorze grupos existentes nos dados.

Figura 7.7 – a) Visualização *Viz3D* do conjunto de dados *Synthetic Data Set 2*, adotando a sequência original dos atributos; b) Visualização *Viz3D* adotando a disposição dos eixos obtida com o *SBAA*.

Os algoritmos *SBAR* e *mSBAR* oferecem uma estratégia para a redução dos atributos que pode ser útil para reduzir o congestionamento visual quando o número de eixos é muito alto (ver Figura 5.3(b)), mas ainda procurando preservar a estrutura dos dados no espaço original. Isso é ilustrado na Figura 7.8, que mostra visualizações por Coordenadas Paralelas do conjunto de dados *Mammals2.data*, com 72 atributos. A visualização em

7.8(a) adota o arranjo dos atributos fornecido no conjunto de dados e a visualização em 7.7(b) adota o arranjo obtido com o *SBAA*, usando a Equação 7.2. Em (c) tem-se a visualização obtida utilizando os 40 atributos mais dissimilares obtidos com o *SBAR* (usando a Equação 7.2) e, em (d), a visualização obtida utilizando apenas os 30 atributos mais dissimilares. Comparando a visualização na Figura 7.8(b), que usa os 72 atributos, com a da Figura 7.8(d), gerada com apenas 30 atributos, percebe-se que esta última preserva satisfatoriamente a estrutura dos dados, sendo capaz de revelar claramente o comportamento dos quatro grupos distintos.

Figura 7.8 – a) Visualizações por Coordenadas Paralelas do conjunto de dados *Mammals2.data* (72 atributos): a) adotando o mesmo arranjo dos atributos fornecido nos dados; b) adotando o arranjo por similaridade dos 72 atributos; c) com apenas 40 atributos obtidos com o *SBAR* (usando a Equação 7.2); d) idem, para 30 atributos.

A Figura 7.9 ilustra os resultados obtidos com o arranjo e a redução de atributos na visualização do conjunto de dados *Synth_Sierp.data*, explorando a técnica *IPC Density Plot*. Em (a) tem-se a visualização convencional por Coordenadas Paralelas, bastante sobrecarregada pela excessiva sobreposição das poligonais. A visualização com o *IPC Density Plot* em 7.9(b) revela melhor a estrutura dos dados; entretanto, ela não favorece a identificação de correlação entre os atributos. O *IPC Density Plot* em 7.9(c) exibe os

atributos na sequência gerada pelo *SBAA*, segundo a medida de similaridade dada pela Equação 7.2. A visualização permite identificar visualmente a correlação entre os atributos mapeados nos eixos 2 e 3 (*Corr_XY* e *X_Sierp*) e também nos eixos 4 e 5 (*Corr_Rand* e *Rand*). Em 7.9(d) tem-se a visualização obtida considerando apenas os atributos *Y_Sierp*, *Corr_XY* e *Corr_Rand*, remanescentes após reduzir a dimensionalidade de cinco para três atributos, aplicando o *SBAR* com a medida de similaridade dada pela Equação 7.2. São mantidos os atributos *Y_Sierp*, *CORR_XY* e *CORR_Rand*. O atributo *X_Sierp*, embora não tenha sido mantido, está ‘embutido’ no atributo *CORR_XY*. Da mesma forma, os atributos *Rand* e *CORR_Rand* são correlacionados e, portanto, a escolha de qualquer um deles é justificada.

Figura 7.9 – Visualizações do conjunto de dados *Synth_Sierp.data* a) Coordenadas Paralelas; b) *IPC Density Plot*; c) *IPC Density Plot* após arranjo dos atributos usando o *SBAA*; d) *IPC Density Plot* após redução de cinco para três atributos.

A Figura 7.10 ilustra o resultado obtido aplicando o *SBAA* com a medida de similaridade da Equação 7.4 (coeficiente de correlação de Pearson). O *IPC Density Plot* obtido considerando os cinco atributos na sequência *Y_Sierp*, *Corr_Rand*, *Rand*, *Corr_XY* e *X_Sierp*, exibido em 7.10(a), permite identificar, visualmente, a correlação entre os pares

de atributos mapeados nos eixos 2 e 3 e nos eixos 4 e 5. Em 7.10(b) tem-se a visualização gerada com os três atributos remanescentes após o processo de redução, usando $mSBAR$ com a medida de similaridade da Equação 7.4.

Figura 7.10 – a) Visualização do conjunto de dados *Synth_Sierp.data* com o *IPC Density Plot*, adotando a sequência dos atributos obtida com o *SBAA* usando o coeficiente de correlação de Pearson como medida de similaridade (Equação 7.4; b) Visualização considerando apenas os três atributos obtidos com o $mSBAR$.

A exploração dos mesmos dados no *Viz3D* é ilustrada na Figura 7.11. Em (a) tem-se a visualização inicial, com os atributos dispostos segundo a sequência dada no conjunto de dados. Em (b) tem-se a mesma visualização após uma rotação que revela melhor a organização espacial dos dados. Em 7.11(c) tem-se a visualização gerada considerando apenas os atributos *Corr_Rand*, *Y_Sierp* e *Corr_XY*, nessa sequência, obtida aplicando o *SBAR* com a medida de similaridade dada pela Equação 7.2. Em (d) tem-se a visualização considerando os atributos *Rand*, *X_Sierp* e *Y_Sierp*, nessa sequência, obtidos com o $mSBAR$, adotando a medida de similaridade dada pela Equação 7.4. Observa-se que os atributos selecionados pelo $mSBAR$ coincidem com os obtidos por Razente [2004] aplicando o *FDR – Fractal Dimension Reduction* [Traina et al., 2000]. As visualizações apresentadas em 7.11(c) e 7.11(d) foram obtidas após algumas operações de rotação.

