TÉCNICAS ESTATÍSTICAS EM DATA MINING

Francisco Louzada-Neto

Carlos Alberto Ribeiro Diniz

Departamento de Estatística

Universidade Federal de São Carlos

CP 676, 13565-905, São Carlos, SP, Brasil

dfln@power.ufscar.br e dcad@power.ufscar.br

2002 - IMCA - Peru

Preface

Data Mining ou Mineração de Dados é um processo de extração de informação de grandes bancos de dados, o qual pode ser visto como uma nova disciplina na interface da estatística, do aprendizado de máquina, do reconhecimento de padrão e da tecnologia de bancos de dados.

Tanto a área acadêmica como vários setores da economia, como por exemplo, o financeiro, o comercial, o industrial e o de marketing, entre outros, podem ser beneficiados através da utilização do processo *data mining*.

Este manuscrito tem, como idéia central, a introdução ao tópico *data maning* do ponto de vista estatístico. Esperamos que ele possa auxiliar nas necessidades práticas de profissionais que trabalham com análise de grandes bancos de dados.

Um tratamento geral sobre o processo *data mining* está fora de contexto e alguns tópicos não são tratados no livro, o que gera a oportunidade para a pesquisa e preparação futura de material sobre o assunto. Assim, seremos gratos a sugestões e críticas que possam contribuir para um novo desenvolvimento do texto.

Esse trabalho foi parcialmente financiado pelo CNPq.

Francisco Louzada-Neto e Carlos Alberto Ribeiro Diniz
São Carlos, junho de 2002

Conteúdo

Capítulo 1

Introdução

Atualmente, com a disponibilidade de avançados recursos computacionais a baixo custo operacional, é extremamente fácil armazenar informações ou dados. Estima-se que a quantidade de informação no mundo dobra a cada 20 meses (Dilly, 1999) e que o tamanho e a quantidade dos bancos de dados crescem com uma velocidade ainda maior.

Entre os fatores que têm colaborado com o crescimento da quantidade de informação disponível em forma de dados estão os avanços na coleta automática de dados científicos, via sensores remotos e satélites; o processamento automático de transações através de códigos de barras; a instrumentação eletrônica e o processamento analítico *on-line* (OLAP).

As organizações, instituições e/ou empresas geram e coletam grandes volumes de dados que são usados ou obtidos em suas operações diárias. Os dados utilizados ou necessários em cada uma destas operações são registrados e mantidos pelo departamento correspondente.

Nos dias atuais, bancos de dados de vários gigabytes ou terabytes são comuns (Hand, 1998). Como exemplo podemos citar a empresa

telefônica americana AT&T, que concentra mais de 100 milhões de clientes e registra mais de 200 milhões de telefonemas em um único dia. A empresa varejista Wal-Mart, com mais de 20 milhões de transações diárias. A empresa UOL que chega a obter 50 gigabytes de dados diários através das transações de seus clientes internautas. A empresa Mobil Oil que tem mais de 100 terabytes de dados relacionados a exploração de óleo. O Sistema de Observação Terrestre da NASA que é projetado para gerar em torno de 50 gigabytes de dados por hora (Fayyad, *et.al.*, 1996).

Um terabyte de dados corresponde a estocagem de 2^{40} bytes de informação e em torno de 51 kilômetros de papel A4 densamente escrito (Potts, 1998).

Neste contexto, algumas perguntas básicas podem ser formuladas. Tendo acumulado uma grande quantidade de dados, o que fazer com eles? Como reverter estas informações em benefícios para a organização? Como os responsáveis pelas decisões podem identificar e utilizar as informações escondidas nos dados, de tal sorte que isto se reverta em vantagens em um curto período de tempo para a organização?

1.1 O Que é *Data Mining*?

Respostas referentes às questões formuladas acima podem ser dadas através de *Data Mining* ou Mineração de Dados. *Data mining* é o processo de extração de informações, sem conhecimento prévio, de um grande banco de dados, e seu uso para a tomada de decisões.

Na realidade, *data mining* é parte de um processo maior conhecido como *Knowledge Discovery in Databases (KDD),* ou Busca de Conhecimentos em Bancos de Dados. O processo *KDD*, que inclui os passos: estruturação do banco de dados, seleção de dados, preprocessamento, transformação e redução, *data mining*, análises,

assimilações, interpretações, avaliações e uso do conhecimento extraído, será discutido detalhadamente no Capítulo 4.

Como enfatizado por Dilly (1999), o termo *data mining* tem sido estendido além do seu limite para poder ser aplicado às mais variadas formas de análise de dados. Vários autores têm tecido definições, muitas vezes conflitantes, sobre *data mining*, o que acrescenta as dificuldades para uma definição única. Três definições, que simplificadamente descrevem o termo *data mining*, são dadas a seguir.

Data Mining é a busca por relações e características globais que estão "escondidas" em uma vasta quantidade de dados, tal como a relação entre dados de pacientes e seus diagnósticos médicos. Estas relações representam valiosos conhecimentos a respeito do banco de dados (Holshemier e Siebes, 1994).

Data Mining refere-se ao "uso de uma variedade de técnicas para identificar informações úteis em bancos de dados e a extração dessas informações de tal maneira que elas possam ser usadas em áreas tais como teoria de decisão, estimação, predição e previsão. Os bancos de dados são geralmente volumosos, e na forma que se encontram nenhum uso direto pode ser feito deles; as informações escondidas nos dados é que são realmente úteis " (Clementine User Guide, 2000).

Data Mining é a busca por informações valiosas em grandes volumes de dados. É um esforço cooperativo entre os seres humanos e os computadores. Os seres humanos planejam os bancos de dados. Os computadores pesquisam através dos dados, procurando por padrões que correspondam às metas (Weiss e Indurkhya, 1998).

Em resumo, *data mining* se relaciona com a análise de dados e o uso de ferramentas computacionais (*softwares*) na busca de características, regras e regularidades em um grande conjunto de dados. Obviamente é extremamente importante a escolha apropriada da ferramenta de *data mining* que será implementada. É esta ferramenta

ou conjunto de ferramentas que auxiliará na extração de características e busca de regras que não são perceptíveis ou não são óbvias nos bancos de dados.

1.2 A Interdisciplinariedade da Técnica

Data Mining é, também, uma área interdisciplinar que relaciona procedimentos científicos distintos, envolvendo técnicas estatísticas, de aprendizado de máquina, de reconhecimento de padrões e de visualização de dados (Cabena, *et.al.*, 1998).

Para a comunidade estatística, entretanto, o termo *data mining* tem uma conotação, de certa forma, pejorativa. Grosso modo, o que se descreve é análise exploratória de dados (Tukey, 1973) com uma certa sofisticação. A diferença reside no imenso tamanho dos bancos de dados.

Os estatísticos tipicamente não têm trabalhado com conjuntos de dados contendo bilhões de registros, que requerem formas especiais de estocagem e manipulação de dados. Como consequência, nota-se o surgimento e o crescimento de especialistas em bancos de dados que tratam desses problemas.

Desta forma, se por um lado o termo *data mining* pode ser relacionado a análise exploratória de dados, por outro, merece ser diferenciado dessa técnica por contemplar grandes conjuntos de dados (ver Capítulo 2).

1.3 Problemas Típicos de *Data Mining*

Existem vários problemas típicos de *data mining,* todavia, nesta seção, atenção será dada somente a alguns que parecem ser os mais importantes. Entre eles, os problemas de sumarização e visualização, segmentação ou agrupamento de bancos de dados, associações,

classificações e sequenciamentos.

Um dos objetivos básicos de *data mining* é a sumarização dos dados para fácil interpretação. Procedimentos descritivos estatísticos tradicionais podem ser empregados para atender tais requisitos. Em um contexto mais geral, técnicas de visualização de dados são imprescindíveis e muitas vezes responsáveis únicas pela descoberta de novas características nos bancos de dados.

Em muitas situações pode ser natural a segmentação ou agrupamento de um banco de dados em vários bancos de dados menores. Em geral, este procedimento é utilizado quando o banco de dados original pode ser decomposto em distintos subgrupos e espera-se que diferentes conclusões possam ser obtidas para cada um deles. Outro problema típico relaciona-se a associação entre diferentes itens presentes na mesma transação. Também pode existir a necessidade de sequenciamento de alguns procedimentos que são executados ao longo do tempo.

1.4 Potenciais Aplicações de *Data Mining*

Várias áreas do conhecimento podem ser beneficiadas através de *data mining*. Abaixo são descritas algumas potenciais aplicações.

Setor Bancário - Estudo do comportamento do uso de cartões de crédito para determinados grupos de clientes, detecç ão de cartões de crédito roubados, estudo do comportamento do uso de cartões de crédito roubados, identificação de clientes fiéis, detecção de correlações "escondidas" entre diferentes indicadores financeiros, *credit scoring* e *behaviour scoring* (relacionados à determinação do comportamento de clientes de risco).

Marketing e Comércio - Identificação do comportamento de compra dos clientes, predição de respostas para campanhas de marketing, determinação de associações entre itens comprados e entre

características demográficas dos clientes.

Seguros e Planos de Saúde - Predição de quais clientes ou grupos de clientes comprariam novas apólices de seguro, ou planos de saúde, identificação de clientes/pacientes de risco, identificação de clientes ou sinistros fraudulentos, verificação de quais procedimentos médicos e/ou odontológicos são utilizados conjuntamente.

Transporte - Análise do comportamento das cargas, determinação da distribuição dos cronogramas de entrega de mercadorias entre as diferentes distribuidoras, detecção de notas fiscais fraudulentas.

Medicina - Identificação de melhores terapias para diferentes doenças, caracterização do comportamento dos clientes para predição do horário das consultas, determinação de padrões em sequências de DNA, sequenciamento de genes.

Indústrias - Determinação da confiabilidade de produtos industriais complexos, confiabilidade de *software*, controle e gestão eficiente da qualidade.

Meio Ambiente - Determinação de risco relacionados a usinas nucleares, determinação do impacto ambiental de instalação de fábricas em uma determinada região, estudo de difusão de poluentes.

Internet - Identificação de determinados padrões em *home pages*, busca e agrupamento de documentos.

1.5 Descrição do Livro

A idéia básica deste livro consiste na introdução do processo *data mining* em uma linguagem acessível a analistas de dados, não necessariamente estatísticos e iniciantes no processo, ou alunos de graduação em Estatística. Para os não estatísticos, alguns capítulos mais específicos exigirão conhecimentos de análise multivariada, análise de regressão e redes neurais.

O livro é composto por 9 capítulos. O Capítulo 1 apresenta a introdução ao processo *data mining.* O Capítulo 2 relaciona *data mining* e Estatística. O Capítulo 3 descreve o armazenamento de dados. O Capítulo 4 detalha todos os passos de um processo *KDD.* Os Capítulos 5 a 8 tratam das técnicas de *data mining,* entre elas, classificação e regressão, análise de associação, análise de *cluster* e redução de dimensionalidade. O Capítulo 9 ilustra alguns exemplos envolvendo *data mining.*

Capítulo 2

Data Mining e Estatística

Frente a um novo procedimento de análise de dados, *data mining*, surge a necessidade de comparação com os conhecidos métodos tradicionais de análise, como é o caso da análise estatística. Neste capítulo as similaridades e diferenças entre essas duas metodologias, questionando as estratégias estatísticas tradicionais, são apresentadas.

