Learning and Nonlinear Models (L&NLM) – Journal of the Brazilian Neural Network Society, Vol. 9, Iss.3, pp. 168-184, 2011.

Mineração de Dados: Uma Introdução

Luis Cláudius Coradine
Roberta Vilhena Vieira Lopes
Andrilene Ferreira Maciel

Universidade Federal de Alagoas - UFAL
Instituto de Computação
Rod. BR 104 N, km 96 - Campus A. C. Simões
CEP 57057-800, Maceió, AL, Brasil
e-mails: lccoral@gmail.com, rv2l@hotmail.com, andrilene.ferreira@gmail.com

Resumo – Mineração de dados pode ser visto como um conjunto de métodos para fazer inferências a partir de dados. Existe uma grande sobreposição entre os objetivos da mineração de dados, e os da estatística multivariada. Contudo, existem algumas diferenças filosóficas e metodológicas importantes. Este breve tutorial busca abordar os principais métodos de análise multivariada e de mineração de dados com critérios de classificação. Alem disso, serão abordados os principais métodos de análise de agrupamentos e de projeção, suas características e aplicações, no escopo de uma revisão da literatura.

Palavras chaves – mineração de dados, *data mining*, descoberta de conhecimento em base de dados, *knowledge discovery in Databases*, KDD, aprendizagem de máquinas, *machine learning*.

1 Introdução

Na década de 70, muitos especialistas foram instruídos a armazenar seus dados em quaisquer recursos tecnológicos (discos rígidos, fitas magnéticas, banco de dados, etc...), que fornecessem segurança. A evolução da tecnologia, o surgimento de novos métodos de armazenamento de dados e a popularização dos Sistemas de Gerenciamento de Banco de Dados (SGBD) como recursos da tecnologia da informação (TI), favoreceram à proliferação da informação. O surgimento dos sistemas de apoio a decisão (SAD) na década de 80 e a necessidade de reduzir o impacto das integrações entre sistemas de diversas plataformas, tanto no que se refere ao custo com a tecnologia da informação quanto ao aumento da velocidade de processamento dos sistemas de informação (SI), fizeram com que novas tecnologias fossem adotadas para esse armazenamento (Inmon, 1997).

Os sistemas de informações construídos para apoiar o processo decisório geralmente armazenam seus dados em sistemas de banco de dados ou até mesmo em grandes repositórios de dados, *data warehouse*. A idéia de se criar um banco de dados (BD) para armazenar os registros do sistema, fez com que o tamanho desses bancos crescesse rapidamente. A tecnologia *data warehouse* permite atender sistemas de informação capazes de produzir transações de alto desempenho com objetivo de armazenar e cruzar grande volume de dados (Inmon, 1997; Kimball e Caserta, 2004).

Segundo Inmon (Inmon, 1997), "um *data warehouse* consiste de um banco de dados especializado capaz de manipular um grande volume de informações obtidas a partir de bancos de dados operacionais e de fontes de dados externas à organização".

Apenas, parte da informação armazenada é transformada em conhecimento, isso quando não é quase que totalmente esquecida nesses repositórios. De acordo com tipo de informação que possa ser extraído desses bancos de dados, o processo de extração pode ser considerado complexo e superar a capacidade humana de analisar essas informações e transformá-las em conhecimento (Adriaans e Zantinge, 1996).

Para que os dados sejam devidamente manipulados, e conseqüentemente, se tenha extração de informações importantes, no sentido de, posteriormente, serem transformadas em conhecimento, faz-se necessário a utilização de técnicas que propiciem a automação desse processo a partir de estruturas artificialmente inteligentes, as quais envolvam técnicas necessárias para a compreensão da linguagem, percepção, raciocínio, aprendizagem e resolução de problemas, buscando a criação de teorias e modelos com capacidade cognitiva e a implementação de sistemas computacionais baseados nestes modelos, objetivando a descoberta dos conhecimentos engendrados no banco de dados, processo conhecido como descoberta de conhecimento em base de dados*, knowledge discovery in databases* - KDD (Adriaans e Zantinge, 1996; Russel e Novic, 1995; Diniz e Louzada-Neto, 2000).

O KDD foi proposto em 1989 (Piatetsky-Shapiro, 1991) para referir-se às etapas que produzem conhecimentos a partir de dados relacionados, sendo a mineração de dados, *data mining*, a etapa que transforma dados em informações. Assim, o KDD

refere-se ao processo de extração da informação relevante ou de padrões nos dados contidos em grandes BD e que sejam não-triviais, implícitos, previamente desconhecidos e potencialmente úteis, objetivando a tomada de decisão (Fayyad et al, 1996a).

Nesse sentido, a mineração de dados provém da análise inteligente e automática de dados para descobrir padrões ou regularidades em grandes conjuntos de dados, através de técnicas que envolvam métodos matemáticos, algoritmos baseados em conceitos biológicos, processos linguísticos e heurísticos, os quais fazem parte do processo KDD responsável pela busca de conhecimentos em banco de dados (Adriaans e Zantinge, 1996; Han e Kamber, 2006; Bigus, 1996; Fayyad et al, 1996a).

Os passos adicionais à mineração de dados no KDD, como preparação de dados, seleção de dados, limpeza de dados, incorporação de conhecimento prévio adequado e interpretação adequada dos resultados da mineração, são essenciais para garantir que se extraia o conhecimento útil a partir dos dados brutos. A figura 1.1 destaca essas etapas (Fayyad et al, 1996a).

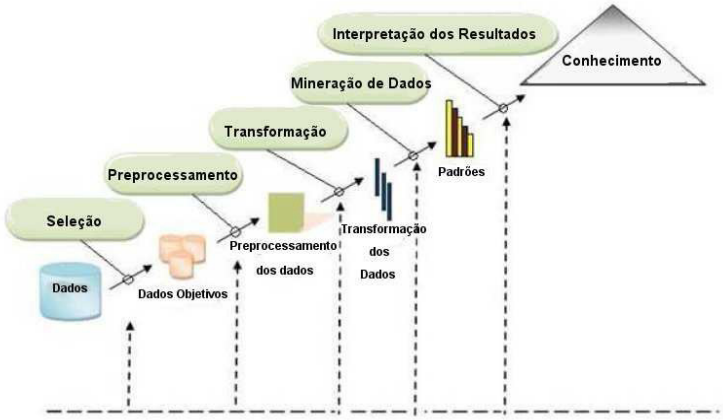

Figura 1.1: Principais tarefas do processo do KDD (Fayyad et al, 1996a)

A figura 1.1 apresenta as principais tarefas do KDD, ou seja, a análise de requisitos, que representa a maneira pela qual os dados e eventos se modificam nas transações do sistema, possibilitando a criação dos elementos necessários para o estudo e compreensão do domínio da aplicação objetivando o refinamento das informações relevantes; a seleção de dados, que busca identificar conjunto de dados relevantes e seus subconjuntos de variáveis objetivando a criação de um conjunto de restrito de dados para a descoberta de conhecimento; o pré-processamento, que envolve a limpeza dos dados, com operações de remoção dos ruídos, elaboração de esquemas e mapeamentos de valores desconhecidos; a transformação dos dados, onde se busca características úteis nos dados, utilizando métodos de redução ou transformação da dimensionalidade dos dados para um o melhor desempenho; a mineração de dados onde se aplica técnicas específicas em dados pré-processados com objetivo de buscar modelos de interesse numa representação incluindo regras de classificação, árvores de decisão, regressão ou agrupamento; e a interpretação dos dados, com a análise dos resultados obtidos, a qual permite avaliar padrões com objetivo de determinar quais as melhores maneiras de usar as informações na tomada de decisão (Fayyad et al, 1996a).