Figura 7.11 – Visualizações *Viz3D* do conjunto de dados *Synth_Sierp.data*: a) posicionamento inicial considerando todos os atributos; b) após operação de rotação; c) usando a sequência de atributos *Corr_Rand*, *Y_Sierp* e *Corr_XY*; d) usando a sequência de atributos *Rand*, *X_Sierp* e *Y_Sierp*.

7.7 Comentários Finais

Apesar da simplicidade dos algoritmos propostos neste capítulo, eles se mostram importantes aliados em tarefas de análise visual de agrupamentos, uma vez que em todos os casos observados, resultaram em uma melhor separação dos grupos do que as visualizações que adotam um arranjo aleatório dos atributos para a exibição. Os tempos de execução dos algoritmos propostos são baixos o suficiente para viabilizar a sua aplicação a grandes conjuntos de dados de alta dimensionalidade.

O congestionamento visual é atenuado em Coordenadas Paralelas, porque o cruzamento excessivo entre as poligonais é reduzido. No *RadViz* e *Viz3D* o posicionamento dos marcadores é melhorado, porque evita que os marcadores sejam mapeados indevidamente no centro da visualização, quando atributos com valores similares são mapeados em eixos opostos no gráfico. O uso dos algoritmos para reduzir a dimensionalidade dos dados também se mostrou eficaz, pois mesmo após a eliminação de

alguns atributos, ainda se consegue identificar as estruturas relevantes nas visualizações obtidas. Uma vantagem importante das propostas introduzidas neste capítulo sobre outras abordagens da literatura é a sua simplicidade, e a possibilidade de substituir com facilidade a medida de similaridade adotada. As medidas de similaridade merecem uma investigação adicional, particularmente no caso de atributos categóricos, que não foram tratados especificamente nesse trabalho. Além disso, a estratégia proposta trata simultaneamente os problemas de arranjar e remover atributos, o que não ocorre nas demais soluções discutidas na Seção 7.2.

Capítulo 8

Conclusões

Técnicas efetivas para a Visualização Exploratória e Mineração Visual de conjuntos de dados volumosos e de alta dimensionalidade ainda representam um enorme desafio para os pesquisadores da área. Apesar dos avanços recentes, técnicas convencionais de visualização direta – em que cada elemento do conjunto de dados é mapeado em um elemento visual na tela – ainda são bastante limitadas nesse contexto. Neste trabalho foram propostas algumas soluções para tratar estas limitações. Técnicas de visualização exploratória foram re-visitadas e enriquecidas com estratégias que permitem gerar representações visuais mais efetivas para apoiar processos interativos de exploração de grandes conjuntos de dados de alta dimensionalidade. As técnicas introduzidas apresentam baixa complexidade computacional, de modo a não prejudicar ou inviabilizar a interação do usuário com as visualizações, na busca por padrões como agrupamentos e regiões de alta densidade nos dados. Quatro contribuições distintas foram descritas:

(i) Foi apresentada uma nova técnica de visualização multidimensional, denominada *Viz3D*, que projeta os dados em três dimensões espaciais, amenizando assim o efeito indesejado da excessiva sobreposição dos marcadores, típica de estratégias bidimensionais, como o *RadViz* e *Star Coordinates*. A técnica, descrita no Capítulo 4, oferece uma alternativa de baixo custo para a projeção de dados multidimensionais no espaço 3D;

(ii) Técnicas de visualização direta, como *Coordenadas Paralelas*, *RadViz* e o próprio *Viz3D*, são enriquecidas com informações de frequência e densidade computadas a partir dos dados, as quais são mapeadas visualmente buscando reduzir o efeito prejudicial de uma excessiva sobreposição dos marcadores, atenuar os efeitos indesejados da presença de ruído nos dados e ainda realçar a presença de padrões nos dados. As técnicas resultantes da incorporação desta abordagem foram introduzidas no Capítulo 5;

(iii) Foi introduzido um procedimento interativo para apoiar a identificação de agrupamentos em grandes conjuntos de dados, o qual é apoiado nas visualizações enriquecidas com informação de frequência e densidade introduzidas anteriormente. A estratégia interativa de agrupamento visual é bastante efetiva para a identificação de agrupamentos em conjuntos de dados de alta-dimensionalidade com o apoio do usuário, como discutido e ilustrado no Capítulo 6;

(iv) Finalmente, no Capítulo 7 foi apresentada uma estratégia simples para configurar a disposição dos atributos de dados mapeados na visualização (tipicamente, para atributos mapeados em eixos da visualização), a qual resulta em visualizações de melhor qualidade e mais informativas no caso de dados de alta dimensionalidade.