2.1 Entendendo as Diferenças

Certa dificuldade na distinção entre *data mining* e análise estatística deve-se basicamente à similaridade entre elas e ao fato de que esse novo procedimento de análise é, geralmente, utilizado conjuntamente com os métodos estatísticos. Entretanto, essa dificuldade pode ser diluída se as técnicas de *data mining* forem diferenciadas, ou pelo menos entendidas como uma adaptação das técnicas estatísticas tradicionais, visando a análise de enormes bancos de dados.

O termo *data mining* parece não ser novo para muitos estatísticos e econometristas e tem sido utilizado para descrever o processo de pesquisa de conjunto de dados na esperança de identificar com-

portamentos ou características comuns. É examente esse ponto que leva muitos pesquisadores ao ceticismo.

Data dredging, *data snooping* e *fishing* podem ser vistos como sinônimos de *data mining*, e têm sido utilizados para nomear a extração de estruturas suspeitas e identificar padrões em conjuntos de dados (Hand, 1998 e Potts, 1998).

Apesar de *data mining* e análise estatística terem o mesmo objetivo: construção de modelos parcimoniosos e compreensíveis, que incorporem as dependências entre as descrições de uma determinada situação e os resultados destas descrições, *data mining* e análise estatística representam dois procedimentos diferentes para análise de dados.

Muitos estatísticos se preocupam com a análise *primária* de dados. Isto é, os dados são coletados com uma questão particular ou um conjunto de questões particulares a priori, e que podem ser traduzidas em forma de hipóteses a serem testadas. Planejamento de experimentos e amostragem são dois exemplos de áreas que tiveram o seu desenvolvimento voltado a propiciar facilidade para responder tais questões.

Por outro lado, *data mining* objetiva não só a análise *primária*, mas também a análise *secundária* dos dados, isto é, minerar o conjunto de dados à procura de relacionamentos não suspeitos, que são de interesse para o pesquisador ou para a organização proprietária do banco de dados.

Enquanto a análise estatística tem como base um procedimento hipotético-dedutivo, *data mining* é, além disso, um processo indutivo (Hand, 1998).

Como exemplo, assuma que um empresário do setor alimentício deseje conhecer a quantidade de arroz vendida em um de seus supermercados. Ele poderia utilizar um pacote gráfico para representar as vendas de arroz em cada um dos seus supermercados em

um determinado período de tempo. Caso necessário, poderia-se utilizar uma regressão para analisar a sensitividade dos preços de uma marca de café nos diferentes supermercados ou comparar as vendas dos vários produtos nos diferentes supermercados utilizando análise multivariada.

Estes cenários, apesar de diferentes, têm algo em comum: todos envolvem hipóteses, tais como, existência de produtos, vendas de produtos, diferentes supermercados. O empresário teria, então, como objetivo determinar o relacionamento entre essas quantidades. Embora útil e essencial, esses procedimentos são clássicos e conhecidos como análises estatísticas de dados. É possível também hipotetizar subjetivamente sobre a venda de um determinado tipo de produto e incorporar essa informação na análise, utilizando conceitos Bayesianos (Box and Tiao, 1974). Novamente uma hipótese estaria presente.

Data mining pode ser visto como o descendente direto da estatística e surge 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. Considerando ainda o exemplo acima, *data mining* permitiria ao empresário obter respostas para questões, em certo sentido, mais sofisticadas como: *Que tipo de clientes poderiam ser atraidos para a compra de um determinado tipo de produto? Por que as promoções não geraram a rentabilidade esperada?*

2.2 Questionando Estratégias

Usualmente, os procedimentos estatísticos de análise são fomentados por bancos de dados pequenos e *limpos* que são coletados para responder uma certa quantidade de questões particulares, e, em geral, são constituídos de variáveis numéricas. As respostas são, geral-

mente, estáticas, independentes, identicamente distribuídas e são obtidas de forma direta. Entretanto, estas situações não se aplicam no contexto de *data mining*.

Hoje, muitos bancos de dados são constituídos por uma enorme quantidade de registros, onde algumas variáveis podem estar *contaminadas*, não ser independentes e identicamente distribuídas , não apresentar estacionariedade e apresentar vício, entre outros.

2.2.1 O Tamanho das Bases de Dados

Os tamanhos dos bancos de dados que estão presentes no contexto de *data mining* geram problemas que não têm sido considerados pela maioria dos pesquisadores da comunidade estatística.

Métodos de estimação sequencial e adaptativos precisam ser desenvolvidos, uma vez que, apesar do dramático crescimento em capacidade de memória computacional experimentada nos últimos anos, a base de dados pode não caber na memória do computador. Esse tipo de preocupação tem sido, de certa forma, enfocada por especialistas em bancos de dados, reconhecimento de padrões e aprendizado de máquina.

Outro problema relaciona-se à forma de estocagem da base de dados. É comum conjuntos de dados compostos de vários subconjuntos que podem inclusive estar fisicamente separados ou terem uma estrutura hierárquica que não permite fácil acesso à base de dados total, tornando, desta forma, a amostragem um processo complicado e demorado. Consequentemente, a estrutura da base de dados pode implicar na impossibilidade de aplicação de simples técnicas estatísticas e, nestes casos, variantes baseadas em estratificação e agrupamento serão necessárias.

Quando existem dados em abundância, as estratégias tradicionais de análises estatísticas, baseadas na fixação do *erro do tipo I* de um

teste de hipóteses, ficam comprometidas. Os resultados de tais testes irão sempre indicar uma forte evidência de que efeitos, mesmo os extremamente pequenos, existem. Dessa forma, *significância* do ponto de vista estatístico torna-se irrelevante ou, pelo menos, prejudicada, e a *significância* subjetiva deste efeito será de maior valia.

2.2.2 Dados Contaminados

Um pré-requisito exigido pela maioria das técnicas estatísticas é a *limpeza* dos dados, que pode ser traduzido basicamente em dois tópicos: detecção de *outliers* e dados *missing*.

Existindo algum questionamento sobre a qualidade dos dados, uma possibilidade de tratamento consiste em verificar a origem dos registros em busca de possíveis explicações. Isso é possível quando a preocupação do pesquisador se concentra na análise primária dos dados, mas impossível ou questionável no contexto de *data mining*, onde a preocupação se concentra, também, na análise secundária dos mesmos. Além disso, outros problemas podem aparecer quando existem dados em grande quantidade. Neste caso, certamente, uma porção dos dados será inválida.

2.2.3 Variáveis IID

Excluindo alguns novos modelos para medidas repetidas, um outro pré-requisito exigido pela maioria das técnicas de análise estatísticas é a suposição de que os dados foram amostrados independentemente e da mesma distribuição — variáveis independentes e identicamente distribuídas (iid). Isto não é a regra no contexto *data mining.* Em períodos diferentes é possível que algumas regiões do espaço de uma variável sejam amostradas com mais intensidade do que outras, induzindo, assim, dúvida à validade de estimativas padrões.

2.2.4 Não Estacionariedade

Chamado por alguns autores por *population drift* (Taylor *et. al.* 1997; Hand, 1998), problemas de não estacionariedade podem aparecer quando a população em estudo modifica-se rapidamente. Este problema pode se intensificar no contexto de *data mining,* pois grandes bases de dados podem ser dinâmicas. Transações com cartões de crédito ocorrem todos os dias durante vinte e quatro horas. Os resultados obtidos em dezembro, levando-se em consideração o que ocorreu em setembro, podem ser irrelevantes para a organização e, assim, processamento dos dados em tempo real pode ser necessário.

2.2.5 Presença de Vício

Em geral, grandes bases de dados são amostras *convenientes* ao invés de amostras aleatórias, como o ideal para muitas técnicas estatísticas. Neste caso, inferências sobre a verdadeira população estarão comprometidas.

Considere uma instituiçao financeira que oferece empréstimos para uma quantidade de clientes de acordo com uma regra prévia. Uma quantidade detalhada de informação estará disponível somente para os indivíduos que já aceitaram o oferecimento. Se existir interesse em inferências sobre o comportamento de futuros clientes de cartão de crédito, baseadas nesta base de dados, erros possivelmente serão introduzidos.

Uma possível solução seria a proposição de um modelo que acomodasse o mecanismo de seleção amostral, o que pode ser difícil.

2.2.6 Variáveis Não Numéricas

Em problemas de *data minig* é comum a existência de diferentes tipos de dados, incluindo dados não numéricos. Entre os conjuntos de dados não usuais encontram-se dados de imagens, áudio, texto e

dados geográficos. Novas técnicas que acomodem estas situações são necessárias.

2.3 Comentários Finais

Neste capítulo algumas similaridades e diferenças entre *data mining* e estatística foram discutidas. Também, questionamentos sobre as estratégias estatísticas tradicionais aplicadas a grandes bancos de dados foram apresentadas.

Excluindo-se os problemas inerentes ao tamanho dos bancos de dados, tratados em *data mining*, os procedimentos utilizados são familiares a estatísticos e econometristas. Entretanto, a validade da vários procedimentos padrões de inferência podem ser violados quando o conjunto de dados tratado é muito grande.

Em comparação às técnicas estatísticas, *data mining* herda a dependência de um banco de dados bem documentado e *limpo*, o que pode, eventualmente, ser considerado uma desvantagem, dado o tamanho das bases de dados que nos referimos.

As promessas e oportunidades comerciais relacionadas a *data mining* são enormes, entretanto, a técnica em si parece não ser um "remédio para todos os males " (Potts, 1998). Enquanto grandes conjuntos de dados têm potencialidade para produção de melhores resultados, não existe garantia de que estes produzam resultados melhores do que os encontrados através de conjuntos de dados menores (Weiss e Indurkhya, 1998).

Uma estratégia ótima de análise de dados possivelmente consideraria ambas metodologias, estatística e *data mining,* como técnicas complementares.

Capítulo 3

Importância do Banco de Dados

Entre as principais características do processo *data mining* se encontram o tamanho dos bancos de dados e o tratamento e manuseio dos mesmos de uma forma sistemática. Assim, é importante a discussão envolvendo a descrição de um banco de dados, o que vem a ser uma característica associada a um banco de dados e a necessidade de um sistema gerenciador. Além da descrição de termos como *data warehouse* e *OLAP* (processamento analítico *on-line*) e seus relacionamentos com *data mining*.

3.1 Características em um Banco de Dados

Uma característica em uma tabela de dados é definida como sendo um conjunto de linhas que compartilham o mesmo valor em duas ou mais colunas. Considere, por exemplo, a Tabela 1 que contém dados sobre clientes de uma instituição financeira, incluindo o sexo, se é ou não proprietário de veículo e o tempo que é cliente. Na tabela, as

linhas correspondem aos diferentes clientes.