KDD evoluiu, e continua a evoluir, a partir da interseção de campos de pesquisa, como aprendizagem de máquina, reconhecimento de padrões, bases de dados, estatísticas, inteligência artificial, aquisição de conhecimento para sistemas especialistas, visualização de dados e computação de alto desempenho. O objetivo é unificar a extração de alto nível de conhecimento a partir de dados de baixo nível, no contexto de grandes conjuntos de dados. O componente de mineração de dados se baseia fortemente em técnicas conhecidas de aprendizagem de máquina, reconhecimento de padrões e estatísticas, para encontrar padrões de dados na etapa de mineração de dados do processo de KDD (Adriaans e Zantinge, 1996; Fayyad et al, 1996a; Han e Kamber, 2006; Diniz e Louzada-Neto, 2000).

2 A Mineração de Dados e a Análise Estatística

A análise estatística pressupõe a coleta, a manipulação e a organização dos dados, buscando a construção de modelos a partir da conceituação de distribuição de probabilidade, como a distribuição normal, da média e da variância. As técnicas de inferência estatística desses modelos, como a estimação e o teste de hipótese, a análise de regressão, análise de dispersão dos dados, análise discriminante, análise de agrupamento e de componentes principais são utilizadas para buscar informações

contidas nos dados, bem como analisar, reduzir dimensões e descobrir relacionamentos entre eles (Hair Jr. et al, 2005).

Dessa forma, as técnicas estatísticas, como análise multivariada de dados se colocam como ferramental importante no contexto de mineração de dados, inclusive quando combinadas com outras técnicas, tendo uma importante função na extração de informações relevantes de um conjunto de dados (Hair Jr. et al, 2005; Johnson, 1992; Field, 2009).

Em mineração de dados, dentre as especificidades, é possível utilizar tanto métodos estatísticos tradicionais como técnicas mais sofisticadas descrita em um ambiente de inteligência computacional. Nesse sentido, a mineração de dados pode ser vista como uma descendente direta da estatística, estando exatamente no limite do que poderia ser encontrado e inferido por métodos tradicionais de análise de dados, tratando de questões que estão além do domínio desses procedimentos (Adriaans e Zantinge, 1996; Han e Kamber, 2006; Diniz e Louzada-Neto, 2000).

3 Funcionalidades da Mineração de Dados

A multidisciplinaridade da mineração de dados pode ser considerada inevitável devido à integração de diversas áreas de conhecimento no processo de análise, abordando áreas de pesquisas que envolvem estatística, matemática e a computação, as quais são disciplinas fundamentais para realização do processo de mineração de dados. A figura 3.1 apresenta esses aspectos multidisciplinares da mineração de dados (Han e Kamber, 2006).

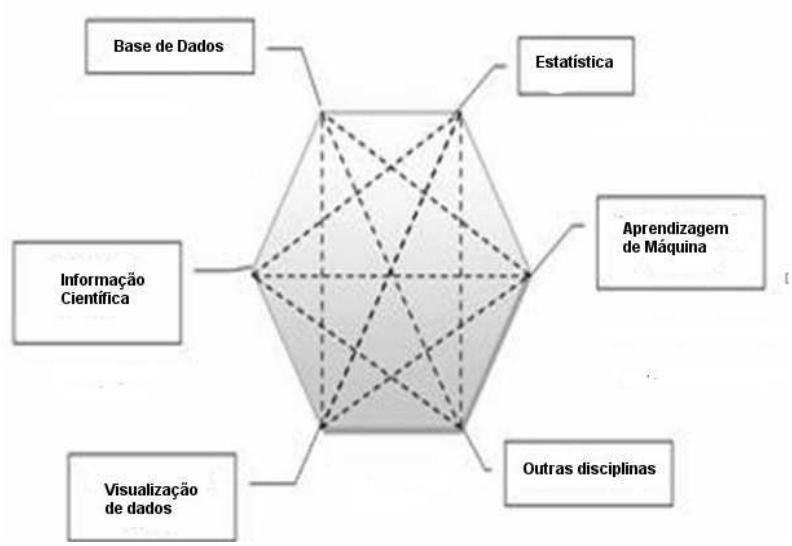

Figura 3.1: Mineração de dados e seu aspecto multidisciplinar (Han e Kamber, 2006).

Algumas limitações podem ser identificadas quando se refere à escolha do melhor método de mineração de dados, uma vez que, existe uma grande dificuldade dos especialistas na identificação desses métodos. Geralmente, os especialistas adotam um método mais adequado para cada problema específico. O processo de mineração de dados pode ser dividido em componentes capazes de favorecer a identificação mais adequada dos algoritmos de mineração quando se leva em consideração algumas informações relevantes tais como: a função do modelo e a representação do modelo (Han e Kamber, 2006).

4 Funções do Modelo

As funções do modelo são utilizadas para especificar o tipo de aplicação do algoritmo minerador. A seguir, serão descritas as funções mais comumente usadas, ou seja:

4.1 Classificação

A função de classificação tem por premissa reconhecer, em um conjunto de dados, as observações que tenham as mesmas características. A tarefa é descobrir se um item vindo do banco de dados pertence a uma das algumas classes, previamente, definidas. O problema é como definir essas classes. Na prática, as classes são muitas vezes definidas usando-se valores específicos de determinados campos nos registros de dados ou alguns derivados desses valores. Por exemplo, se um registro de

dados contém o campo *Região*, então algum dos valores típicos do campo, por exemplo, *Norte*, *Sul*, *Leste* ou *Oeste*, pode definir a classe (Devedzic, 2001).

A classificação de dados pode ser vista um processo em duas etapas. Na primeira etapa, um modelo é construído descrevendo um conjunto pré-determinado de classes de dados ou conceitos. O modelo é construído através da análise de uma lista ordenada de elementos do banco de dados, descrita por atributos. Cada lista é suposta pertencer a uma classe pré-definida, conforme determinado por um dos atributos, chamado atributo rótulo de classe. No contexto da classificação, os dados listados também são referidos como amostras, exemplos ou objetos.

As listas de dados analisadas para construir o modelo formam coletivamente um conjunto de treinamento de dados. As listas individuais que compõem o conjunto de treinamento são referidas como amostras de treinamento e são selecionados aleatoriamente da população de amostra. Como o rótulo de classe de cada amostra de treinamento é fornecido, esta etapa é vista como um aprendizado supervisionado. Na segunda etapa, o modelo é usado para a classificação. Primeiro, a precisão da previsão do modelo (classificador) é estimada. O método de validação usa um conjunto de testes das amostras de treinamento (Han e Kamber, 2006).

Uma vez que o algoritmo classificador tenha sido desenvolvido de forma eficiente, ele poderá ser usado de forma preditiva para classificar novos registros naquelas mesmas classes pré-definidas (Han e Kamber, 2006; Diniz e Louzada-Neto, 2000).

As técnicas de classificação mais usadas são a análise discriminante, através de árvore de decisão, baseadas em regras de decisão, por associação, baseadas em modelos adaptativos e evolutivos e a classificação bayesiana (Johnson, 1992; Apté e Weiss, 1997; Michie et al, 1994; Curram e Mingers, 1994; Haykin, 2009).

4.2 Regressão

A função de regressão é, basicamente, um conjunto de métodos que permite a interpretação da relação funcional entre variáveis com boa aproximação, considerando a existência de uma relação entre essas as variáveis, de modo que a medida possa estabelecer modelos utilizados para fins de predição. A análise de regressão pode ser considerada uma ferramenta analítica simples e eficiente, dentro de uma determinada relação de vizinhança, para explorar todos os tipos de relações de dependência. No caso mais simples, a regressão é uma função de aprendizado que mapeia um dado item a uma variável de predição de valor real. No caso geral tem-se a predição de uma ou mais variável dependente, considerada como resposta, a partir de conjunto de variáveis independentes, os preditores. Essa relação de múltiplas variáveis independentes como preditores é denominada *análise de regressão multivariada* (Hair Jr. et al, 2005; Johnson, 1992; Härdle e Simar, 2003).

Os métodos de regressão podem ser utilizados em diversas áreas de conhecimento como, por exemplo, para previsão da economia nacional com base em certas informações (níveis de renda, investimentos, etc...), para a verificação de quais fatores ajudam a manter a qualidade dos serviços oferecidos ou na medida de viabilidade de um novo produto. Também na construção de séries temporais onde as variáveis de entrada são versões atrasadas da variável de predição.