A técnica *Viz3D*, apresentada no Capítulo 4, projeta dados multidimensionais em um espaço tridimensional, e oferece uma alternativa interessante às técnicas convencionais de visualização e de redução de dimensionalidade, devido à sua simplicidade e baixo custo computacional. As projeções geradas no *Viz3D* têm qualidade semelhante às obtidas com estratégias como a *PCA* (Análise de Componentes Principais) seguida de exibição no espaço 3D e o *FastMap*. Entretanto, o cálculo das projeções no *Viz3D* demanda tempos de processamento significativamente menores (particularmente em relação à *PCA*), o que torna a técnica conveniente para a manipulação interativa de grandes conjuntos de dados multidimensionais. Em comparação com o *RadViz*, que inspirou a abordagem, o *Viz3D* minimiza deficiências como o congestionamento na área central da visualização, causado pela alta dimensionalidade e pelo processo de projeção, e as oclusões decorrentes do grande número de marcadores, que dificultam a análise. Analogamente ao *RadViz*, o *Viz3D* garante que registros próximos entre si no espaço original sejam projetados próximos no espaço de saída. Entretanto, registros distantes entre si no espaço original podem ser indevidamente projetados em posições próximas no espaço 3D. De fato, qualquer que seja a estratégia de projeção adotada, sempre haverá alguma perda de informação decorrente da projeção, sendo impossível preservar completamente a estrutura dos dados no espaço original. Ainda assim, visualizações de vários conjuntos de dados de alta dimensionalidade geradas com o *Viz3D* revelaram-se bastante eficazes para identificar estruturas de interesse. Adicionalmente, como os agrupamentos no espaço de entrada original são preservados na projeção, a técnica é eficaz para apoiar processos interativos de agrupamento visual, como apresentado no Capítulo 6.

As técnicas de visualização enriquecidas com informação de frequência e densidade introduzidas no Capítulo 5 – denominadas *Viz3D Density Plot*, *IPC Frequency Plot* e *IPC Density Plot* – reduzem significativamente o efeito da sobreposição dos marcadores em visualizações sobrecarregadas, conseguindo realçar estruturas que não poderiam ser observadas utilizando técnicas diretas convencionais. Os algoritmos para computar frequência e densidade são de baixa complexidade computacional, o que viabiliza a sua aplicação a grandes conjuntos de dados multidimensionais, em tempos de processamento aceitáveis. No caso do *Viz3D*, realizar o cálculo da informação de frequência e densidade no espaço tridimensional da projeção representa uma simplificação que reduz consideravelmente o processamento necessário. Seria impraticável calcular essas informações nos espaços multidimensionais originais, em vista da necessidade de manipular muitos registros e de exibir o resultado em tempos adequados para proporcionar uma interação satisfatória ao usuário. Da mesma forma, no caso dos *IPC Frequency Plot* e *IPC Density Plot*, o cálculo da informação de frequência e densidade nos espaços bidimensionais definidos pelos pares de eixos paralelos consecutivos também é uma simplificação para reduzir o processamento, mas não implica em perda de qualidade para a visualização. A opção por calcular a informação de frequência e densidade em espaços discretos, especificados pelo usuário – uma vez que determinados pela resolução dos gráficos a serem gerados – não compromete a qualidade das visualizações porque considera a sobreposição dos marcadores para a exibição na resolução corrente. O usuário pode alterar a resolução para obter visualizações com diferentes níveis de detalhamento. De fato, mecanismos de interação, como o *zoom* no *Viz3D* e no *Viz3D Density Plot*, são importantes para apoiar processos exploratórios. Nos *IPC Frequency* e *Density Plots*, o usuário também pode variar a resolução dos gráficos, computando a informação de frequência em diferentes níveis de resolução. Foram descritos vários recursos de interação associados a essas técnicas, que podem ser muito úteis para a exploração de conjuntos de dados volumosos. Em particular, foram propostas operações simples de limiarização que permitem filtrar registros de interesse com base em informação de frequência.

A exploração visual de dados por meio da interação do usuário com visualizações geradas pelas técnicas *Viz3D Density Plot*, *IPC Frequency* e *IPC Density Plots*, abordada no Capítulo 6, dá controle ao usuário e insere o potencial humano de perceber padrões no processo de identificação de agrupamentos. O processo de agrupamento é controlado pelo usuário por meio da identificação, utilizando o mouse, de regiões de alta densidade na

visualização. A identificação de agrupamentos relevantes nas visualizações por *IPC Frequency* e *IPC Density Plots* pode ser executada satisfatoriamente, mesmo sem grande precisão nas seleções efetuadas pelo usuário sobre cada eixo (dimensão), pois a determinação consecutiva de intervalos de interesse sobre vários eixos (ainda que com baixa precisão) é capaz de caracterizar os grupos. Adicionalmente, os registros identificados em uma seleção inicial podem ser novamente exibidos e a marcação da seleção pode ser refinada com a mesma abordagem. Da mesma forma, o processo de marcação de regiões no *Viz3D Density Plot* pode envolver várias marcações consecutivas, intercaladas com operações de rotação e re-posicionamentos, aumentando progressivamente a precisão da seleção. Ainda assim, é importante dispor de mecanismos para refinar a identificação de grupos, por exemplo, analisando informações complementares como estatísticas ou métricas descritivas dos grupos obtidos. A estratégia de utilizar técnicas de visualização diferentes, apoiando-se mutuamente, também mostra-se útil para revelar registros inseridos indevidamente nos grupos.

Os algoritmos para definir uma sequência de exibição de atributos e reduzir a dimensionalidade das visualizações, introduzidos no Capítulo 7 – *SBAA*, *SBAR* e *SBAR_m* mostraram-se grandes aliados em tarefas de exploração visual. Estes algoritmos se destacam de estratégias similares apresentadas na literatura por sua simplicidade e porque conseguem tratar, simultaneamente, os dois problemas de definir uma sequência de exibição e sugerir atributos candidatos a serem removidos da visualização. Os experimentos que ilustram a aplicação destas estratégias apontaram uma redução no congestionamento visual em Coordenadas Paralelas, reduzindo o cruzamento entre as linhas poligonais. Da mesma forma, resulta em uma melhora no posicionamento dos marcadores nas técnicas *RadViz* e *Viz3D*, evitando a concentração de marcadores na região central, que ocorre quando atributos correlacionados são mapeados em eixos opostos no gráfico. A redução da dimensionalidade da visualização por meio destes algoritmos também se mostrou eficaz, pois analisando resultados obtidos com vários conjuntos de dados verificou-se que a estrutura dos dados é satisfatoriamente preservada nas visualizações resultantes. Outra vantagem importante dos algoritmos propostos sobre outras abordagens é a sua flexibilidade. É possível, por exemplo, substituir facilmente a medida de similaridade adotada. Diferentes medidas de similaridade merecem investigação adicional, particularmente no caso de atributos categóricos, que não foram tratados especificamente nesse trabalho.