.

Tabela 1 - Dados de Clientes de um Banco

Linha #	Sexo	Veículo	Tempo em anos
1	F	Sim	5
2	F	Sim	4
3	F	Sim	5
4	F	Não	1
5	M	Não	2

Nesta tabela temos três linhas (linhas 1, 2 e 3) que compartilham dos mesmos valores em duas colunas (sexo e proprietário de veículo). A **confiança (**não no sentido estatístico) com a qual se refere a este fato é de 75%. A confiança aqui é facilmente calculada dividindo-se o número de linhas que têm F e Sim pelo número de linhas que têm F. Pode-se observar também que: *A maioria dos clientes é do sexo feminino.*

A palavra "maioria" é usada aqui simplesmente pelo fato que a confiança é maior que 50%. Se a confiança for 100% usa-se o termo "todos". Por exemplo, *todos os clientes são mulheres.* Por outro lado, se a confiança for menor que 50% pode-se, ainda, apresentar a característica usando a porcentagem. Por exemplo, "*40% dos clientes são mulheres".*

Outro parâmetro usado para diferenciar características é o número de linhas que confirmam a característica. Por exemplo, o fato "*A maioria dos clientes é do sexo feminino"* é sustentado por três linhas. Uma característica que é confirmada por um grande número

de linhas é mais *poderosa* que uma corroborada por um pequeno número de linhas (característica *fraca*).

Uma vez encontradas as características, pode-se também procurar por **exceções.** Por exemplo, a linha 4 representa uma exceção ou pode ser um erro. Esta é a única linha onde uma cliente não possui automóvel. Esta exceção torna-se muito mais interessante nos casos onde, digamos, de 1000 clientes do sexo feminino, 999 delas possuem automóveis e uma não. Uma exceção é definida como uma linha de característica fraca.

Em *data mining* **fatos e regras** são casos especiais de características ou são características com confianças maiores que 50%. A diferença entre fatos e regras reside apenas na apresentação. As duas frases a seguir representam a mesma característica:

A maioria dos clientes é do sexo feminino.

Se cliente = feminino então Veículo = Sim.

A primeira representa um fato, que é mais adequado para fazer observações e análises típicas de aplicações de *data mining*, enquanto a segunda representa uma regra que é orientada para situações de condição-ação utilizadas em sistemas inteligentes.

3.2 *Data Warehouse*

Data warehouse, ou depósito de dados, é um sistema de gerenciamento de banco de dados relacional, desenvolvido especificamente para as necessidades de sistemas de processamentos de transações e por esta razão tem uma associação muito forte com *data mining.* As possibilidades de mineração de dados podem ser intensificadas se os dados estiverem armazenados em um *data warehouse.*

Data warehouse é uma poderosa técnica, capaz de extrair dados operacionais arquivados e superar as inconsistências entre diferentes

formatos de dados. Além de integrar dados de toda uma empresa, independente da localização, formato ou condições de comunicação, a técnica permite, também, a incorporação de informações adicionais.

Inmon (1992) apresenta uma outra definição bem aceita de *data warehouse:* é uma coleção de dados com quatro características**: tópico-orientado**, **integrado**, **tempo-variante** e **não-volátil.**

Tópico-orientado - Os dados são definidos e organizados em assuntos de negócios, em vez de aplicações. Por exemplo, um empresa de seguros usando um *data warehouse* organizaria seus dados por consumidor, prêmios e sinistros, em vez de organizá-los por diferentes produtos (seguros de automóveis, de vida, residenciais etc). Os dados organizados por tópicos contêm somente a informação necessária para apoiar o processo de decisão.

Integrado - Quando os dados residem em várias aplicações separadas no ambiente operacional é provável que exista uma codificação inconsistente dos mesmos. Por exemplo, em uma aplicação, o sexo do cliente pode ser codificado como "f" e "m", em outra como "0" e "1". Quando os dados são transferidos de ambiente operacional para o *data warehouse*, eles assumem um código convencional consistente, isto é, os dados são transformados em "f" e "m", digamos.

Tempo-variante - *Data warehouse* contém um espaço para armazenar dados antigos (de 5 a 10 anos, por exemplo), que podem ser usados em comparações, tendências e previsões. Estes dados não são atualizados.

Não-volátil - Uma vez que os dados entram no *data warehouse* não são atualizados ou mudados, são somente carregados ou acessados.

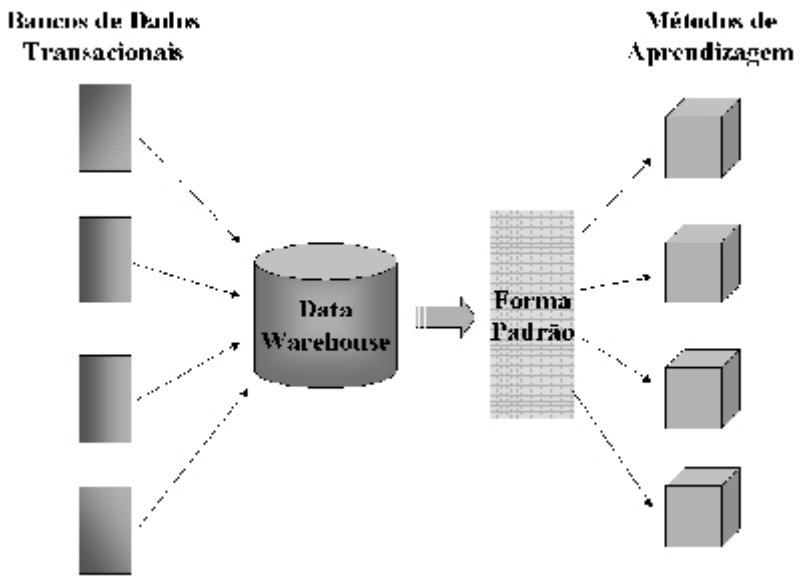

Figura 3.1: Modelo de *data warehouse* para *data mining*.

3.3 *Data Warehouse* para *Data Mining*

Um modelo de armazenamento de dados para *data mining* é ilustrado na Figura 3.1. Geralmente, bancos de dados em grandes organizações estão pulverizados entre as várias unidades subsidiárias, onde cada uma destas unidades registra, potencialmente, um número maciço de transações. Nestes casos, *data warehouse* deve ser desenvolvido de maneira que os dados sejam extraídos destes vários bancos de dados operacionais e registrados de uma forma centralizada. A idéia principal é disponibilizar a informação de tal forma que possa ser usada para futuros processamentos analíticos e tomadas de decisões.

Os dados armazenados nos bancos de dados não estão necessariamente na forma ideal para a mineração de dados. Existe um grande salto para que bancos de dados transacionais configurem-se em um recurso centralizado, isto é, um *data warehouse*, onde os dados de apoio às tomadas de decisões estão armazenados. Um número de tarefas, ilustradas na Figura 3.2, são usualmente executadas para transferir dados de fontes dispersivas para um "depósito" centralizado.

As tarefas incluem extração, transformação, limpeza e integração.

Extração - Dados são extraídos de diferentes fontes em diferentes formatos;

Transformação - Dados brutos são transformados em dados mais
qualificados para apoio à decisão. Por exemplo, cada venda pode ser registrada, mas um resumo de vendas por dia pode ser a transformação mais adequada a ser armazenada em um *data warehouse;*

Limpeza - Os campos de dados são verificados procurando-se por inconsistências ou por valores faltantes. Registros errados são solucionados ou eliminados.

Integração - Dados de múltiplos bancos de dados e outras fontes são integrados em um *warehouse* central.

Ter um *data warehouse* não resolve necessariamente todos os pontos pertinentes a preparação de dados. Para uma apresentação uniforme e padronizada, compatível com *data mining*, os dados extraídos do *data warehouse* podem precisar de transformações adicionais.

3.4 OLAP - Processo Analítico *On-line*

Um problema importante relacionado com o processamento de enormes bancos de dados, de complexidade crescente, diz respeito ao

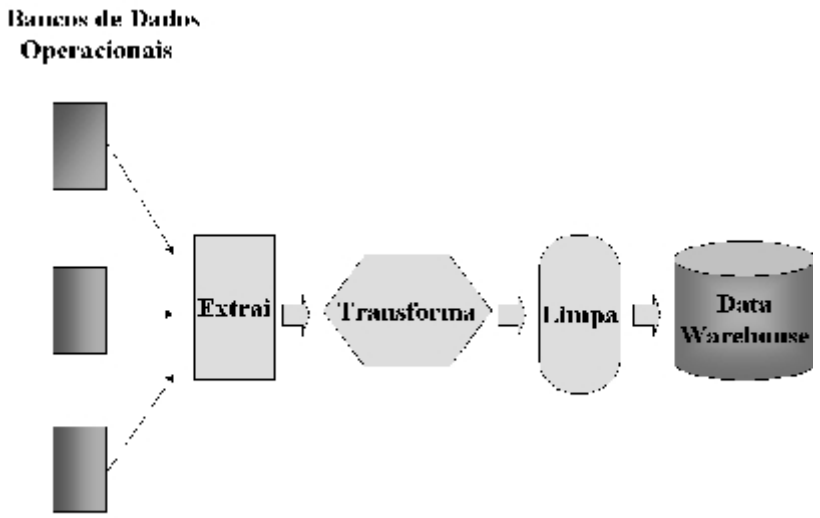

Figura 3.2: Movendo Dados Operacionais para o *data warehouse*

processamento de informação sem sacrificar o tempo de resposta. Uma arquitetura computacional capaz de minimizar este problema é conhecida como *On-line Analytical Processing* (OLAP) e pode ser vista como uma extensão de um *data warehouse.* O termo OLAP foi criado por E.F. Codd em 1993 e definido como a síntese, análise e consolidação dinâmica de grandes volumes de dados multidimencionais.

A principal característica de uma arquitetura OLAP é ser *on-line,* de tal forma que o sistema possa acessar grandes quantidades de dados, promover a análise das relações entre os diferentes tipos de variáveis, agregar os dados de forma adequada para análise, apresentar os dados em diferentes perspectivas, e responder rapidamente

às perguntas do usuário (Dilly, 1998 e Inmon, 1996).

Para ilustrar, considere um banco de dados OLAP composto de vendas que foram agregadas por região, tipo de produto e formas de venda. O sistema OLAP deve estar apto a pesquisar multi-gigabytes de dados de todas as vendas dos produtos em cada região, para cada tipo de produto. Refinando a análise, o sistema deve estar apto a determinar o volume de vendas para cada forma de venda, dentro das classificações em produtos por regiões. Refinando ainda mais a análise, pode-se fazer uma comparação ano a ano das vendas para diferentes formas de vendas. Esse processo de pesquisa no banco de dados deve ser feita *on-line* com tempo de resposta rápido (Dilly, 1998).