4.3 Análise de Associação

Análise de associação tem como objetivo elaborar uma representação explícita entre os objetos, visando determinar relacionamentos entre conjuntos de itens de associação. Ela gera redes de interações e conexões presentes nos conjuntos de dados usando as associações item a item. Nesse sentido, a presença de um item implica necessariamente na presença do outro item na mesma transação. Em geral, uma regra de associação pode ser representada formalmente através do tipo, *se X então Y*, considerados corpo e cabeça da regra, respectivamente (Diniz e Louzada-Neto, 2000).

A figura 4.1 exemplifica uma regra de associação voltada a identificar afinidades entre itens de um subconjunto de dados, dentro de um conjunto de valores (produtos comprados por um cliente, sintomas apresentados por um paciente, etc.), destacando, ainda, dois fatores importantes, o de *confidência* e o de *suporte*.

Figura 4.1: Exemplo de uma regra de associação (Diniz e Louzada-Neto, 2000).

4.4 Análise de Seqüência

Análise de seqüência constitui-se de uma variação da análise associativa objetivando extrair e registrar desvios e tendências no tempo. As regras identificadas são usadas para reconhecer seqüências relevantes que possam ser utilizadas para prever comportamentos, modelar processos gerando uma seqüência ou relatar tendências de um processo ao longo do tempo (Adriaans e Zantinge, 1996).

Assim, por exemplo, seja um conjunto de dados ordenado pelo sobrenome do consumidor e pelo período de transação de compras (Oliveira* visitou a loja em dois dias consecutivos, comprou cerveja no primeiro dia e vodka no segundo dia). A tabela 4.1, mostra as seqüências de transações de consumidores organizadas segundo o tempo, onde cada conjunto de parênteses indica uma transação que inclui um ou mais itens (Diniz e Louzada-Neto, 2000).

Tabela 4.1: Seqüências de transações dos consumidores data (Diniz e Louzada, 2000)

Consumidor*	**Seqüência diária de compras de bebidas**			
Oliveira	(Cerveja)	(Vodka)		
Soares	(Guaraná, Suco)	(Cerveja)	(Água, Licor, Vinho)	(Gin, Licor)
Tenório	(Cerveja)	(Água, Gin, Vinho)	(Vodka, Soda)	
Zacaria	(Vodka)			

***sobrenomes fictícios**

As técnicas de busca de característica seqüencial detectam características entre as transações de tal forma que a presença de um conjunto de itens é seguida por outro conjunto de itens em um banco de dados de transações em um período de tempo (Adriaans e Zantinge, 1996). Vê-se, no caso da tabela 4.1, que a característica seqüencial "*cerveja é comprada em uma transação anterior a que a vodka é comprada*" ocorre em dois dos quadros de consumidores. A técnica também determina a freqüência de cada combinação de transações que pode ser produzida nas seqüências de consumidores e disponibiliza as características seqüenciais cujas ocorrências relativas são maiores que um determinado nível de suporte mínimo requerido (Adriaans e Zantinge, 1996). A tabela 4.2, apresenta as características seqüenciais com suporte maior que 40%.

Tabela 4.2: Características Seqüenciais com Suporte > 40% (Diniz e Louzada, 2000)

Características Seqüenciais (com Fator de Sustentação > 40%)	**Consumidores de Apoio**
(Cerveja), (Vodka)	(Oliveira, Tenório)
(Cerveja), (Vinho, Água)	(Soares, Tenório)

4.5 Sumarização

A função de sumarização visa obter uma descrição compacta de um conjunto de dados, bastante usada em análise exploratória de dados. Geralmente, a sumarização não é usada para a resolução de problemas, mas possibilita identificar características no conjunto de dados que possa estar contaminadas por ruídos, que interfiram no processo de análise, ou redundantes, gerando uma tendência errônea à análise. A sumarização é usada, principalmente, no pré-processamento dos dados, onde valores inválidos, no caso de variáveis quantitativas, são determinados através do cálculo de medidas estatísticas e, no caso de variáveis categóricas, através da distribuição de freqüência dos valores (Hair Jr. et al, 2005; Johnson, 1992). O objetivo da sumarização em mineração de dados é propiciar a limpeza dos dados facilitando a análise e a geração automatizada de relatórios (Bigus, 1996; Diniz e Louzada-Neto, 2000).

Em base de dados com informações complexas, outras formas complementares de sumarização podem ser implementadas, tais como a análise de componentes principais (Johnson, 1992) ou de componentes independentes (Hyvärinen, Karhunen e Oja, 2001), mais sofisticadas, destinando-se, inclusive, a auxiliar em técnicas de visualização de dados (Günsel, Tekalp e Van Beek, 1998), as quais têm sido parte integrante da análise estatística tornando-se de extrema importância para se obter informações a partir de um entendimento, muitas vezes, indutivo do conjunto de dados. É importante destacar, neste item, a caracterização, a qual descreve qualidades relevantes a partir da análise quantitativa, propiciando uma descrição compacta do conjunto, podendo generalizar, resumir e inclusive contrastar características de dados. Nesse sentido, sumarização e caracterização tendem a ser complementares (Härdle e Simar, 2003).

4.6 Visualização

As técnicas de visualização podem ser consideradas ferramentas eficientes para se analisar grandes quantidades de dados. Em muitas situações, elas são suficientes para a extração das respostas de interesse, descobrindo padrões, tendências, estruturas e relações, dentro de um conjunto de dados (Han e Kamber, 2006; Günsel, Tekalp e Van Beek, 1998). O método de visualização escolhido para análise dependerá basicamente do tipo de conjunto de dados disponível e como esses dados podem ser modelados, por exemplo, se o conjunto de dados envolve chamadas telefônicas feitas em um intervalo de tempo específico, então uma representação visual desta informação poderia ser sumarizada através de um simples diagrama de associação, disponibilizando todas as relações entre as chamadas, conforme tabela 4.3 (Diniz e Louzada-Neto, 2000).

Tabela 4.3 Representação tabular das chamadas telefônicas (Diniz e Louzada-Neto, 2000)

De	1	1	2	4	4	8	7	8
Para	2	3	6	6	7	6	5	6
Horário	07:45	08:00	08:36	09:16	09:48	11:22	11:51	12:03
De	7	6	3	2	8	6	2	6
Para	4	2	2	6	6	2	6	7
Horário	14:03	14:18	14:53	15:34	16:19	16:38	17:05	17:28

A figura 4.2, apresenta a visualização de várias camadas entre certos pares de telefones. As linhas mais grossas no diagrama representam os números maiores de chamadas. A partir desse diagrama, é possível detectar, rapidamente, quais números requerem uma análise mais detalhada, enquanto, no formato tabular, cálculos adicionais são necessários para a análise de freqüência. Também é possível verificar ocorrências de associações com outros números, para se obter a mesma informação.

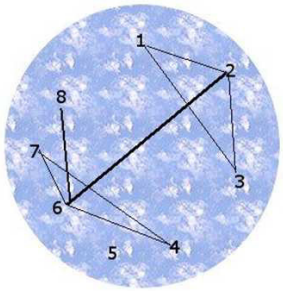

Figura 4.2: Diagrama de associação das chamadas telefônicas (Diniz e Louzada-Neto, 2000)

Representações de métodos de visualização bem comum são, por exemplo, os métodos de visualização simples de dados, os quais se baseiam em gráficos ou resumos rápidos que, de alguma forma, representam ou resumem características dos conjuntos de dados. Os gráficos podem ser plotados em formatos bidimensionais, tridimensionais e etc., proporcionando o relacionamento entre atributos de dados ou integrando expressões matemáticas a partir da média, da potenciação, do logaritmo, etc. Esses gráficos formam uma espécie de "descrição sucinta" dos conjuntos de dados, cuja análise preliminar, possibilitaria um melhor entendimento dos dados e evitaria a aplicação negligente de técnicas de mineração de dados, o que muitas vezes leva a resultados sem sentido (Fayyad et al, 1996a).