Na medida do possível, as propostas apresentadas foram comparadas a trabalhos relacionados existentes na literatura, e avaliadas por meio de sua aplicação a diversos conjuntos de dados reais e sintéticos, de vários tamanhos e dimensionalidades. A inexistência de conjuntos de dados contendo diferentes padrões de interesse conhecidos, que pudessem ser utilizados como *'benchmarks'*, dificultou a realização de comparações mais precisas e extensivas. Adicionalmente, implementações das técnicas encontradas na literatura nem sempre estão disponíveis, dificultando a realização de novos experimentos com outros conjuntos de dados. Todas as estratégias apresentadas neste trabalho foram implementadas em um ambiente denominado *MDV (Multidimensional Visualization)*, o qual está disponível para *download*¹, assim como os conjuntos de dados utilizados para validar as propostas apresentadas. Os resultados indicam que, apesar das estratégias propostas não eliminarem os problemas associados à oclusão e congestionamento visual na visualização de grandes conjuntos de dados, elas oferecem soluções efetivas e de baixo custo que ampliam a escalabilidade visual de técnicas tradicionais, viabilizando tarefas de exploração e mineração que seriam impossíveis, ou muito dificultadas, sem esses recursos.

Um aspecto que merece investigações adicionais é a proposta de novos mecanismos de interação com as técnicas de visualização introduzidas, bem como o refinamento das interações já disponíveis. Por exemplo, seria interessante estudar se o processo iterativo de identificação de agrupamentos pode ser facilitado com interações que permitem uma delimitação mais precisa das regiões de interesse que especificam uma seleção. Em particular, a Ligação (*Linking*) entre múltiplas visualizações também pode ser útil. Nos exemplos descritos neste trabalho, a apresentação por uma outra técnica de registros selecionados em uma determinada visualização é feita salvando os registros selecionados em um arquivo, e exibindo o conteúdo deste arquivo na segunda visualização, o que dificulta a tarefa de exploração. O efeito da manipulação de diferentes parâmetros, como resolução e largura do filtro utilizados no cálculo de densidade, no resultado das visualizações e no processo de identificação de agrupamentos pode ser investigado mais detalhadamente sob a perspectiva dos usuários das técnicas.

Para validar o processo iterativo de agrupamento baseado nas visualizações por frequência e densidade, bem como ampliar a confiabilidade de potenciais usuários nesse processo, seria importante comparar resultados desse processo de agrupamento direcionado pelo usuário com resultados obtidos por outras técnicas de mineração de agrupamentos.

¹ www.unoeste.br/~almira/research/visualization

com ou sem suporte visual. Mecanismos para avaliar resultados de agrupamentos são abordados em Halkidi, et al., [2000; 2001], e poderiam oferecer um ponto de partida. Uma dificuldade é, justamente, o acesso a diferentes implementações de diferentes técnicas, bem como a conjuntos de dados representativos – particularmente em se tratando de dados reais.

A estratégia para gerar uma disposição adequada dos atributos e reduzir a dimensionalidade da visualização poderia ser estendida para incorporar diferentes métricas de similaridade – sendo que diferentes métricas podem ser adequadas a diferentes tipos de dados. Dados categóricos, por exemplo, não mereceram qualquer tratamento especial e ocorrem hoje com muita frequência. A heurística utilizada para definir o *layout* dos atributos é muito simples, e estratégias mais sofisticadas poderiam ser adotadas, por exemplo, a técnica poderia examinar o impacto que a incorporação de um certo atributo na sequência de atributos mais similares vai causar no passo seguinte, quando for anexado o próximo atributo na sequência. De fato, várias heurísticas propostas para resolver o problema do caixeiro viajante poderiam ser investigadas nesse contexto.

A experiência resultante do *Viz3D* pode direcionar outras iniciativas no sentido de estender outras técnicas bidimensionais para o espaço tridimensional. A vantagem do espaço tridimensional é a possibilidade de acomodar um maior número de marcadores, permitindo ainda recursos simples de interação simples aos quais usuários típicos estão bastante acostumados, devido à experiência com jogos ou ambientes gráficos 3D. Informações de frequência e densidade, calculadas no espaço discreto das visualizações também podem ser acrescentadas para melhorar a qualidade visual das visualizações.