3.5 Comentários Finais

As arquiteturas de bancos de dados descritas neste capítulo servem como suporte na aplicação de *data mining*, facilitando o emprego das técnicas relacionadas a ele. A não existência de um *warehouse* pode implicar em muitas dificuldades quando da tentativa de extração de informações de grandes massas de dados, particularmente se essa informação tiver de ser extraída de múltiplos bancos de dados separados fisicamente.

Para não comprometer o tempo de resposta usualmente necessário em tomadas de decisões, um sistema OLAP deve ser utilizado.

Na verdade, pode-se dizer que apesar de terem como objetivo comum, o auxílio na aplicação de *data mining*, OLAP e *data warehouse* são ambientes diferentes. Em geral, um sistema OLAP contém dados padronizados, enquanto um *data warehouse* é mais genérico.

Capítulo 4

O Processo KDD

A busca por características em um banco de dados é somente um passo no processo KDD. A Figura 4.1 ilustra todos os passos que devem ser executados para que um processo KDD esteja completamente identificado. Embora os passos devam ser executados na ordem em que são apresentados, o processo é extremamente interativo e iterativo (com várias decisões sendo feitas pelo próprio usuário e *loops* podendo ocorrer entre quaisquer dois ou mais passos). É importante ressaltar que o processo é inteiramente controlado pelos objetivos da empresa e somente através de um problema presente na empresa é que teremos a base para que um projeto *data mining* se estabeleça.

Como descrito na Figura 4.1, a sequência de passos é: Banco de Dados ⇒ Seleção (Dados Selecionados) ⇒ Preprocessamento (Dados Preprocessados) ⇒ Transformação (Dados Transformados) ⇒ Data Mining (Características, Informações Extraídas) ⇒ Análises, Assimilações, Interpretações e Avaliações (Conhecimentos Assimilados)

Os passos não compartilham do mesmo peso em termos de tempo e esforços consumidos. A preparação dos dados, por exemplo, que envolve a *seleção, preprocessamento* e *transformação dos dados*, ne-

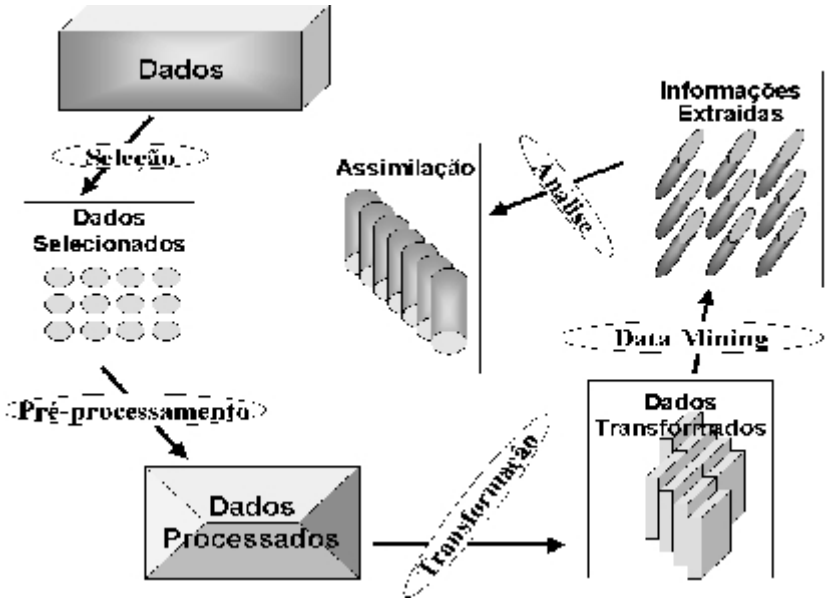

Figura 4.1: Passos de um processo KDD.

cessita entre 60 e 80% do tempo utilizado em todo o processo, com a maior parte do tempo consumido na "limpeza" dos dados.

Cada passo no processo KDD será discutido de forma detalhada nas seções seguintes. O passo *data mining* entretanto, por ser o principal tópico deste livro e a novidade deste processo, será tratado, de forma destacada, nos próximos capítulos.

4.1 Seleção dos Dados

O objetivo principal do passo *seleção dos dados* é, na verdade, identificar a origem interna e externa da informação e extrair o subconjunto de dados necessário (selecionar as variáveis de interesse) para a

aplicação de *data mining*. Certamente a seleção de dados vai variar de acordo com os objetivos da empresa. As variáveis selecionadas podem ser do tipo **categórica** ou do tipo **quantitativa.**

As variáveis categóricas assumem valores finitos, diferem na forma e podem ser nominais ou ordinais. Ao contrário da variável nominal, existe uma ordem entre os possíveis valores de uma va- riável categórica ordinal. Exemplos de variáveis nominais são estado civil (solteiro, casado, divorciado, desconhecido) e sexo (masculino, feminino). Exemplos de variáveis ordinais são graus de instrução (primeiro grau, segundo grau, superior) e escore de crédito pessoal (ruim, regular, bom).

As variáveis quantitativas assumem valores numéricos e podem ser do tipo contínua (os possíveis valores são números reais) ou discreta (os possíveis valores fazem parte de um conjunto finito ou infinito enumerável). Exemplos de variáveis contínuas são receita e taxa. Exemplos de variáveis discretas são número de empregados e número de filhos.

As variáveis selecionadas para *data mining* são denominadas *variáveis ativas* uma vez que elas são ativamente usadas para distinguir segmentos, fazer predições ou desenvolver outras operações específicas de *data mining*.

No passo de seleção de dados é importante considerarmos o fato que mudanças em andamento em circunstâncias externas podem afetar a eficiência da mineração. Por exemplo, é comum que mudanças de emprego ou de cargo ocorram em uma porcentagem dos clientes a cada ano. Qualquer análise onde o tipo de emprego é um fator deve ser re-examinada periodicamente. Similarmente, variáveis demográficas ou de estilo de vida costumam ter esperanças de vida curtas.

4.2 Preprocessamento dos Dados

Como em qualquer análise quantitativa, a qualidade dos dados é essencial para a obtenção de resultados confiáveis. Portanto, dados limpos e compreensíveis são requisitos básicos para o sucesso da mineração dos dados.

O preprocessamento dos dados tem como objetivo assegurar a qualidade dos dados selecionados. Esta fase inicia-se com uma revisão geral da estrutura dos dados e algumas medidas de sua qualidade. Isto pode ser feito utilizando-se uma combinação de métodos estatísticos e técnicas de visualização de dados (ver Capítulo 10).

Na situação em que temos um grande volume de dados selecionados, podemos considerar apenas uma amostra representativa dos mesmos no processo de revisão. Uma maneira simples de entender o conteúdo dos dados, no caso de variáveis categóricas, é através da distribuição de frequência dos valores e/ou através de ferramentas gráficas tais como histogramas e diagrama de setores. No caso de variáveis quantitativas, uma forma de determinar a presença de valores inválidos é através do cálculo de medidas estatísticas tais como, mínimo, máximo, média, moda, mediana e desvio padrão amostral. Analisando os valores mínimo e máximo pode-se detectar rapidamente valores espúrios nos dados e, atráves das diferentes medidas de tendência central e de dispersão, pode-se verificar o grau de ruído nos dados.

O *boxplot* e o diagrama de dispersão são ferramentas gráficas extremamente úteis no caso de variáveis quantitativas. Os *boxplots* podem ser utilizados para comparação da média ou do desvio de duas ou mais variáveis, enquanto o diagrama de dispersão é um gráfico simples bi-dimensional que representa a relação entre duas variáveis contínuas. Ambas ferramentas gráficas resumem, de forma rápida e eficiente, a estrutura e a qualidade dos dados. Maiores detalhes

sobre esses dois tipos de representações gráficas serão apresentados no Capítulo 10.

Dados com erros e valores faltantes (*missing)* são dois problemas que naturalmente são resolvidos no passo pré-processamento dos dados.

4.2.1 Dados com Erros

Valores que são significantemente fora do esperado para uma ou mais variáveis podem ser devido a erros, ou *outliers*. Os *outliers* podem indicar uma boa ou má notícia. Uma boa notícia se indicarem uma nova tendência de resultados para as variáveis em questão e uma má notícia se realmente forem apenas dados inválidos.

Diferentes tipos de *outliers* devem ser tratados de forma diferente. Um tipo comum diz respeito ao resultado de um erro humano, como um registro de compra em supermercado ser da ordem de milhões de reais ou o registro de idade de uma pessoa ser maior que 200 anos ou o valor de uma receita ser negativo. Estes registros devem ser corrigidos se valores razoáveis ou válidos estão disponíveis, caso contrário, estes registros devem ser excluídos da análise.

Outro tipo de *outlier* é criado quando alguma mudança no sistema operacional ainda não tenha sido refletida no ambiente da mineração dos dados. Por exemplo, novos códigos de produtos, que aparecerão no ambiente como sendo *outliers*. Nestes casos a atualização dos sistema deve se feita.

4.2.2 Valores *Missing*

Valores missing incluem os valores que simplesmente não estão presentes no conjunto selecionado e os valores inválidos que foram eliminados durante a detecção de *outliers.* Os valores *missing* podem ocorrer devido a erros humanos,ou porque a informação não está

disponível no momento do levantamento dos dados, ou quando os dados são selecionados considerando-se diferentes origens, gerando informações contraditórias.

Uma forma de tratamento de valores *missing* é simplesmente eliminar toda a linha (ou coluna) de observações que contenha valores faltantes. Este procedimento é simples mas tem como consequência a perda de informação. Embora esta perda não seja um problema nas situações onde o volume de dados é grande, certamente trará algum efeito nos casos onde o volume de dados é pequeno ou onde os objetivos se concentram em fraude ou controle de qualidade. Se existir um grande número de valores *missing* para uma determinada variável, a saída, talvez, seria retirar a variável da análise, o que, novamente poderia ter sérias consequências se esta variável é chave para a solução do problema.

Portanto, a decisão de eliminar observações ou variáveis não é fácil e não se pode predizer as consequências de tais atitudes. Felizmente existem várias técnicas que permitem substituir os valores *missing*. Para variáveis quantitativas, a mais simples é o uso da média ou da moda. Para variáveis categóricas, pode-se utilizar a moda ou um novo atributo criado para a variável, como por exemplo, usar a denotação *Desconhecido*. Técnicas mais sofisticadas, para ambos os tipos de variáveis, como modelos de predição e técnicas de imputações, estão também disponíveis (Hair, *et.al.*, 1998).

4.2.3 Sumarização

A sumarização é responsável pela descrição compactada de um conjunto de dados. É utilizada, principalmente, no pré-processamento dos dados, onde valores inválidos, no caso de variáveis quantitativas, são determinados através do cálculo de medidas estatísticas tais como mínimo, máximo, média, moda, mediana e desvio padrão amostral e,

no caso de variáveis categóricas, através da distribuição de frequência dos valores.

Outras técnicas de sumarização mais sofisticadas são as chamadas técnicas de visualização. Visualização de dados tem sido parte integrante da análise estatística de dados e foi primeiramente apresentada nos trabalhos de Tukey (1973), como parte da análise exploratória dos dados. Estas técnicas são de extrema importância para se obter um entendimento, muitas vezes intuitivo, do conjunto de dados. Se o volume de dados é grande, que é o caso dos conjuntos de dados presentes em *data mining*, técnicas de visualização tornam-se imprescindíveis.