A classificação dos métodos de visualização de dados pode ser resumida a partir de histogramas, gráficos relacionando atributos ou resumos destes entre si ou representações icônicas, onde normalmente associa-se um atributo de dado a um atributo de uma figura que o representará (Everitt, Landau, e Leese, 2001). Outros métodos de visualização de dados incluem: diagramas baseados em proporções, diagramas de dispersão, histogramas, box plots entre outros (Härdle e Simar, 2003).

Os modelos que incluem a representação dos dados através de figuras poligonais podem ser visualizados a partir das "faces de Chernoff ", considerada uma técnica para ilustrar tendências em dados multidimensionais. As faces representadas por este modelo ilustram características para representar dados em diferentes dimensões, capazes de representar tendências em termos de valores nos dados e podem ser utilizadas para visualizar graficamente dados multivariados complexos (Chernoff, 1973). As faces apresentadas na figura 4.3, ilustram o usuário na detecção de padrões, agrupamentos e correlações entre os dados, os quais são simplificados a partir de desenhos provenientes da face humana.

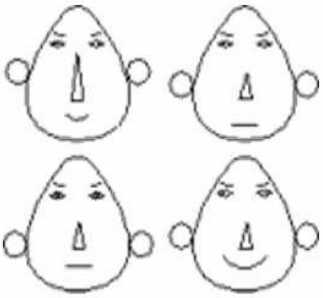

Figura 4.3: Faces de Chernoff (Chernoff, 1973).

5 Representação do Modelo

As funções do modelo têm um papel importante na análise e modelagem do problema. Porém, a integração da função e a representação do modelo podem ser consideradas um dos fatores de grande relevância, uma vez que, os modelos representados a partir de algoritmos de mineração de dados, podem determinar a flexibilidade do mesmo em representar o conjunto de dados e a sua interpretação. Os modelos mais complexos podem ajustar melhor os dados, entretanto, ficam mais difíceis de serem interpretados (Diniz e Louzada-Neto, 2000). Representações mais tradicionais incluem árvore de decisão (Adriaans e Zantinge, 1996), conjunto de regras (Han e Kamber, 2006), métodos de agrupamento (Fayyad et al, 1996a), modelos lineares (Hair Jr. et al, 2005; Johnson, 1992) e não lineares (Härdle e Simar, 2003; Haykin, 2009), os quais são descritos a seguir.

5.1 Árvores de Decisão e Regras de Decisão

Quando o processo de mineração de dados é direcionado à classificação, o método de árvore de decisão pode ser conveniente quando o objetivo se relaciona à categorização dos dados. As árvores de decisão são ferramentas eficientes e populares para classificação e diagnóstico. A árvore é formada por nós e o primeiro, nó raiz, envolve todo o conjunto de dados, onde o processo de classificação se inicia. A estrutura de uma árvore de decisão pode ser ilustrada conforme figura 5.1, onde cada nó interno identifica um dos atributos de previsão. Cada linha que sai desse nó identifica um valor assumido por tal nó e cada nó terminal (folha) identifica o resultado da previsão ou objetivo (Apté e Weiss, 1997).

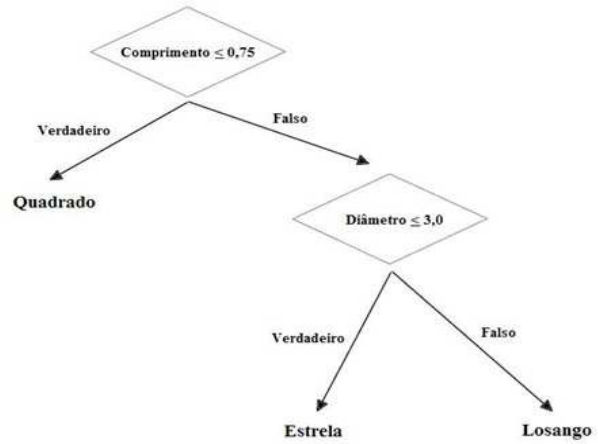

Figura 5.1: Classificação por árvore de decisão do formato de pinos dados comprimento e diâmetro (Apté e Weiss, 1997).

No exemplo da figura 5.1, o nó raiz testa todos os itens para o comprimento ≤ 0,75. Os que satisfazem o teste são considerados verdadeiros, indo para uma folha, indicando que todos os pinos pertencem a classe *Quadrado*. A linha de falsos, que sai do nó raiz encaminha todos os casos que falharam no teste inicial. Esses pinos ainda não pertencem a uma só classe e, portanto, futuros testes serão necessários ao nó intermediário. O teste nesse nó é para o diâmetro ≤ 3,00. Os pinos que satisfazem o teste vão para a classe *Estrela* e os que falharam, para a classe *Losango*, ou seja, o teste conduziu a outras folhas.

Considerando ainda o exemplo anterior, onde estão disponíveis dados sobre o comprimento e diâmetro de uma série de pinos, que podem ter o formato de quadrado, de estrela ou de losango, uma classificação que caracterize a variedade do pino como uma função do comprimento e do diâmetro pode ser útil para se entender como essas variedades diferem. Os dados são ilustrados na figura 5.2, que apresenta duas linhas paralelas aos eixos, uma no comprimento 0.75 e outra no diâmetro 3,00, particionando as três variedades em três sub-áreas. Métodos de solução por decisão por árvore fornecem automaticamente estas partições de eixos paralelos (Apté e Weiss, 1997).

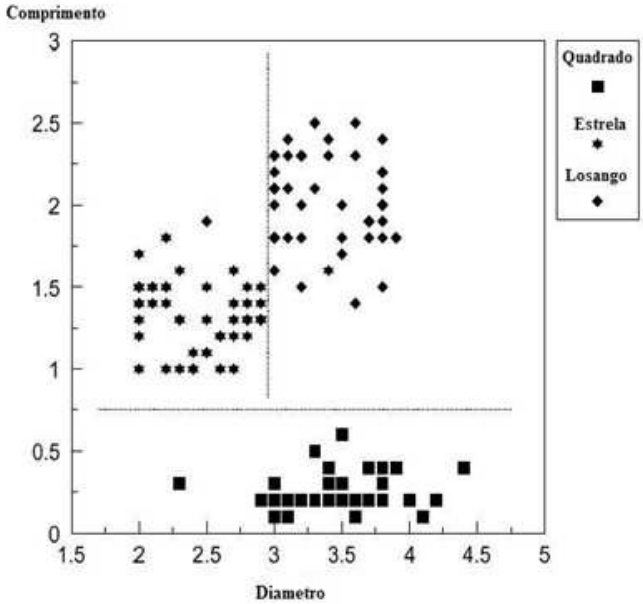

Figura 5.2: Dados dos pinos: classificação que caracteriza a variedade dos pinos (Apté e Weiss, 1997).

As regras de decisão podem ser consideradas um processo para analisar uma série de dados e a partir dela gerar padrões. Também podem ser vistas como a expressão verbal das árvores de decisão. Nesse sentido, a integração dos métodos de árvore

de decisão e regra de decisão pode ser considerada ferramenta fundamental em previsão. Uma regra pode ser construída através da formação de um conjunto de testes que ocorre nos caminhos entre nó raiz e os nós terminais da árvore. A coleção de todas as regras obtidas em cada caminho do nó raiz a um nó terminal é uma solução baseada em regras para a classificação (Han e Kamber, 2006).

No exemplo dos pinos, figuras 5.1 e 5.2, pode-se utilizar, para ilustrar a decisão por árvore, a solução por regra (Diniz e Louzada-Neto, 2000):

Se (comprimento ≤ 075)

Então Quadrado

Se (não (comprimento ≤ 075)) & (diâmetro ≤ 3,00)

Então Estrela

Se (não (comprimento ≤ 075)) & (não (diâmetro ≤ 3,00))

Então Losango

Gerada uma solução utilizando árvore de decisão ou regra de decisão, esta pode ser usada para estimar ou predizer a resposta ou classe variável para um novo caso (Han e Kamber, 2006).