Referências Bibliográficas

- AGGARWAL, C., A Human-Computer Cooperative System for Effective High Dimensional Clustering, *Proceedings ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pp. 221-226, 2001.
- AGRAWAL, R., SRIKANT, R., Fast Algorithms For Mining Association Rules, *Proceedings 20th Int. Conf. Very Large Data Bases (VLDB)*, Morgan Kaufman, Editor: Bocca, J.B., Jarke, M., Zaniolo, C., pp. 487-499, 1994.
- ANDERSON, T.W., *An Introduction to Multivariate Statistical Analysis*, Wiley, 1958.
- ANKERST M., KEIM D.A., KRIEGEL H.P.: Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets, *Proc. Visualization '96*, San Francisco, Ca, 1996.
- ANKERST, M., BERCHTOLD, S., KEIM, D.A. Similarity Clustering of Dimensions for Enhanced Visualization of Multidimensional Data, *Proc. IEEE Information Visualization 1998*, pp. 52-60, 1998.
- ANKERST M., Visual Data Mining, *Ph.D. Thesis*, University of Munich, 2000.
- ANKERST, M., Visual Data Mining with Pixel-Oriented Visualization Techniques, *Url: CiteSeer.Nj.Nec.Com/Ank2001.Html*, 2001.
- ANKERST, M., BREUNIG, M., KRIEGEL, H.P., SANDER, J. *OPTICS: Ordering Points To Identify the Clustering Structure*. ACM SIGMOD Conference, 1999.
- ARTERO, A.O., OLIVEIRA, M.C.F., LEVKOWITZ, H., Uncovering Clusters in Crowded Parallel Coordinates Visualizations, *Proc. IEEE Symposium on Information Visualization (InfoVis2004)*, pp. 81-88, 2004.
- ARTERO, A.O., OLIVEIRA, M.C.F., *Viz3D: Effective Exploratory Visualization of Large Multidimensional Data Sets*, *Proc. XVII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI2004)*, pp. 340-347, 2004.
- ASIMOV, D., The Grand Tour: A Tool For Viewing Multidimensional Data, *Siam J. Science & Statistical Computing*, Vol. 6, pp. 128 -143, 1985.
- BELTRAMI, E., Sulle Funzioni Bilineari [on bilinear function], *Giornale di Matematiche ad Uso degli Studenti delle Università*, 11, pp.98-106, 1873.
- BERTIN, J., *Graphics and Graphic Information Processing*, De Gruyter Publishers, 1981.
- BERKHIN, P., Survey Of Clustering Data Mining Techniques, Accrue Software, San Jose, CA, *http://citeseer.ist.psu.edu/berkhin02survey.html*, 2002.
- BESHIERS, C., FEINER, S., Autovisual: Rule-Based Design Of Interactive Multivariate Visualizations, *IEEE Computer Graphics & Applications*, pp. 41-49, July, 1993.
- BOLTON, R. J., KRZANOWSKI, W. J., A Characterization of Principal Components for Projection Pursuit, *Tas The American Statistician*, *American Statistical Association Amstat Online*, Vol. 53, N. 2, 1999.
- BRUNSDON, C., FOTHERINGHAM, A. S., CHARLTON, M. E., An Investigation of Methods For Visualizing Highly Multivariate Datasets, In *Report Case Studies of Visualization In The Social Sciences* Edited By Unwin, D, Fisher, P., 1999.
- BUJA, A., COOK, D, SWAYNE, F., Interactive High-Dimensional Data Visualization, *Journal Of Computational And Graphical Statistics*, vol. 5, pp. 78-99, 1996.

- CABENA, P., HADJINIAN, P., STADLER, R., VERHEES, J., ZANASI, A. *Discovering Data Mining – From Concept to Implementation*, Prentice Hall, 1998.
- CARD, S. K., MACKINLAY, J. D., SHNEIDERMAN, B., *Readings in Information Visualization – Using Vision to Think*, Morgan Kaufmann, p. 712, 1999.
- CHEN, M. S.; HAN, J.; YU, P. S. Data Mining: An Overview from a Database Perspective, *IEEE Trans. Knowledge & Data Engineering*, vol. 8, n. 6, pp. 866-883, 1996.
- CHERNOFF, H., The use of Faces to Represent Points in K-Dimensional Space Graphically, *Journal of American Statistical Association*, vol. 68, pp. 361-368, 1973.
- CHOU, S. Y., LIN, S. W., YEH, C. S., Cluster Identification with Parallel Coordinates, *Patterns Recognition Letters*, vol. 20, pp. 565-572, 1999.
- CLEVELAND, W. S., *Visualizing Data*, Hobart Press, Summit, 1993.
- COOK, D., BUJA, A., CABRERA, J., HURLEY, C., Grand Tour and Projection Pursuit, *Journal of Computational and Graphical Statistics*, vol. 4, pp. 155-172, 1995.
- DEMBÉLÉ, D., KASTNER, P., Fuzzy C-means method for clustering microarray data, *Bioinformatics*, vol.19, n.8, pp.973-980, 2003.
- DORIGO, M., GAMBARDELLA, L. M., Ant Colony System: A Cooperative Learning Approach to Traveling Salesman Problem, *IEEE Transaction on Evolutionary Computation*, vol. 1, n. 1, pp. 53-66, 1997.
- EICK, S. G., Data Visualization Sliders, *Proceedings UIST'94*, pp. 119-120, 1994.
- EICK, S. G., STEFFEN, J. L., SUNNER, E. E., Seesoft – A Tool for Visualizing Line Oriented Statistics Software, *IEEE Transactions On Software Engineering*, vol. 18, n. 11, pp. 957-968, 1992.
- ENEVA, E., KUMARASWAMY, K., MATTEUCCI, M., WEKKEM: A Study in Fractal Dimension and Dimensionality Reduction. In *Workshop on Fractals and Self-similarity in Data Mining: Issues and Approaches (KDD-2002)*, 2002.
- ESTER, M., KRIEGEL, H. P., SANDER, J., XU, X., A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp.226-231, 1996.
- FALOUTSOS, C., LIN, K. L., Fastmap: A Fast Algorithm For Indexing, Data Mining and Visualization of Traditional and Multimedia Datasets, *ACM SIGMOD Proceedings*, pp. 163-174, 1995.
- FASULO, D., An Analysis of Recent Work on Clustering Algorithms, *Technical report UW-CSE01-03-02*, University of Washington, 1999.
- FAYYAD, U., SHAPIRO, P.G., SMYTH, P., *From Data Mining to Knowledge Discovery: An Overview*, In Fayyad, U., Piatetsky-Shapiro, G., Amith, Smyth, P., and Uthurusamy, R. (eds.), *Advances in Knowledge Discovery and Data Mining*, MIT Press, pp. 1-36, Cambridge, 1996.
- FEINER, S., BESHES, C., Worlds Within Worlds: Metaphors for Exploring N-Dimensional Virtual Worlds, *Proceedings UIST'90*, pp. 76-83, 1990.
- FISHKIN, K., STONE, M. C., Enhanced Dynamic Queries Via Movable Filters, *Proc. Human Factors in Computing Systems Chi'95 Conf.*, pp.415-420, 1995.
- FODOR I. K., A Survey of Dimension Reduction Techniques, *LLNL-Lawrence Livermore National Laboratory*, Technical Report, 2002.
- FRIEDMAN, J. H., Exploratory projection pursuit, *Journal of the American Statistical Association*, vol. 82, pp. 249-266, 1987.
- FRIEDMAN, J. H., TUKEY, J. W., A Projection Pursuit Algorithm for Exploratory Data Analysis, *IEEE Transactions on Computers*, vol. 23, pp. 881-890, 1974.
- FUA, Y. H., WARD, M. O., RUNDENSTEINER, A. Hierarchical Parallel Coordinates for Exploration of Large Datasets, *Proc. IEEE Visualization' 99*, pp.43-50, 1999.