Tabela 4.1 - Representação tabular de chamadas telefônicas

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

As diferentes medidas estatísticas descritas acima são facilmente encontradas em qualquer livro básico de Estatística e já devem ser familiares aos leitores. A novidade, portanto, na sumarização de um grande conjunto de dados, reside no uso de algumas técnicas de visualização. É imprecindível a utilização de sistemas específicos em visual *data mining* para que a visualização esperada seja alcançada.

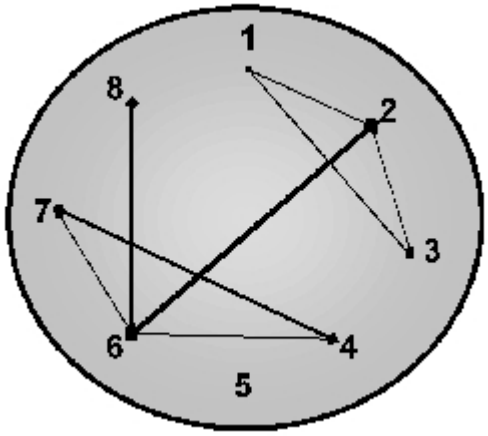

Figura 4.2: Diagrama de associação das chamadas telefônicas.

4.3 Técnicas de Visualização

Técnicas de visualização compreendem uma poderosa ferramenta que pode ser utilizada de forma rotineira para sumarizar grandes quantidades de dados.

Estas técnicas são, em muitas situações, suficientes para a extração das respostas de interesse, descobrindo padrões, tendências, estruturas e relações dentro de um conjunto de dados. Entretanto, é importante salientar que as técnicas de visualização devem ser aplicadas em combinação com técnicas analíticas, com o intuito de extração de resultados interpretáveis.

O método de visualização escolhido para a análise dependerá basicamente do tipo de conjunto de dados disponível e como estes dados podem ser modelados. Se, por exemplo, o conjunto de dados envolve chamadas telefônicas feitas em um específico intervalo de tempo,

como apresentado na Tabela 4.1, então uma representação visual desta informação poderia ser sumarizada através de um simples diagrama de associação, como mostrado na Figura 4.2, disponibilizando todas as relações entre as chamadas.

Pelo diagrama de associação é possível ver imediatamente que existem várias chamadas entre certos pares de telefones e poucas ou nenhuma entre outros. As linhas mais grossas no diagrama denotam números maiores de chamadas. A vantagem da visualização destes dados através do diagrama de associação em relação a representação tabular é clara. Examinando o diagrama é possível detectar rapidamente quais números merecem uma análise mais detalhada, enquanto que, no formato tabular, adicionais cálculos de frequências e ocorrências de associações com outros números seriam necessários para obter a mesma informação.

Uma outra forma de representação para esses dados seria por meio da matriz de associação. A cada linha e a cada coluna da matriz associa-se um número de telefone. Os elementos da matriz correspondem ao número de vezes que os telefones correspondentes foram conectados. A linha e a coluna externa a matriz mostram, respectivamente, o número total de chamadas recebidas e feitas por cada telefone. Porém, quando o número de telefones é grande a matriz de associação torna-se inviável e, nestes casos, é sugerido o uso do diagrama de associação.

Matriz de Associação.

		Para							
		1	2	3	4	5	6	7	8	
	1		1	1						2
	2						3			3
	3		1							1
De	4						1	1		2
	5									0
	6		2					1		3
	7				1	1				2
	8						3			3
		0	4	1	1	1	7	2	0	

Outros tipos de diagramas (ou gráficos), além dos diagramas de associação, que são utilizados para visualização de dados (Parsaye *et. al.* 1993) podem ser categorizados como: mapas, diagramas baseados em coordenadas, baseados em proporções, híbridos, icônicos e diagramas hierárquicos. Além disso, existem os diagramas especializados em cada domínio, como, por exemplo, o mostrado na Figura 4.2.

Figura 4.3: Exemplo de um diagrama do tipo mapa.

4.3.1 Mapas

Mapas ou panoramas são usados para representar informações de posicionamento espacial. Os objetos nestes casos são usualmente modelados para incluírem informações tais como latitude/longitude ou outras coordenadas que auxiliem a posicionar o objeto dentro do panorama. É comum mapas serem usados em combinação com outros tipos de gráficos, tal como, diagramas de barras (ver Figura 4.3).

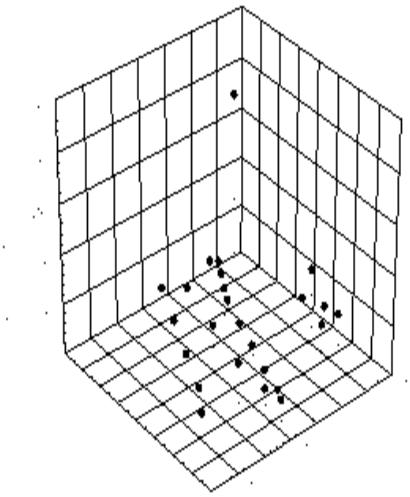

Figura 4.4: Sistema de coordenadas de um diagrama tri-dimensional.

4.3.2 Diagramas Baseados em Coordenadas

Os diagramas baseados em coordenadas referem-se a um sistema de coordenadas, onde cada eixo representa diferentes atributos. Os valores de cada atributo são representados ao longo dos eixos. Informações simultâneas sobre os valores relativos das variáveis podem ser obtidas projetando cada ponto em um eixo ortogonal. Variantes deste tipo de diagrama são os diagramas de dispersão, como mostrado na Figura 4.4, e os diagramas de linhas, úteis na detecção de tendências.

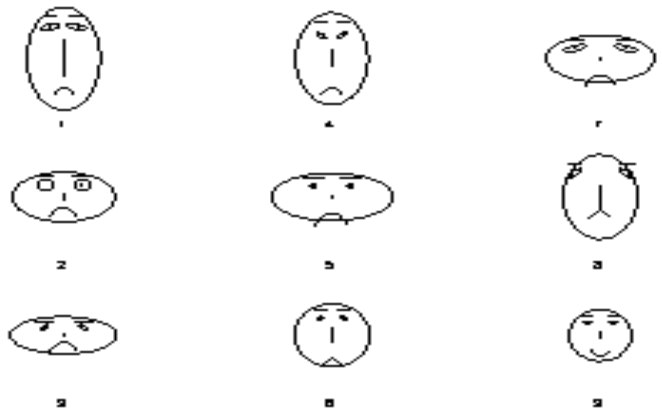

Figura 4.5: Um exemplo de faces de Chernoff.

4.3.3 Diagramas Baseados em Proporções

Os diagramas baseados em proporções enfatizam dados de proporções. Entre as variantes deste tipo de diagrama existem os diagramas setoriais e os perfis multivariados. Cada segmento do diagrama setorial corresponde ao valor da proporção de um atributo. Os diagramas perfis multivariados, que são na realidade múltiplos diagramas setoriais, são utilizados para fornecer visualização de dados multivariados. Uma outra forma de representar os perfis seria através de diagramas de barras.

Livros básicos de Estatística apresentam uma série de ilustrações de diagramas baseados em proporções, ver por exemplo, Bussab e Morettin (1985).

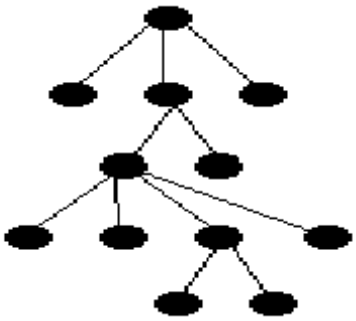

Figura 4.6: Diagrama hierárquico.

4.3.4 Diagramas Híbridos

Os diagramas híbridos são uma mistura de diagramas baseados em coordenadas e os baseados em proporções. Entre os diferentes diagramas híbridos estão os diagramas de barras, os histogramas e os *boxplots*. Diagramas de barras são considerados híbridos quando uma variável é representada por um eixo numérico e o outro atributo é representado pelos níveis da categoria. Quando o diagrama de barras é usado para representar a distribuição de uma variável, fazendo com que as barras fiquem bastante fina e numerosa, o diagrama é referido como histograma. *Boxplot* é usado para visualizar a distribuição de uma variável segundo os valores de uma outra variável (*boxplot* bi-dimensional) ou duas variáveis (*boxplot* tri-dimensional). Exemplos destes diagramas podem, também, ser encontrados em livros básicos de Estatística.

4.3.5 Diagramas Icônicos

Os diagramas icônicos são caracterizados pela presença de glifos que são usados para descrever os valores de múltiplos atributos. As faces de Chernoff (ver Figura 4.5) são um tipo especial de glifos que utiliza característica facial humana para representar graficamente um conjunto de dados.

4.3.6 Diagramas Hierárquicos

Quando a posição de um objeto está baseada na sua relação com outro objeto forma-se uma hierarquia na apresentação. Um exemplo de apresentação hierárquica é mostrado na Figura 4.6. Esta forma de diagrama é útil para representar um conjunto de dados com classes de objetos hierárquicas dentro de outras.

4.4 Transformação de Dados

Um dos objetivos principais da transformação de dados é converter o conjunto bruto de dados em uma forma padrão de uso. Técnicas como discretização (converter variáveis contínuas em categóricas), 1 a N (converte variáveis categóricas em uma representação numérica) e técnicas de redução de dimensionalidade (combinar várias variáveis em uma única) são comumente usadas.

4.5 Data Mining

O objetivo principal do passo *data mining* no processo *KDD* é a aplicação de técnicas de mineração nos dados pré-processados. Isto envolve ajuste de modelos e/ou determinação de características nos dados. Em outras palavras, envolve o uso de algoritmos de *data mining*.

Deve-se destacar que cada técnica de *data mining* ou cada implementação específica dos algoritmos que são utilizados para conduzir as operações *data mining*, adapta-se melhor a alguns problemas que a outros, o que impossibilita a existência de um método de *data mining* universalmente melhor. Para cada particular problema tem-se um particular algoritmo. Portanto, o sucesso de uma tarefa *data mining* está diretamente ligada à experiência e intuição do analista.

Um grande número e uma grande variedade de algoritmos de *data mining* estão descritos na literatura de estatística, reconhecimento de padrão, inteligência artificial, redes neurais e bancos de dados. Em geral, os algoritmos de *data mining* consistem em uma mistura de três componentes: o modelo, o critério de preferência e o algoritmo de busca.

4.5.1 O Modelo

Existem dois fatores relevantes neste componente: a função do modelo e a representação do modelo.

Funções do Modelo

As funções mais comuns utilizadas na prática de *data mining* incluem:

•**Classificação** — Associa ou classifica um item em uma ou várias classes predefinidas.

•**Regressão** — Associa um item a uma ou mais variáveis de predição de valores reais.

•**Modelos de Dependência** — Descreve dependências significantes entre variáveis.