6 Análise de Agrupamento

A prática de classificar objetos de acordo com similaridades percebidas pode ser considerada a base inicial para vários aspectos da ciência. O principal objetivo da análise de agrupamento (*cluster*) está relacionado ao processo de agrupar elementos de dados mediante o particionamento de uma população heterogênea em subgrupos mais homogêneos. Nesse sentido, a análise de agrupamento é o estudo formal dos algoritmos e dos métodos para agrupar ou classificar objetos (Jain e Dubes, 1988).

No agrupamento, não há classes pré-definidas, os elementos são agrupados de acordo com a semelhança, o que a diferencia da tarefa de classificação, buscando reunir indivíduos ou objetos em grupos tais que os objetos no mesmo grupo são mais parecidos uns com os outros do que com os objetos de outros grupos. A sua abordagem se dá em diversas áreas de conhecimento tais como ciências biológicas (por exemplo, criar a taxonomia biológica para a classificação de vários grupos de animais), ciências humanas (por exemplo, analisar vários perfis psiquiátricos) e outros. A idéia é maximizar a homogeneidade de objetos dentro dos grupos, ao mesmo tempo em que maximiza a heterogeneidade entre os grupos (Han e Kamber, 2006).

A figura 6.1 mostra uma árvore abordando diferentes métodos de agrupamento aplicados ao problema de classificação de forma simplificada, os quais são descritos a seguir:

Figura 6.1: Classificação simplificada dos métodos de agrupamentos (Jain e Dubes, 1988).

6.1 Métodos de Agrupamento

O método de agrupamentos é uma técnica analítica para desenvolver subgrupos significativos de indivíduos ou objetos e têm como objetivo classificar uma amostra de entidades em um pequeno número de grupos mutuamente exclusivos, com base nas similaridades entre eles (Hair Jr. et al, 2005; Johnson, 1992). Essa técnica pode ser dividida em três etapas: a primeira relaciona-se a medida de similaridade ou associação entre as entidades para determinar quantos grupos realmente existem na amostra; a segunda refere-se ao processo de busca do agrupamento, no qual entidades são particionadas, e o último passo busca estabelecer o perfil das variáveis para determinar sua composição (Jain e Dubes, 1988). A figura 6.2 mostra os métodos.

Figura 6.2: Classificação simplificada dos métodos de agrupamentos adaptados (Jain e Dubes, 1988).

O critério de classificação utilizando os métodos intrínsecos é a essência da análise de agrupamento (Jain e Dubes, 1988). Existem várias adaptações ao modelo simplificado aos métodos de agrupamentos, representados graficamente a partir da figura 6.2 (Jain e Dubes, 1988; Rencher, 2002).

6.2 Agrupamento Hierárquico

A análise de agrupamentos corresponde ao grupo de técnicas multivariadas de dados cuja finalidade primária é agregar objetos com base nas características que eles possuem. Entre essas técnicas encontram-se o procedimento hierárquico, que opera para formar um intervalo inteiro de soluções de agrupamentos (Hair Jr. et al, 2005; Johnson, 1992), ou ainda, que opera como um método aglomerativo, objetivando fundir agrupamentos individuais (inicialmente, cada grupo contém um único objeto) em partições maiores até a obtenção de uma única partição contendo todos os objetos do conjunto, onde os agrupamentos são formados pela combinação de outros já existentes (Rencher, 2002) .

O procedimento hierárquico trata o conjunto de dados como uma estrutura de partições, cada uma correspondendo a um agrupamento, hierarquicamente organizadas segundo a similaridade entre seus objetos (Jain e Dubes 1988). A maioria dos métodos de análise de agrupamento requer uma medida de similaridade entre os elementos a serem agrupados, a qual é normalmente expressa por uma função distância ou métrica. Por exemplo, a similaridade pode ser medida a partir de uma associação, onde os coeficientes de correlação positivos maiores medidos representam maior similaridade (Hair Jr. et al, 2005). A proximidade entre cada par de objetos pode avaliar a similaridade onde medidas de distância ou de diferença são empregadas e as menores distâncias ou diferenças representam maior similaridade (Rencher, 2002; Han e Kamber, 2006).

A tabela 6.1, exibe a representação formal de algumas medidas de similaridade que são usadas na análise de agrupamentos. A métrica mais utilizada é a distância euclidiana, principalmente, quando não há nenhuma outra informação prévia existente acerca dos dados de entrada que possa afetar diretamente na quantidade de formação dos grupos encontrados pelos algoritmos de agrupamento (Jain e Dubes, 1988; Rencher, 2002; Kohonen, 1997; Haykin, 2009).

Tabela 6.1: Distâncias: (A) Euclidiana, (B) Quadrado da Euclidiana, (C) Manhattan e (D) Chebychev.

Medidas de Similaridade	Representação Formal
Objetos	$X=[x_1, x_2, ...,x_n]$ *e* $Y = [y_1, y_2, ...,y_n]$

A	$d_{xy} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
B	$d_{xy} = \sum_{i=1}^{m}(x_i - y_i)^2$
C	$d_{xy} = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$
D	$d_{xy} = \text{máx}\ (\lvert x_1 - y_1 \rvert + \lvert x_2 - y_2 \rvert + ... + \lvert x_n - y_n \rvert)$

6.3 Métodos Aglomerativos

Dentre os métodos hierárquicos que envolvem a construção de uma hierarquia numa estrutura em árvore, encontram-se as técnicas aglomerativas, utilizadas para descobrir agregados e são divididas em (Kaufman e Rousseeuw, 1990):

Ligação Individual ou Simples: Procedimento baseado na distância mínima. Ele encontra os dois objetos separados pela menor distância e os coloca primeiro no agrupamento. Em seguida, a próxima distância mais curta é determinada, e um terceiro objeto se junta aos dois primeiros para formar um agregado, ou um novo agrupamento de dois membros é formado. O processo continua até que todos os objetos formem um só agregado. Esse procedimento também pode ser chamado de abordagem do vizinho mais próximo.

Ligação Completa: Procedimento baseado na distância máxima. Por essa razão, às vezes é chamado de abordagem do vizinho mais distante ou de método do diâmetro. A distância máxima entre indivíduos em cada agregado representa a menor esfera (diâmetro mínimo) que pode incluir todos os objetos em ambos os agrupamentos. Esse método é chamado de ligação completa porque todos os objetos em um agrupamento são conectados um com o outro a alguma distância máxima ou similaridade mínima. Podemos dizer que a similaridade interna se iguala ao diâmetro do grupo. Esta técnica elimina o problema de encadeamento identificado na ligação individual.

Ligação Média: Procedimento baseado na distância média de todos os indivíduos em um agrupamento aos demais em outro. Esta técnica não depende de valores extremos, como ocorre na ligação individual ou completa, pois a partição é baseada em todos os elementos dos agregados, ao invés de um único par de membros extremos. Abordagens de ligação média tendem a combinar agregados com pequena variação interna tendendo a produzir agregados com aproximadamente a mesma variância.

A figura 6.3 ilustra esse dois casos. O critério de ligação simples define d_{AB} como a menor distância entre todos os pares *(x, y)* de dois objetos onde $x \in A$ e $y \in B$ e o critério de ligação completa define d_{AB} como a maior distância entre todos os pares *(x, y)*. Assim, depois de calculadas as distâncias entre os agrupamentos, os algoritmos promovem a união dos agrupamentos com a menor distância entre si.