- FUA, Y. H., WARD, M. O., RUNDENSTEINER, E. A., Hierarchical Parallel Coordinates for Visualizing Large Multivariate Data Sets, *IEEE Conf. on Visualization '99*, pp. 43-50, 1999.
- GAMERO, E. J. T., MINGHIM, R., Improved Visual Clustering of Large Multi-Dimensional Data Sets. Institute of Mathematics and Computer Science – USP. *ICMC's Technical Reports*, Number 215, 2003.
- GENNARI, J. H., LANGLEY, P., FISHER, D., *Models of incremental concept formation*, Artificial Intelligence 40-1(3), pp.11-61, 1989.
- GRINSTEIN, G. G., HOFFMAN, P. E., PICKETT, R. M., Benchmark Development for the Evaluation of Visualization for Data Mining, in: *Information Visualization in Data Mining and Knowledge Discovery*, edited by: FAYYAD, U., GRINSTEIN, G. G., WIERESE, A., Morgan Kaufmann Publishers, 2001.
- GUHA, S., RASTOGI, R., SHIM, K., Cure: An Efficient Clustering Algorithm For Large Databases, *SIGMOD*, vol. 27, n. 2, pp. 73-84, 1998.
- HALKIDI, M., VAZIRGIANNIS, M., AND BATISTAKIS, I., Quality Scheme Assessment in the Clustering Process, In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 265 – 276, 2000.
- HALKIDI, M., VAZIRGIANNIS, M., BATISTAKIS, I., Clustering Validation Techniques, *Intelligent Information Systems J.*, vol. 17, n.2-3, pp. 107-145, 2001.
- HAN, J., CERCONE, N., Ruleviz: A Model for Visualizing Knowledge Discovery Process, *ACM Conf. on Knowledge Discovery in Data, Proc. Sixth ACM SIGKDD, Int. Conf. on Knowledge Discovery and Data Mining, Kdd2000*, pp. 244-253, 2000.
- HAN, J., KAMBER, M., *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers 550p, 2000.
- HAND, D. J., MANNILA, H., SMYTH, P., *Principles of Data Mining*, MIT Press, 425p, 2001.
- HINNEBURG, A., KEIM, D. A., An Efficient Approach To Clustering In Large Multimedia Databases With Noise, *Proceedings 4th Int. Conf. On Knowledge Discovery And Data Mining*, AAAI Press, pp. 58-65, 1998.
- HINNEBURG, A., KEIM, D. A., WAWRYNIUK, M., *HD-Eye: Visual Mining of High-Dimensional Data*, *IEEE Computer Graphics and Applications*, pp. 22-31, 1999.
- HINNEBURG, A., KEIM, D. A., Optimal Grid-Clustering: Towards Breaking the Curse of Dimensionality in High-Dimensional Clustering, *Proc. 25th International Conference on Very Large Databases*, pp. 506-517, 1999.
- HOFFMAN, P. E., Table Visualizations: A Formal Model and its Applications. *Doctoral Diss.*, Computer Science Department, University Of Massachusetts, Lowell, Ma, 1999.
- HOTELLING, H., Analysis of a Complex of Statistical Variables into Principal Components, *Journal of Educational Psychology*, 24, pp. 417-441; 498-500-520, 1933.
- INSELBERG, A. The Plane with Parallel Coordinates, *The Visual Computer (Special Issue on Computational Geometry)*, vol. 1, n. 2, pp. 69-92, 1985.
- INSELBERG, A., Multidimensional Detective, *Proc. IEEE Symposium on Information Visualization (InfoVis' 97)*, Phoenix, AZ, pp. 100-107, 1997.
- INSELBERG, A., DIMSDALE, B., Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry, *Proc. IEEE Visualization '90*, pp. 361-375, 1990.
- JAIN, A. K., MURTY, M. N., FLYNN, P. J., Data Clustering: A Review, *ACM Computing Surveys*, vol. 31, n. 3, September, 1999.
- KANDOGAN, E. Visualizing Multi-Dimensional Clusters, Trends, and Outliers using Star Coordinates, *Proc. ACM Int. Conf. Knowledge Discovery and Data Mining*, pp.107-116, 2001.