•**Análise de Associação** — Determina relações entre campos de um banco de dados. Por exemplo, o modelo de associação pode

descrever quais itens são usualmente comprados conjuntamente com outros itens em um supermercado.

•**Análise de Sequência** — Determina características sequenciais, como, por exemplo, em dados com dependência no tempo. A idéia é extrair e registrar desvios e tendências no tempo.

•**Cluster** — Associa um item com um ou vários agrupamentos determinados pelos dados. Observe que, ao contrário de classificação, onde as classes são predeterminadas, os *clusters* são definidos através de agrupamentos naturais dos dados, baseados em medidas de similaridade ou modelos probabilísticos.

•**Sumarização** — Descreve de forma compacta um subconjunto de dados. Exemplos simples seriam o cálculo da média e do desvio padrão.
Exemplos mais sofisticados seriam o uso de técnicas de visualização e a determinação de relações funcionais entre variáveis. Funções de sumarização são frequentemente usadas na análise exploratória de dados, com geração automatizada de relatórios.

Representação do Modelo

As representações mais populares de modelos incluem decisão por árvore, regras de decisão, modelos lineares, modelos não-lineares (entre eles, redes neurais), método do vizinho-mais-próximo e modelos de dependência, entre outros. A representação do modelo determina a flexibilidade do mesmo em representar os dados e a sua interpretabilidade. Os modelos mais complexos podem ajustar melhor os dados, entretanto, ficam mais difíceis de serem interpretados.

4.5.2 O Critério de Preferência

Conhecido pelos estatísticos como seleção de modelos, o critério de preferência determina se um particular modelo e seus parâmetros

cumprem os critérios do processo KDD. Usualmente, existe um critério explícito quantitativo embutido nos algoritmos de busca (critério de máxima verossimilhança para estimar os parâmetros), e um critério implícito (refletindo o vício subjetivo do analista em relação aos modelos que serão inicialmente considerados).

4.5.3 Algoritmo de Busca

Os algoritmos de busca podem ser classificados em dois tipos: i) busca (estimação) de parâmetros, dado um modelo e ii) busca (ajuste) de um modelo sob o espaço de modelos. A busca dos "melhores" parâmetros é, frequentemente, um problema de otimização (encontrar o máximo global de uma função não-linear no espaço paramétrico).

4.5.4 Técnicas de *Data Mining*

Várias Técnicas de *data mining* serão discutidas com detalhes nos Capítulos 5 a 10.

4.6 Assimilação de Conhecimento

Este passo envolve o aproveitamento dos resultados da aplicação das técnicas de *data mining* dentro da administração da empresa ou na pesquisa acadêmica. A idéia básica é apresentar as descobertas obtidas, no passo *data mining*, de forma convincente e determinar quais são as melhores maneiras de utilizar tais informações na tomada de decisão.

Nesta fase define-se também, quais foram as vantagens e desvantagens do projeto, aproveitando o aprendizado para uma redefinição do mesmo ou criação de um novo projeto.

Capítulo 5

Classificação e Regressão

5.1 Introdução

As técnicas, atendendo os objetivos principais de *data mining,* são utilizadas para descrever (visualizar) e predizer. Qual técnica deve ser usada para um particular problema, como visto no capítulo anterior, depende da experiência do analista.

Neste capítulo, as técnicas utilizadas em predição, envolvendo classificação e regressão, são discutidas. O modelo de classificação é apresentado através da perspectiva de decisão por árvore, regra de decisão e análise discriminante. O modelo de regressão é apresentado através do modelo de regressão logística. Exemplos ilustrando o uso das técnicas são esboçados.

5.2 Modelos de Predição

Nas operações *data mining* usam-se modelos de predição para determinar características essenciais sobre os dados. Para isto, é necessário que o conjunto de dados apresente observações válidas e completas

Figura 5.1: Modelo de predição.

para que o modelo possa predizer de forma precisa. O modelo deve ser capaz de predizer corretamente situações já conhecidas antes de iniciar a predição de novas situações. Quando um algoritmo trabalha desta forma, a abordagem é chamada *aprendizado supervisionado.* Fisicamente, o modelo pode ser um conjunto de regras do tipo SE ENTÃO (IF THEN) em algum formato particular, como, por exemplo, um segmento de códigos em linguagem C.

A Figura 5.1 ilustra uma abordagem de análise via modelo de predição. Aqui uma empresa de seguros está interessada em entender a taxa de crescimento de insatisfação do cliente. Um modelo de predição determinou que apenas duas variáveis são de interesse - o período de tempo que o cliente permanece com a companhia

(**Período**) e o número de serviços da empresa que o cliente usa (**Serviço**). A árvore de decisão (ver Seção 5.3.1) apresenta a análise de uma maneira intuitiva. Os clientes que estiveram com a empresa por menos que dois anos e meio e usam somente um ou dois serviços são os mais prováveis a abandoná-la.

Os modelos são desenvolvidos em duas fases: treinamento e teste. Treinamento refere-se a construção de um novo modelo, usando dados históricos. Teste refere-se a utilização do modelo em um conjunto de dados ainda não explorado para determinar sua precisão. O treinamento do modelo é feito em uma proporção razoável do total de dados disponíveis, enquanto o teste é feito em uma pequena parte dos dados que foram reservados exclusivamente para este fim.

Data mining considera dois tipos de modelos de predição: classificação e regressão. Um modelo de predição com classificação é usado para estabelecer uma específica classe para cada registro do banco de dados. A classe deve ser de um conjunto finito de possíveis e predeterminados valores de classe. Considerando-se o exemplo da empresa de seguros, a variável de interesse seria a classe de clientes com dois valores possíveis: **Permanece** e **Abandona**.

Um modelo de predição com regressão é usado para predizer um valor numérico contínuo que está associado a um registro do banco de dados. Por exemplo, uma empresa financeira pode estar interessada em predizer o tempo até o abandono de um específico cliente ou em estimar a probabilidade do cliente saldar um empréstimo. As soluções destes problemas envolvem técnicas conhecidas como *behavior scoring* e *credit scoring,* respectivamente, que por sua vez utilizam uma técnica estatística conhecida como regressão logística (ver Seção 5.5 e 5.6).

Na linguagem estatística, os modelos de predição com classificação e com regressão são chamados, respectivamente, árvore de classificação e árvore de regressão.

5.3 Decisão por Árvore e Regra de Decisão

Nesta seção o modelo de classificação através da perspectiva de decisão por árvore e regra de decisão (Apté e Weiss, 1997) são discutidos. Para melhor assimilação destas técnicas, um simples e hipotético exemplo é apresentado.

Suponha que 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 caracteriza a variedade do pino como uma função do comprimento e do diâmetro pode ser útil para se entender como estas variedades diferem. Os dados são ilustrados na Figura 5.2. A Figura também mostra duas linhas paralelas aos eixos, uma no *comprimento* = 0.75 e outra no *diâmetro* = 3,00, que parecem particionar as três variedades em três diferentes sub-áreas.

Métodos de solução por decisão por árvore fornecem automaticamente estas partições de eixos paralelos. A Figura 5.3 ilustra uma decisão por árvore que corresponde a partição mostrada na Figura 5.2.

A árvore é formada por nós e é no primeiro deles, o nó raiz, que envolve todo o conjunto de dados, onde o processo de classificação inicia. O teste no nó raiz testa todos os itens para *comprimento* $\leq$ 0,75. Itens que satisfazem este teste vão para o arco abaixo pela esquerda (Verdadeiro), permanecendo em um nó (nó terminal ou folha), indicando que todos os pinos pertecem a uma classe (Quadrado) e nenhum teste será mais necessário.

O arco à direita (Falso) do nó raiz recebe todos os casos que falharam no teste inicial. Estes pinos ainda não pertecem a uma só classe e, portanto, futuros testes serão necessários neste nó intermediário. O teste neste nó é para *diâmetro* $\leq$ 3,00. Pinos que

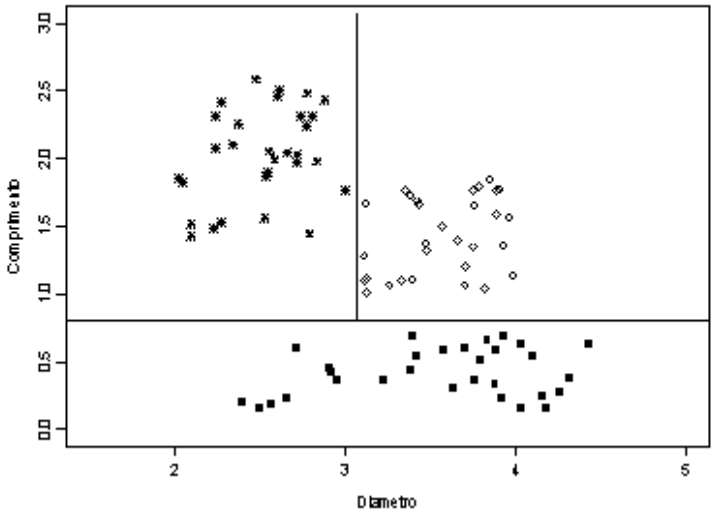

Figura 5.2: Dados dos pinos. Classificação que caracteriza a variedade do pino como uma função do comprimento e do diâmetro.

satisfazem este teste estão todos em uma só classe (Estrela) e aqueles que falharam também estão em uma classe (Losango), e assim ambos arcos deste nó intermediário conduziram a nós terminais.

Este exemplo ilustra o caso de um classificador estruturado por árvore *binária* (Breiman, *et.al.*, 1984), onde cada nó intermediário pode-se subdividir em, no máximo, duas sub-árvores. Decisão por árvore pode também ser não-binária, bastando para isto que em cada nó ocorra uma subdivisão em mais que duas sub-árvores. Isto é, os testes apresentam mais que dois resultados possíveis.

Regras de decisão são soluções alternativas às baseadas em decisão por árvore. Uma regra pode ser construída através da formação

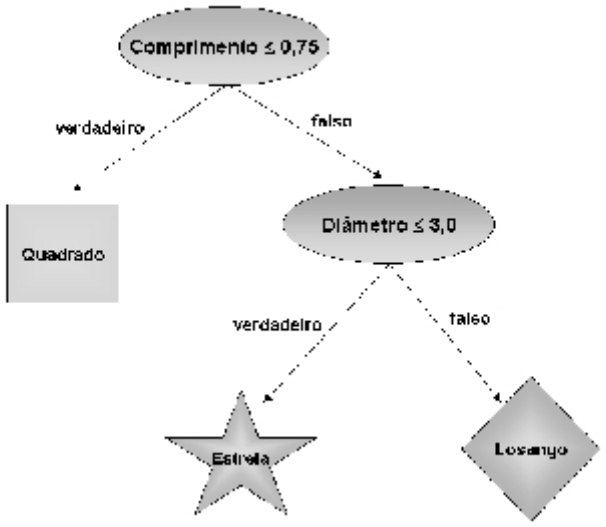

Figura 5.3: Classificando pinos através de decisão por árvore.

de um conjunto de testes que ocorrem nos caminhos entre o nó raiz e os nós terminais. A coleção de todas tais regras obtidas considerando cada caminho do nó raiz a um nó terminal é uma solução, baseada em regras, para a classificação. No exemplo dos pinos, utilizados para ilustrar a decisão por árvore, a solução por regra é dado por:

`Se` (`Comprimento` $\leqslant 0,75$)
`Então Quadrado`
`Se` (`Não` (`Comprimento` $\leqslant 0,75$)) & (`Diâmetro` $\leqslant 3,00$)
`Então Estrela`
`Se` (`Não` (`Comprimento` $\leqslant 0,75$)) & (`Não` (`Diâmetro` $\leqslant 3,00$))
`Então Losango`

Gerada uma solução por decisão por árvore ou por regra de decisão, esta pode ser usada para estimar ou predizer a resposta ou

variável classe para um novo caso.