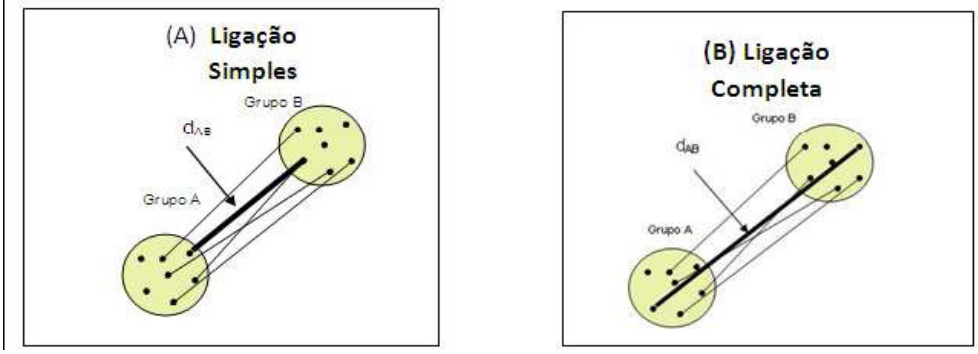

Figura 6.3 Ilustração do Critério de Ligação Simples e Completa (Kaufman e Rousseeuw, 1990).

Além dos procedimentos de ligação, a agregação ainda pode ser listada por (Kaufman e Rousseeuw, 1990; Hair Jr. et al, 2005):

Método de Ward: Procedimento onde a distância entre dois agrupamentos é a soma dos quadrados entre os dois agrupamentos, feita sobre as variáveis. Em cada estágio do procedimento de agrupamento, a soma interna de quadrados é minimizada sobre todas as partições (o conjunto completo de agrupamentos disjuntos ou separados) que podem ser obtidas pela combinação de dois agregados do estágio anterior. Este procedimento tende a combinar agrupamentos com um pequeno número de observações. Ele também tende a produzir agregados com aproximadamente o mesmo número de observações.

Método Centróide: Procedimento onde a distância entre dois agrupamentos é a distância (geralmente euclidiana quadrada ou euclidiana simples) entre seus centróides. Centróides são valores médios das observações na variável estatística de agrupamento. Neste método, toda vez que indivíduos são reunidos, um novo centróide é computado. Os centróides migram quando ocorrem fusões de agregados. Em outras palavras, existe uma mudança no centróide do agrupamento toda vez que, um novo indivíduo ou grupo de indivíduos é acrescentado a um agregado já existente. Esses algoritmos diferem na forma como a distância entre os agrupamentos é computada. A saída gerada por esses algoritmos podem proporcionar a criação de gráficos demonstrando o processo de formação desses agrupamentos, conforme figura 6.4. Os diagramas representados na figura 6.4, exibem uma representação gráfica de diferentes tipos de clusters. As representações mais comuns são do modelo *dendrograma, do grego dendro* (árvore), ou seja, diagrama em árvore, que representa as junções sucessivas de partições e que pode gerar agrupamentos diferentes conforme o nível em que é seccionada (Witten e Frank, 1999; Rencher, 2002).

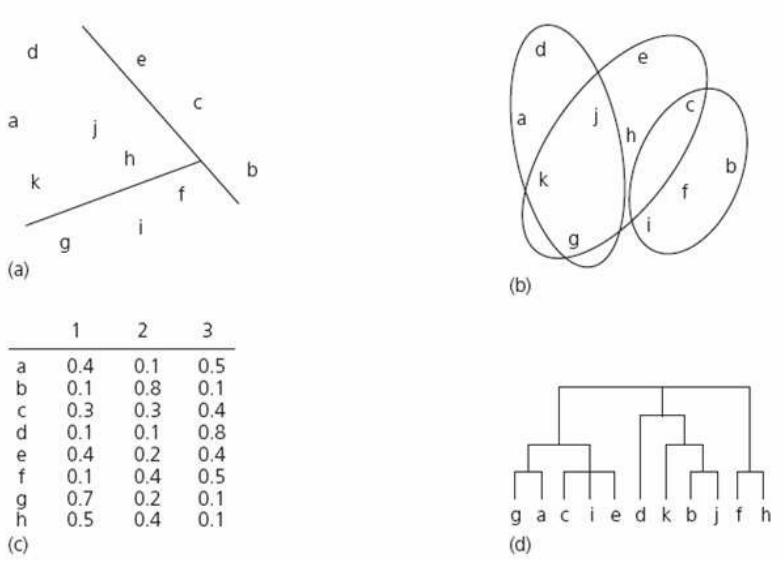

Figura 6.4 Diferentes tipos de representação de agrupamentos.

6.4 Métodos Divisivos

Quando o processo de classificação utiliza métodos divisivos ocorre uma inversão com relação ao método hierárquico aglomerativo. Pode-se observar que um conjunto contendo todos os dados é particionado a partir de um aglomerado unificado. Os métodos divisivos começam o processamento a partir de um grande agregado que contém todas as observações (objetos). Em passos sucessivos, as observações mais diferentes entre si são separadas e transformadas em agrupamentos menores, ou seja, pode-se considerar a existência de uma única partição (o próprio conjunto de dados), subdividindo esta partição em uma série de partições alinhadas. Os métodos hierárquicos são bastante utilizados no processo de análise multivariada de dados, porém, em bases de dados de grande porte, esses métodos podem se tornar impraticável por ser difícil visualizar as informações contidas no dendrograma e nos demais tipos de gráficos resultados dessa classificação, tornando-os pouco indicáveis em mineração de dados (Jain e Dubes 1988; Hair Jr. et al, 2005).

6.5 Métodos Particionais

Os métodos particionais têm como objetivo dividir um conjunto de objetos em um número pré-estabelecido de clusteres. O método mais popular é conhecido como *k-means*. Esse algoritmo divide o conjunto de dados em partes disjuntas, satisfazendo as seguintes recomendações: a) objetos de uma mesma parte estão próximos, de acordo com um critério de dado; b) objetos de partes distintas estão longe, de acordo com este mesmo critério. A subdivisão realizada pelo algoritmo *k-means*, como método particional, é feita da seguinte maneira: cria-se uma partição inicial aleatória de K partes e posteriormente, em um processo iterativo, os elementos das partes vão sendo realocados para outras partes, de modo a melhorar o particionamento a cada iteração, isto é, de modo que cada parte contenha, realmente, objetos que estão próximos e objetos em partes distintas estejam longe um do outro. Dessa forma, os métodos particionais dividem o conjunto dos N objetos em K agrupamentos sem relacioná-los hierarquicamente entre si, como o fazem os métodos hierárquicos (Jain e Dubes 1988; Rencher, 2002).

Normalmente as partições são obtidas por um processo de otimização, com a busca a partir de um critério local, definido sobre um subconjunto de objetos, ou de um critério global, na forma de uma função objetivo. A aproximação dos objetos pode ser

analisada a partir da formação de uma matriz de distância ou das similaridades de acordo com uma métrica pré-estabelecida (no caso da distância euclidiana). Assim, pode-se descrever os objetos do conjunto de dados, por uma simples representação matemática, onde $X = [x_1, x_2, x_3, ...x_n]$ representa o conjunto de objetos de um número $K \leq X$ representando o número de *clusters* que se deseja formar.

Seja $C = [C_1, C_2, C_3, ..., C_n]$ uma partição do conjunto de dados em *k clusters* e sejam os elementos escolhidos em cada um dos *clusters*, representando os centros dos mesmos, os centróides. Cada entidade influencia o grupo, cujo protótipo $V = [v_1, v_2, v_3, ...,v_k]$ são os elementos escolhidos em cada um dos *clusters*, representando os centros de área ou centróides. Cada entidade x_i $(i = 1, ..., n)$, influencia o grupo C_k $(k = 1, ..., K)$,cujo protótipo v_k está mais próximo. Os centróides constituem valores médios dos objetos contidos no agrupamento sobre cada variável usada nas variáveis estatísticas de agrupamento ou no processo de validação.