- KEIM, D. A., Visual Database Exploration Techniques, *Proc. Tutorial KDD'97 Int. Conf. on Knowledge Discovery and Data Mining*, Newport Beach, CA, <http://www.informatik.uni-halle.de/~keim/PS/KDD97.pdf>, 1997.
- KEIM, D. A., Designing Pixel-Oriented Visualization Techniques: Theory And Applications, *IEEE Transactions on Visualization and Computers Graphics*, vol. 6, n.1, pp.59-78, 2000.
- KEIM, D. A., Visual Exploration of Large Databases, *Communication ACM*, vol.44, n.8, pp. 48-44, 2001.
- KEIM, D. A., Information Visualization and Visual Data Mining, *IEEE Transactions on Visualization And Computers Graphics*, vol. 8, n.1, pp. 1-8, 2002.
- KEIM, D. A., KRIEGEL, H. P., VisDB: Database Exploration using Multidimensional Visualization, *IEEE Computer Graphics and Applications*, vol. 14, n. 5, September, pp. 40-49, 1994.
- KEIM, D. A., KRIEGEL, H. P., Visualization Techniques for Mining Large Databases: A Comparison, *IEEE Trans. Knowledge & Data Engineering*, vol. 8, n. 6, pp. 923-936, 1996.
- KOHAVI, R., BRODLEY, C., *Knowledge Discovery & Data Mining Cup 2000*, Data Provided By Blue Martini Software And Gazelle.Com., <http://www.ecn.purdue.edu/kddcup/>, 2000.
- KRUSKAL, J. B., Nonmetric Multidimensional Scaling: A Numerical Method, *Psychometrika*, vol. 29, pp.115-129, 1964.
- KRUSKAL, J. B., WISH, M., *Multidimensional Scaling*, SAGE publications, 1978.
- LAMPING, J., RAO, R., PIROLLI, P., A Focus + Context Technique Based on Hyperbolic Geometry For Visualizing Hierarchies, *Proc. Human Factors in Computing Systems CHI 95 Conf.*, pp. 401-408, 1995.
- LEBLANC, J., WARD, M. O., WITTELS, N., Exploring N-Dimensional Databases, *Proc. IEEE Visualization'90, IEEE CS Press*, pp. 230-237, 1990.
- LEUNG, Y. K., APPERLEY, M. D., A Review and Taxonomy of Distortion-Orientation Presentation Techniques, *ACM Transactions on Computer-Human Interaction*, 1(2), pp. 126-160, 1994.
- MACKINLAY, J. D., ROBERTSON, G. G., CARD, S. K., The Perspective Wall: Detail and Context Smoothly Integrated, *Proceedings Human Factors in Computing Systems CHI 91, Conf.*, pp. 173-179, 1991.
- MANCO, G., PIZZUTI, C., TALIA, O.D., Eureka!: An Interactive and Visual Knowledge Discovery Tool, *Journal of Visual Languages and Computing*, Elsevier, vol. 15, n. 1, pp. 1-35, 2004.
- MANLY, B. F. J., *Multivariate Statistical Methods*, Chapman & Hall, 1986.
- MEHTA, M., AGRAWAL, R., RISSANEN, J., SLIQ: A Fast Scalable Classifier for Data Mining, In *Proc. 1996 International Conference on Extending Database Technology (EDBT'96)*, 1996.
- MILLER, J. J., WEGMAN, E. J., Construction of Line Densities for Parallel Coordinate Plots, *Computational Statistics and Graphics*, eds. A. Buja, P. Tukey, Springer-Verlag, pp. 107-123, 1990.
- MOLINA, L. C., BELANCHE, L., NEBOT, A., Feature Selection Algorithms: An Survey and Experimental Evaluation, *IEEE International Conference on Data Mining*, vol. 1, pp. 306-313, 2002.
- MUNZNER, T., BURCHARD, P., Visualizing the Structure of the World Wide Web in 3D Hyperbolic Space, *Proc. Virtual Reality Modeling Language (VRML'95) Symp.*, pp. 33-38, 1995.
- NASON, G., Three-Dimensional Projection Pursuit, *Applied Statistics*, 44, pp. 411-430, 1995.
- NG, R., HAN, J., Efficient and Effective Clustering Method For Spatial Data Mining, In *Proc. 1994 International Conference On Very Large Data Bases*, pp. 144-155, 1994.