Estimar a verdadeira acurácia de um modelo de decisão por árvore ou de regra de decisão é um dos aspectos mais importante do processo de modelagem. Esta é a razão principal de empregar uma estratégia dupla no processo de geração do modelo: treinamento e teste, como discutido anteriormente. O segundo passo que envolve o teste da solução proposta em casos independentes, às vezes utiliza um processo de poda para compensar o super-ajustamento ocorrido no primeiro passo.

O problema a ser solucionado com a poda pode ser especificado como segue: dada uma amostra de itens, S, onde cada item é composto das características observadas e da classe, o problema seria encontrar o melhor modelo, RS_{best}, tal que a razão de erro em novos itens seja mínima. Dado um modelo super ajustado, RS, para um conjunto de itens S é necessário determinar uma derivativa de RS que satisfaça o critério acima.

Várias técnicas de poda estão disponíveis na literatura. Estas técnicas usualmente empregam ou a abordagem de validação cruzada ou a abordagem de treinamento-teste. Validação cruzada é preferida se a modelagem estiver sendo feita com amostras pequenas. Nestes casos, dividimos os dados, repetidamente, em diferentes combinações de partições de treinamento e teste. A partição treinamento é usada para gerar um modelo super ajustado, enquanto a partição teste é usada para generalizar este modelo para a melhor derivativa possível. O melhor modelo, RS_{best}, dentre os vários candidatos disponíveis, é selecionado considerando uma média entre as várias combinações de treinamento-teste. Quando o conjunto de dados disponível é grande, uma simples partição treinamento-teste é suficiente para avaliar e selecionar RS_{best} usando a abordagem de poda.

5.3.1 Modelagem Através de Decisão por Árvore

A aplicação da técnica decisão por árvore a um conjunto de itens é um processo de decisão de cima para baixo, controlado pela avaliação de testes e tomando uma ramificação apropriada, iniciando no nó raiz e terminando quando um nó é alcançado. Se todos os itens pertencessem a mesma classe então nenhuma decisão futura seria necessária para particionar os itens e, portanto, a solução estaria completa. Se itens neste nó intermediário pertencessem a duas ou mais classes, então um teste seria elaborado e este nó resultaria em uma divisão.

O processo é repetido recursivamente para cada um dos nós intermediários até que uma árvore discriminante completa seja obtida. Uma decisão por árvore neste estágio é potencialmente uma solução super ajustada, isto é, podem existir componentes que foram criados não por razões estruturais, mas sim por ruídos ou por *outliers* presentes nos dados de treinamento. Para relaxar este super ajustamento, vários métodos de decisão por árvore utilizam o recurso de poda, que tenta generalizar a árvore eliminando sub-árvores que parecem ser muito específicas.

CART

CART – Classification And Regression Tree – (Breiman *et. al.*, 1984) é um método de decisão por árvore binária que tem sido usado extensivamente (Lael, 1996; Ferreira, 1999). Nesta técnica uma função de avaliação usada para dividir os nós, ou para medir a homogeneidade ou a impureza de determinado nó, é o índice *Gini*.

Para um dado nó t, este índice é definido como

$$Gini(t) = 1 - \sum_i p_i^2,$$

onde p_i é a probabilidade da i-ésima classe no nó t. No caso de existir

apenas duas classes, o índice *Gini*I reduz-se a $Gini(t) = 2p(1|t)(1 - p(1|t))$, onde t é um nó arbitrário, $p(1|t)$ é a probabilidade da variável resposta ser classificada na categoria 1 no nó t e $(1 - p(1|t))$ é a probabilidade de ser classificada na categoria 2 no nó t.

Poda do CART

O mecanismo de poda do CART é uma tentativa de conseguir o tamanho correto da árvore que minimiza o erro de classificação errônea. Uma árvore completa tem razão de erro próxima a zero nos dados de treinamento (dados considerados para a construção da árvore). Todavia, sua razão de erro, medida através da avaliação das classificações errôneas quando a árvore é aplicada ao conjunto de dados de teste, pode ser muito maior. O objetivo do processo de poda é o de encontrar a sub-árvore que produz o menor erro, considerando o tamanho (complexidade) da árvore. A importância das sub-árvores geradas é expressa em termos de uma medida denotada custo-complexidade. O custo-complexidade de uma árvore é uma função da classificação errônea da árvore nos dados de treinamento e do seu tamanho.

Utilizando-se a formulação de custo-complexidade pode-se derivar uma sequência de árvores com custo-complexidade decrescente, começando na árvore de maior tamanho. Esta sequência é recursivamente criada selecionando a última árvore na sequência (inicialmente, a árvore de maior tamanho), examinando cada uma de suas sub-árvores, escolhendo aquela com a menor medida de custo-complexidade e fazendo desta a próxima sub-árvore na sequência. O processo termina quando a última sub-árvore é exatamente o nó raiz.

Um critério alternativo para identificar a melhor sub-árvore gerada no processo de poda é discutido em Zhang et al. (1996). E um processo alternativo à poda, denotado encolhimento (*shrinking*) é discutido em Clark e Pregibon (1992).

C4.5

C4.5 (Quinlan, 1993) é outro popular método de decisão por árvore. Este método é uma extensão de um sistema de modelagem por decisão por árvore conhecido como ID3. O método ID3 utiliza critério de entropia para dividir os nós. Dado um nó t, o critério de divisão usado é

$$Entropia(t) = \sum_i -p_i \log p_i,$$

onde p_i é a probabilidade da i-ésima classe dentro do nó t. Em C4.5, dado um nó t, o critério de divisão usado é

$$Razão\ Ganho(t) = \frac{ganho(t)}{Informacão\ da\ Divisão(t)}.$$

Esta razão expressa a proporção de informação gerada pela divisão, que é útil para desenvolver a classificação e que pode ser pensada como um ganho de informação normalizada ou como medida de entropia para o teste. O numerador nesta razão (ganho de informação) é a diferença de entropia de informação no nó t (Quinlan, 1993). O teste que maximiza esta razão é selecionado.

Regra de Classificação Bayesiana

A regra de classificação Bayesiana designa um item à classe com a probabilidade condicional mais alta. Esta regra é conhecida como sendo ótima, isto é, a regra minimiza o erro de classificação errônea. Formalmente, a regra designa um item à classe C_i se $P(C_i|x) > P(C_j|x)$, para todo $i \neq j$.

Outros Critérios

Outros critérios para a divisão das árvores incluem o *Mean Posterior Improvement* (MPI) discutido Taylor e Jones(1996) e a técnica

desenvolvida por Segal (1988, 1992) para variável resposta sujeita à censura.

Qualquer que seja a técnica utilizada, o processo de subdivisão termina quando o nó resultante for homogêneo o suficiente ou possuir um número reduzido de observações.

5.3.2 Modelagem Através de Regras de Decisão

Regras de decisão podem ser induzidas dos dados de treinamento da mesma forma que árvores de decisão, isto é, de cima-para-baixo e de geral-para-específico ou podem, também, utilizar a forma de baixo-para-cima e do específico-para-geral.

A coleção de todos os itens individuais em um conjunto de dados de treinamento forma o estado inicial de uma solução via regra de decisão. Cada um desses itens pode ser imaginado como uma regra de decisão altamente especializada. Vários sistemas de modelagem via regra de decisão empregam um processo de busca que faz com que este conjunto altamente específico desenvolva-se para regras mais gerais. Este processo de busca é iterativo e, usualmente, termina quando as regras não podem mais ser generalizadas ou um critério de parada é satisfeito.

Similar a construção de árvores de decisão, dados com *ruídos* podem conduzir a uma regra de decisão super-ajustada. Para contornar este problema, vários mecanismos de poda têm sido desenvolvidos.

Métodos de regras induzidas têm como objetivo encontrar um conjunto de regras de "cobertura" que particione completamente os itens em suas corretas classes. Este conjunto é encontrado através da busca de uma regra simples "ótima" que cobre os casos de apenas uma classe. Tendo encontrado a regra "ótima" para a classe C, digamos, esta regra é adicionada ao conjunto de regras. Os casos que satisfazem esta regra "ótima" são removidas de considerações

futuras. O processo de busca é repetido até que não exista nenhum caso para ser coberto.

AQ

A família de algoritmos AQ (Michalski *at. al.*, 1986) é influenciada e motivada pelos métodos usados em Engenharia Elétrica para simplificação de circuitos lógicos. Usando a terminologia AQ, um teste em um atributo é chamado seletor, um conjunto de testes é chamado complexo e uma subdivisão de complexos é chamado cobertura. Se uma regra satisfaz um item, esta regra é chamada uma cobertura para o item. Inicialmente cada item é, ele próprio, um complexo no modelo. Os complexos são, então, examinados e os seletores são retirados na medida em que o complexo resultante permaneça consistente, isto é, associando apenas os itens da mesma classe. E os complexos são, desta forma, produzidos um de cada vez.

No processo de busca para a criação de complexos uma função de avaliação é usada para ordenar e determinar quais seletores devem ser retirados ou generalizados. Uma função frequentemente usada é a razão de itens classificados corretamente por um complexo pelo total de itens classificados por este complexo.

CN2

Ao contrário da técnica AQ, o sistema CN2 executa uma busca do tipo geral-para-específico. Em cada passo ou um novo seletor é adicionado a um complexo ou um complexo é inteiramente removido. Dois métodos são utilizados na busca pelo complexo ótimo, um método que leva em conta um determinado limite, tal que qualquer complexo abaixo deste limite não será considerado na seleção do complexo ótimo e, um segundo método, que leva em conta uma medida da qualidade do complexo que é usado para dar uma ordenação

dos complexos candidatos para inclusão na cobertura final.

Com mais detalhe, no primeiro método deve-se calcular a estatística

$$2\sum_{i=1}^{n} p_i \log(p_i/q_i), \tag{5.1}$$

onde $p_1, \ldots p_n$ são as distribuições de frequências observadas dos itens entre classes, satisfazendo um dado complexo, e $q_1, \ldots q_n$ são as distribuições de frequências esperadas do mesmo número de itens, sob a restrição que os complexos selecionam itens aleatoriamente. Esta estatística representa uma medida de distância entre as duas distribuições. Qualquer complexo cuja estatística (5.1) esteja abaixo de um limite predeterminado é rejeitado. No segunto método, a medida comumente utilizada é o estimador do erro dado por

$$(n - n_c + k - 1)/(n + k),$$

onde n é o número total de itens cobertos pela regra, n_c é o número de itens positivos cobertos pela regra e k é o número de classes no conjunto de dados.