Toda vez que os objetos são reunidos, um novo centróide é computado e os objetos são realocados a partir de v_k (Hair Jr. et al, 2005; Rencher, 2002), onde:

$$v_k = \frac{1}{m}\sum_{j=1}^{m} x_j^k$$

Uma característica relevante do *k-means* está no emprego da função objetivo *erro quadrático total*, definida para um número *K* de agrupamentos, representada pela equação (Jain e Dubes 1988; Rencher, 2002):

$$e_k^2 = \sum_{j=1}^{K}\sum_{i=1}^{N} \| x_i^{(j)} - v_k \|^2$$

O k-means recebe como entrada um número *K* de agrupamentos e atribui aleatoriamente um objeto como sendo o centróide inicial de cada agrupamento. Sucessivamente, cada objeto é associado ao agrupamento mais próximo e o centróide de cada agrupamento é então recalculado levando-se em conta o novo conjunto de objetos a ele pertencentes (Jain e Dubes 1988).

A figura 6.5 mostra as iterações do algoritmo *k-means* utilizando $K = 3$. Os objetos selecionados em círculos representam a escolha aleatória dos *k* protótipos, nas iterações seguintes os centróides são marcados pelo sinal de +. Em seguida, pode-se visualizar o critério de convergência do algoritmo que pode ser analisado a partir das observações referentes as trocas de objetos.

O algoritmo *k-means* converge quando ocorrem poucas trocas de objetos entre os grupos ou quando o valor de e_k^2 é minimizado, ou até mesmo quando v_k não se altera em duas iterações consecutivas. O método *k-means* possui sua maior vantagem quando atua sobre um conjunto de dados com elevado número de objetos pertencentes (Rencher, 2002).

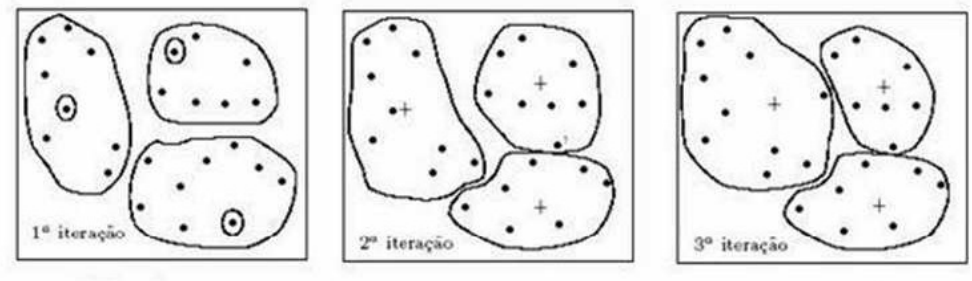

Figura 6.5 Método *k-means* na formação de *clusters* (Rencher, 2002).

7 - Grafos

Os agrupamentos de dados baseado em grafos utilizam algoritmo baseado na construção de uma árvore geradora mínima (*Minimum Spanning Tree - MST*) e têm como objetivo à geração de um grafo, de modo que os objetos não possuam ciclos e sejam conectados por um arco, ou seja, uma árvore. Assim, os agrupamentos são obtidos a partir do primeiro arco de maior comprimento de produção dos aglomerados. Os agrupamentos obtidos pelo algoritmo MST são sub-grafos bastante similares ao método de agrupamento aglomerativo, especificamente as ligações simples e completa (Jain e Dubes 1988).

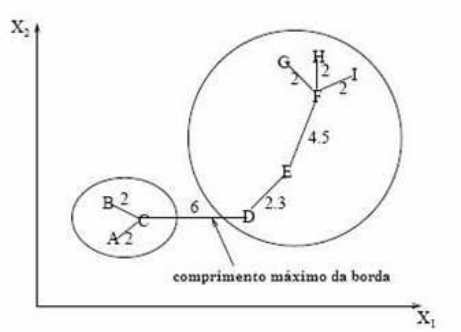

Figura 6.6 Algoritmo de caminho mínimo abrangendo a árvore de *clusters* (Jain e Dubes 1988).

8 – Métodos de Projeção

Os operadores de projeção procuram realizar um conjunto de testes a partir do mapeamento de objetos de um espaço *N* - dimensional em um espaço *M* - dimensional, onde $M < N$. O principal objetivo da realização desses testes é permitir a análise visual dos dados utilizando técnicas que possam exibir uma estrutura do espaço original o mais fielmente possível no hiperplano de projeção, possibilitando assim uma análise de agrupamentos que pode ser realizada visualmente, nos casos $M=2$ ou $M=3$, e que poderá servir para validar resultados obtidos por outros métodos de mineração de dados (Jain e Dubes 1988).

Quando esses testes são executados a partir de uma projeção linear, pode-se obter novas características a partir da combinação linear das características originais dos dados na dimensão *N* (Jain e Dubes 1988). Nesse caso, o mapeamento pode ser representada por:

$$\mathbf{y}_i = \mathbf{A}\,\mathbf{x}_i\ ,\ i = 1, \ldots, N$$

onde **A** é uma matriz $P \times D$ que gera os vetores $\boldsymbol{y} = [y_1, y_2, y_2, \ldots y_P]^T \in \boldsymbol{R}^P$, podendo ser representado por uma combinação linear de suas colunas $\mathbf{a}_j \in \boldsymbol{R}^P$, tal como:

$$y_j = \sum_{j=1}^{D} \mathbf{a}_j\ x_j$$

Os algoritmos de projeção lineares são bastante simples, e os diferentes tipos são definidos a partir das colunas da matriz **A.**

Um dos mais populares algoritmos é a projeção de auto-vetores, que também pode ser conhecida como método de *Karhunen–Loeve* ou Análise de Componentes Principais, PCA (Jain e Dubes 1988; Johnson, 1992; Diamantaras e Kung, 1996). Essa técnica estatística tem como objetivo condensar dados originais obtidos a partir de um conjunto de variáveis com dimensão elevada em um conjunto menor de variáveis com uma perda mínima de informação. O PCA é um método de identificar padrões nos dados, visando expressar os mesmos de modo a salientar as similaridades e diferenças existentes. Essas diferenças podem ser denominadas de processo de seleção ou extração de características (Diamantaras e Kung, 1996; Rencher, 2002; Härdle e Simar, 2003; Haykin, 2009).

De acordo com Haykin (Haykin, 2009) "*a seleção de característica se refere a um processo no qual um espaço de dados é transformado em um espaço de característica que, em teoria, tem exatamente a mesma dimensão que o espaço original de dados. Entretanto, a transformação é projetada de tal forma que o conjunto de dados pode ser representado por um número reduzido de características "efetivas" e ainda reter a maioria do conteúdo de informação intrínseco dos dados; em outras palavras, o conjunto de dados sofre uma redução de dimensionalidade*".

O método PCA toma um conjunto $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ onde $\mathbf{x}_n = [x_{n1}, x_{n2}, \ldots, x_{nD}]$, numa base ortogonal e encontra uma nova base ortonormal $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_D\}$ capaz de gerar o espaço original. Essa nova base é rotacionada de forma que o primeiro eixo coincida com a direção de maior variância dos dados, e assim sucessivamente com os demais eixos ortogonais ao primeiro. Para as *n* soluções possíveis para o vetor $\boldsymbol{q}$, podem ser constatadas a existência de *n* projeções possíveis do vetor de dados **x** a serem considerados por:

$$\mathbf{a}_{i\,=}\,\mathbf{q}_i^{\,T}\,\mathbf{x},\ i = 1,2, \ldots, n$$

onde

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_N] \quad \textit{(autovetores)}$$

e os $\mathbf{a}_i$´*s* são considerados as projeções de **x** sobre as direções principais representadas pelos vetores de entrada, também conhecidos como componentes principais e possuem as mesmas dimensões físicas que o vetor de dados **x**. A equação acima pode ser vista como uma fórmula de análise.

Para reconstruir exatamente o vetor de dados original ***x*** a partir das projeções $\mathbf{a}_i$, considerando $[\mathbf{a}_i \quad i = 1, 2, ..., n]$, combinações do conjunto de projeções em um único vetor a partir de:

$$\mathbf{x} = \mathbf{Qa}$$

Neste sentido, os vetores representam uma base no espaço de dados, ou seja, não é nada mais do que uma transformação de coordenadas, de acordo com a qual um ponto *x* no espaço de dados é transformado em um ponto *a* correspondente um espaço características. Desta forma, a praticidade na análise de componentes principais constitui em fornecer uma técnica efetiva para redução de dimensionalidade, descartando as combinações lineares que têm variâncias pequenas.