- OLIVEIRA, M. C. F., LEVKOWITZ, H., From Visualization to Visual Data Mining: A Survey, *IEEE Transactions on Visualization and Computer Graphics*, vol. 9, n. 3, pp. 378-394, 2003.
- PARK, J. S., CHEN, M. S., YU, P. S., An Effective Hash Based Algorithm for Mining Association Rules, *Proceedings of ACM SIGMOD*, pp. 175-186, 1995.
- PARSAYE, K., CHIGNELL, M., *Intelligent Database Tools and Applications*, John Wiley & Sons, 1993.
- PEARSON, K., On Lines and Planes of Closest Fit to System of Points in Space, *Philosophy Magazine*, vol. 6, pp. 559-572, 1901.
- PENG, W., WARD, M. O., RUNDESTAINER, E. A., Clutter Reduction in Multi-dimensional Data Visualization Using Dimension Reordering, *IEEE Symposium on Information Visualization*, pp. 89-96, 2004.
- PICKETT, R. M., GRINSTEIN, G. G., Iconographic Displays for Visualizing Multidimensional Data, *Proc. IEEE Conf. Systems, Man and Cybernetics, IEEE Press*, pp. 514-519, 1988.
- PIROLI, P., RAO, R., TableLens as a Tool for Making Sense of Data, In Catarci, T., Costabile, M.F., Levialdi, S., Santucci, G (Eds.), *Workshop On Advanced Visual Interfaces: AVI-96*, pp. 67-80, 1996.
- QUINLAN, J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.
- RAO, R., CARD, S. K., The TableLens: Merging Graphical and Symbolic Representation in a Interactive Focus+Context Visualization for Tabular Information, *Proc. Human Factors in Computing Systems CHI 94 Conf.*, pp. 318-322, 1994.
- RAZENTE, H. L., Análise Visual em Processos de Redução de Dimensionalidade para Mineração em Sistemas de Bases de Dados, *Dissertação de Mestrado*, ICMC-USP, 85p, 2004.
- RHYNE, T. M., TREINISH, L., Data Signatures and Visualization of Scientific Data Sets, *IEEE Computer Graphics and Applications*, pp. 12-15, Mar/Apr, 2000.
- RODRIGUES JR, J. F., TRAINA, A. J., TRAINA JR, C., Frequency Plot and Relevance Plot to Enhance Visual Data Exploration, *Proc. XVI Brazilian Symp. on Computer Graphics and Image Processing*, pp. 117-124, 2003.
- SAMMON, J. W., A Nonlinear Mapping for Data Structure Analysis, *IEEE Transaction on Computers*, vol. C-18, n. 5, 1969.
- SCOTT, D. W., Averaged Shifted Histograms: Effective Nonparametric Density Estimation in Several Dimensions, *The Annals of Statistics*, vol. 13, n. 3, pp. 1024-1040, 1985.
- SHEPARD, R., The Analysis of Proximities: Multi Dimensional Scaling with an Unknown Distance Function, *Psychometrika*, vol. 27, pp. 125-140, 219-246, 1962.
- SHNEIDERMAN, B., Tree Visualization with *Treemaps*: A 2D Space-Filling Approach, *ACM Trans. Graphics*, vol. 11(1), pp. 92-99, 1992.
- SHNEIDERMAN, B., Dynamic Queries for Visual Information Seeking, *IEEE Software*, vol 11, n. 6, pp. 70-77, 1994.
- SHIRTOLA, H., Direct Manipulation of Parallel Coordinates, *Proceedings of the International Conference On Information Visualization (IV 2000)*, 2000.
- SILVERMAN, B. W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, 176p, 1990.
- SOBOL, M. G., KLEIN, G., New Graphics As Computerized Displays For Human Information Processing, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, n. 4, 1989.
- STOLLE, C., TANG, D., HANRAHAN, P., Polaris: A System for Query, Analysis and Visualization of Multidimensional Relational Databases, *IEEE Transactions on Visualization and Computers Graphics*, vol. 8, n.1, pp. 52-65, 2002.

- STROE, I. D., RUNDENSTEINER, E. A., WARD, M. O., Scalable Visual Hierarchy Exploration, *Technical Report Tr-00-06*, Worcester Polytechnic Institute, Computer Science Department, 2000.
- TOU, J. T., GONZALEZ, R. C., *Pattern Recognition Principles*, Addison-Wesley, Reading, 1974.
- TRAINA JR., C., TRAINA, A. M. J., WU, L., FAIOUTSOS, C., Fast feature selection using fractal dimension, in *XV Brazilian Database Symposium*, pp. 158-171, 2000.
- TUFTE, E., *The Visual Display of Quantitative Information*, Graphics Press, USA, 1983.
- YANG, J., WARD, M. O., RUNDENSTEINER, E. A., Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets, *ACM International Conference Proceeding Series Proceedings of the symposium on Data Visualisation 2003*, pp. 19-28, 2003.
- YANG, J., PENG, W., WARD, M. O., RUNDENSTEINER, E. A., Interactive Hierarchical Dimension Ordering, Spacing and Filtering for Exploration of High Dimensional Datasets, *IEEE Information Visualization (InfoVis2003)*, pp. 105-112, 2003.
- YANG, J., WARD, M. O., RUNDENSTEINER, E. A., HUANG, S., Visual Hierarchical Dimension Reduction for Exploration of High Dimensional Datasets, *VisSym 2003*.
- WARD, M. O., XmdvTool: Integrating Multiple Methods for Visualizing Multivariate Data, *Proceeding IEEE Visualization '94*, pp.326-336, 1994.
- WEGMAN, E. J., Hyperdimensional Data Analysis using Parallel Coordinates, *JASA- Journal of the American Statistical Association*, vol. 85, n. 411, pp. 664-675, 1990.
- WEGMAN, E. J., Huge Data Sets and the Frontiers of Computational Feasibility, *Journal Of Computational Statistics*, vol. 4, n. 4, pp. 281-295, 1995.
- WEGMAN, E. J., LUO, Q., High Dimensional Clustering using Parallel Coordinates And The Grand Tour. Conf. German Classification Society, Freiburg, Germany, [url: Citeseer.Nj.Nec.Com/Wegman96high.html](http://citeseer.nj.nec.com/Wegman96high.html), 1996.
- WITTEN, I. H., FRANK, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java implementations*, Morgan Kaufmann, San Francisco, CA, 2000.
- WONG, P. C., BERGERON, R. D., 30 Years Of Multidimensional Multivariate Visualization, In Nielson, G. M., Hagen, H., Muller, H., (Eds): *Scientific Visualization Overviews, Methodologies And Techniques*, *IEEE Computer Society Press*, Los Alamitos, Ca, pp. 3-33, 1997.
- ZHANG, T., RAMAKRISHNAN, R., LIVNY, M., Birch: An Introductory Data Clustering Method for Very Large Datasets, *SIGMOD*, vol. 25, n. 2, pp. 103-114, 1996.