Haykin (Haykin, 2009) aborda a análise de componentes principais utilizando redes neurais na busca dos coeficientes. Outros métodos são vistos na literatura, se destacando a análise discriminante (Jain e Dubes 1988; Johnson, 1992).

9 – Conclusão

Este artigo buscou uma introdução a mineração de dados, não tendo a intenção de ser uma obra completa. Neste contexto, outras técnicas podem ser vista na literatura, com detalhes, como forma complementar aos métodos de mineração de dados. Nesse sentido pode-se citar as ferramentas supervisionadas e não supervisionadas com o uso da teoria de redes neurais (Principe et al, 2000; Haykin, 2009; Kohonen, T. K. 1997), dos algoritmos bio-inspirados (Holland, 1992, Jain et al, 1999), dos conjuntos nebulosos (Zadeh, 1996), da análise de componentes independentes (Hivärinen, 2001) e outros.

No contexto da classificação, trabalhos referenciados podem ser vistos em Metha et al (1996) e Shafer et al (1996). Novos métodos de agrupamento *fuzzy cluster* podem ser vistos em Silvanandam et al (2007). Outros trabalhos importantes são citados em Fayyad et al (1996b), Costa (1999), Zuchini (2003).

10 Agradecimentos

Os autores agradecem à CAPES, no contexto do projeto PROCAD NF 2009, pelo apoio financeiro.

11 Referências

Adriaans, P. e Zantinge, D. (1996). **Data Mining**. Addison Wesley, England.

Apté, C. e Weiss, S. (1997). *Data mining with decision trees e decision rules*, Future Generation Comp. Sys., 13: 197-210.

Bigus, J. P. (1996). **Data mining with neural network solving business problems from applications development to decision support**, McGraw-Hill, New York.

Chernoff, H. (1973). *The use of faces to represent points in k-dimensional space graphically*. Journal of the American Statistical association, vol 342, 68: 361–368.

Costa, J.A.F. (1999), **Classificação automática e análise de dados por redes neurais auto-organizáveis**, tese de doutorado, Unicamp, S.P.

Curram, S.P. e Mingers, J. (1994). *Neural networks, decision tree induction e discriminant analysis: An empirical comparison*, J. Operational Research Society, 45: 440-450.

Devedzic, V. (2001). *Knowledge Discovery e Data Mining in Databases*, in Handbook of Software Engineering e Knowledge Engineering.

Diamantaras, K.I. e Kung, S.Y. (1996), **Principal Component Neural Networks**, John Wiley, New York.

Diniz, C.A.R. e Louzada-Neto, F. (2000), Hair Jr., J.F., Anderson, R.E., Tatham, R.L. e Black, W.C. (2005). **Análise Multivariada de Dados**, Tradução, 5ª ed., Bookman, Porto Alegre.

Han, J. e Kamber, M. (2006). **Data Mining: Concepts e Techniques**, 2ª ed., Morgan Kaufmann, San Francisco.

Härdle, W. e Simar, L. (2003). **Applied Multivariate Statistical Analysis**. Springer-Verlag, Berlin

Haykin, S. (2009). **Neural Networks e Learning Machine**, Prentice Hall, New Jersey.

Data Mining: Uma Introdução, ABE - Associação Brasileira de Estatística, São Carlos – SP.

Everitt. B., Landau, S. e Leese, M. (2001). **Cluster Analysis**, 4ªth edition. Wiley, London.

Fayyad, U.M., Piatetsky-Shapiro, G.. e Smyth, P. (1996a). **From Data Mining to Knowledge Discovery in Databases**. AI Magazine 17(3): 37-54.

Fayyad, U.M., Piatetsky-Shapiro, G.., Smyth, P. e Uthurusamy, R. (1996b), *Advances In Knowledge Discovery e Data Mining*,AAI Press/ The MIT Press.

Field, A. (2009). **Descobrindo a Estatística Usando SPSS**, Tradução de Lorí Viali, Artmed, Porto Alegre.

Günsel, B., Tekalp, A.M. e Van Beek, P.J.P. (1998). *Content-based access to video objects: Temporal Segmentation, visual summarization, e feature extraction*, Signal Processing, vol 66, issue 2 (30): 261-280.

Hair Jr., J.F., Anderson, R.E., Tatham, R.L. e Black, W.C. (2005). **Análise Multivariada de Dados**, Tradução, 5ª ed., Bookman, Porto Alegre.

Han, J. e Kamber, M. (2006). **Data Mining: Concepts e Techniques**, 2ª ed., Morgan Kaufmann, San Francisco.

Härdle, W. e Simar, L. (2003). **Applied Multivariate Statistical Analysis**. Springer-Verlag, Berlin.

Haykin, S. (2009). **Neural Networks e Learning Machine**, Prentice Hall, New Jersey.

Hivärinen, A., Karhunen, J. e Oja, E. (2001), **Independent Component Analysis**, Jonh Wiley, New York.

Inmon, W. H. (1997). **Como construir um Data Warehouse**, Tradução de Ana M. N. Guz, 2ª ed., Campus, Rio de Janeiro.

Jain, A.K. e Dubes, R.C. (1988). **Algorithms for Clustering Data**, Prentice Hall, New Jersey.

Jain, A.K., Murty, M.N. e Flynn, P.J. (1999), *Data clustering: A review*, ACM Computing Surveys, 31:264–323.

Johnson, R.A. e Wichern, D.W. (1992). A**pplied Multivariate Statistical Analysis**, Prentice-Hall, New Jersey.

Liu, B., Hsu, W. e Ma, Y. (1998). *Integrating classification e association rule mining*. In Proc. 1998 Int. Conf. Knowledge Discovery e Data Mining (KDD'98), pages 80-86, New York.

Kaufman, L. e Rousseeuw, P.J.(1990), **Finding Groups in Data: An Introduction to Cluster Analysis,** Wiley, London.

Kimball, R. e Caserta, J. (2004). **The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, e delivering data,** Wiley, USA.

Kohonen, T. K. (1997). Self-Organizing Maps, 2nd extended edition, Springer-Verlag, Berlin, Heidelberg.

Metha, M.; Agraval R. e Rissanen, J. SLIQ: A Fast Scalable Classifier for Data Mining, IBM Almaden Research Center, 1996.

Michie, D., Spiegelhalter, D.J. e Taylor, C.C. (1994). **Machine Learning, Neural e Statistical Classification**, Ellis Horwood, 1994.

Piatetsky-Shapiro, G. (1991), *Knowledge Discovery in Real Databases*, A Report on the IJCAI-89 Workshop, AI Magazine 11(5): 68–70.

Principe, J. C., Euliano, N. R. e Lefebvre, W. C. (2000). Neural Adaptive Systems: Fundamentals Through Simulations, John Willey & Sons, New York, NY.

Rencher, A.C. (2002), **Methods of Multivariate Analysis**, 2ª ed., John Wiley e Sons, Canada.

Shafer, J.; Agraval, R.; Mehta, M. SPRINT: A scalable parallel classifier for data mining. In Proc. Of the 22nd VLDB Conference, 1996.

Silvanandam, S.N., Sumatri, S. e Deepa, S.N., (2007). Introduction to Fuzzy Logic Using Matlab. Berlin.

Van Trees, H.L. (1968), **Detection, Estimation, e Modulation Theory, Part I**, John Wiley e Sons, New York.

Witten, I.H. e Frank. E. (1999), **Data Mining: pratical machine learning tools e techniques with Java implementation**, São Francisco, California.

Zadeh. L.A., Klir, G.J., Yuan. B. (1996), **Fuzzy Sets, Fuzzy Logic, e Fuzzy Systems**, World Scientific Publishing, New Jersey.

Zuchini, M.H. (2003), **Aplicações de mapas auto-organizáveis em mineração de dados e recuperação de informações**, dissertação de mestrado, Unicamp, S.P.