RAMP

O sistema de geração de regras RAMP (Hong, 1997) gera regras de classificação "minimal", considerando conjuntos de dados tabulados com uma das colunas correspondendo à variável "classe" e as colunas restantes correspondendo às características explanatórias. Antes das gerações das regras, o conjunto de dados deve estar completamente discretizado, ou seja, as características contínuas devem ser categorizadas para um conjunto finito de valores discretos.

O objetivo principal da abordagem RAMP é encontrar uma regra "minimal" que seja, ao mesmo tempo, completa e consistente com os dados de treinamento. Completa no sentido que as regras cobrem todos os itens dos dados de treinamento e consistente no sentido que as

regras não cobrem nenhum falso-item para as suas classes sugeridas. O sistema RAMP utiliza uma metodologia de minimização lógica, conhecida como R-Mini, para gerar regras "minimais" completas e consistentes.

5.4 Análise Discriminante

Uma técnica estatística apropriada para discriminação e classificação é a análise discriminante (Hair *et. al.,* 1998; Johnson *et. al.* 1998). Os objetivos imediatos desta técnica envolvem a descrição, gráfica ou algébrica, das características diferenciais das observações de várias populações, além de classificar as observações em uma ou mais classes predeterminadas. A idéia é derivar uma regra que possa ser usada para classificar de forma otimizada uma nova observação à uma classe já rotulada.

Análise discriminante é adequada nas situações onde se pretende separar duas ou mais classes de objetos (pessoas, clientes, empresas, produtos, entre outros) ou alocar um novo objeto a uma das classes existentes ou, ainda, se pretende, conjuntamente, separar as classes e alocar um novo objeto. As classes poderiam ser, digamos, risco de crédito ruim e bom de clientes de uma instituição financeira. O vetor de variáveis medidas $\mathbf{x}$ teria componentes, tais como renda, idade, número de cartões de créditos, tamanho da família, saldo bancário, tempo como cliente. Uma vez determinada a regra de classificação, dado o conhecimento de seu vetor de variáveis medidas, um futuro cliente poderia ser classificado como mau ou bom pagador.

5.4.1 Função Discriminante de Fisher – Dois Grupos

Considere, inicialmente, duas populações ou grupos π_1 e π_2. A função discriminante de Fisher (Fisher,1938) é construída sem as-

sumir qualquer forma paramétrica para os grupos. Isto é, sem assumir a exis- tência de uma função de probabilidade associada a cada grupo.

A idéia de Fisher é procurar por uma regra, sensível o suficiente, que possa discriminar entre as duas populações. Sua sugestão para isto foi procurar a função linear $\mathbf{a'x}$ a qual maximiza a razão entre a soma de quadrados entre grupos e a soma de quadrados dentre grupos. A função linear $\mathbf{a'x}$ é chamada *função discriminate linear de Fisher.* O vetor $\mathbf{a}$ é o autovetor da matriz $\mathbf{W}^{-1}\mathbf{B}$ que corresponde ao máximo autovalor, onde $\mathbf{W}$ e $\mathbf{B}$ são as matrizes das somas de quadrados e produtos cruzados dentre grupos e entre grupos respectivamente (ver Mardia, 1989). As médias amostrais $\overline{\mathbf{x}}_i$ terão escores $\mathbf{a'}\overline{\mathbf{x}}_i$, e, no caso de apenas dois grupos, $\mathbf{a} = \mathbf{W}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$. Uma vez que a função discriminante tenha sido determinada, um novo objeto pode ser alocado a uma das duas populações levando em conta o "escore discriminante" $\mathbf{a'x} = (\overline{\mathbf{x}}_\mathbf{1} - \overline{\mathbf{x}}_\mathbf{2})'\mathbf{W}^{-\mathbf{1}}\mathbf{x}$. O objeto com variáveis medidas $\mathbf{x}_0$ é alocado àquela população cujo escore médio, $\mathbf{a'}\overline{\mathbf{x}}_i$, é próximo a $\mathbf{a'x}_0$. Ou seja,

aloca-se o objeto ao grupo π_1 se $(\overline{\mathbf{x}}_\mathbf{1} - \overline{\mathbf{x}}_\mathbf{2})'\mathbf{W}^{-\mathbf{1}}\{\mathbf{x}_0\text{-}\frac{1}{2}(\overline{\mathbf{x}}_\mathbf{1} + \overline{\mathbf{x}}_\mathbf{2})\} > 0$

,

e ao grupo π_2 caso contrário.

Considerando a suposição que as duas populações, quaisquer que sejam suas formas, têm uma matriz de variâncias-covariâncias comum, a função discriminante linear de Fisher é desenvolvida substituindo a matriz $\mathbf{W}$ pela matriz $\mathbf{S}_{pooled}$, uma combinação de dois estimadores da matriz de variâncias-covariâncias comum, definida como

$$\mathbf{S}_{pooled} = [\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}]\mathbf{S}_1 + [\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}]\mathbf{S}_2,$$

onde

$$\mathbf{S}_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (\mathbf{x}_{1j} - \overline{\mathbf{x}}_{\mathbf{1}})(\mathbf{x}_{1j} - \overline{\mathbf{x}}_{\mathbf{1}})'$$

e

$$\mathbf{S}_2 = \frac{1}{n_2 - 1} \sum_{k=1}^{n_2} (\mathbf{x}_{2k} - \overline{\mathbf{x}}_{\mathbf{2}})(\mathbf{x}_{2k} - \overline{\mathbf{x}}_{\mathbf{2}})'.$$

Os vetores $\mathbf{x}_{1j}$, $j = 1, 2, \ldots n_1$ e os vetores $\mathbf{x}_{2k}$, $k = 1, 2, \ldots n_2$ correspondem a amostras de n_1 vetores de variáveis medidas do grupo 1 e a n_2 vetores de variáveis medidas do grupo 2, respectivamente.

Assim, uma regra de classificação baseada na função discriminante de Fisher para duas populações, com matriz de variâncias-covariâncias comum é dada por,

aloca-se o objeto ao grupo π_1 se $(\overline{\mathbf{x}}_{\mathbf{1}} - \overline{\mathbf{x}}_{\mathbf{2}})' \mathbf{S}_{pooled}^{-1} \{\mathbf{x}_0 - \frac{1}{2}(\overline{\mathbf{x}}_{\mathbf{1}} + \overline{\mathbf{x}}_{\mathbf{2}})\} > 0$

,

caso contrário aloca-se ao grupo π_2.

5.4.2 Função Discriminante de Fisher – Vários Grupos

Fisher também propôs uma extensão do seu método discriminante para várias populações. A extensão assume que as matrizes $p \times p$ de variâncias-covariâncias das g populações são iguais e de posto completo. Seja $\mathbf{S}_{pooled}$ um estimador da matriz de variâncias-covariâncias comum. Sejam $\lambda_1, \ldots, \lambda_s > 0$ denotando os $s \leq \min(g - 1, p)$ autovalores não nulos de $\mathbf{W}^{-1}\mathbf{B}$ e sejam $\mathbf{e}_1, \ldots \mathbf{e}_s$ os correspondentes autovetores de tal forma que $\mathbf{e}'\mathbf{S}_{pooled}\mathbf{e} = 1$. O vetor de coeficientes $\mathbf{a}$ que maximiza a razão entre a soma de quadrados entre grupos e a

soma de quadrados dentre grupos

$$\frac{\mathbf{a}'\mathbf{B}\mathbf{a}}{\mathbf{a}'\mathbf{W}\mathbf{a}} = \frac{\mathbf{a}'\left(\sum_{i=1}^{g} n_i'(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})(\overline{\mathbf{x}}_i - \overline{\mathbf{x}})'\right)\mathbf{a}}{\mathbf{a}'\left(\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{1j} - \overline{\mathbf{x}}_i)(\mathbf{x}_{1j} - \overline{\mathbf{x}}_i)'\right)\mathbf{a}}$$

é dado por $\mathbf{a}_1 = \mathbf{e}_1$. A combinação linear $\mathbf{a}_1'\mathbf{x}$ é chamada primeiro discriminante amostral. O segundo discriminante amostral é dado por $\mathbf{a}_2'\mathbf{x}$, onde $\mathbf{a}_2 = \mathbf{e}_2$. O k-ésimo discriminante amostral é dado por $\mathbf{a}_k'\mathbf{x}$, onde $\mathbf{a}_k = \mathbf{e}_k$, $k \leq s$. Desta forma, baseado nos primeiros $r \leq s$ discriminates amostrais, o objeto com variáveis medidas $\mathbf{x}_0$ é alocado à população π_k se

$$\sum_{j=1}^{r}\left[\mathbf{e}_j'\left(\mathbf{x}_0 - \overline{\mathbf{x}}_k\right)\right]^2 \leq \sum_{j=1}^{r}\left[\mathbf{e}_j'\left(\mathbf{x}_0 - \overline{\mathbf{x}}_i\right)\right]^2, \qquad \text{para todo } i \neq k.$$

5.5 Regressão Logística

O modelo de regressão logística, também conhecido por modelo logístico, é, em geral, utilizado para tratar problemas relacionados a dados dicotômicos em várias áreas do conhecimento. Especificamente em *data mining* é de interesse saber qual a probabilidade de um indivíduo pertencer a um determinado grupo.

Este modelo estabelece uma relação entre a probabilidade de ocorrência dos resultados de uma variável resposta dicotômica (em geral chamada de variável dependente), que normalmente é representada pelos termos sucesso e fracasso, e variáveis explicativas categóricas ou contínuas (conhecidas como variáveis independentes). A idéia básica consiste em estabelecer uma relação linear entre as variáveis explicativas (ou alguma transformação dessas) e uma transformação, denominada logito (logit), da variável resposta. Este modelo é re-

presentado por (Hosmer e Lemeshow, 1989; Arminger et al., 1997)

$$\log\left[\frac{P\{Y(\mathbf{x})=1\}}{P\{Y(\mathbf{x})=0\}}\right] = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p, \tag{5.2}$$

onde $P\{Y(\mathbf{x})=1\}$ representa a probabilidade de sucesso para a variável resposta, $P\{Y(\mathbf{x})=0\}$ representa a probabilidade de fracasso, β_0 denota o intercepto da regressão e $\mathbf{x}'=(x_1, x_2, ..., x_p)$ é um vetor de variáveis explicativas com coeficientes $\beta_1, \beta_2, ..., \beta_p$.

Dessa forma, de (5.2), a probabilidade de sucesso para a variável resposta é dada por

$$p(\mathbf{x}) = P\{Y(\mathbf{x})=1\} = \frac{\exp\{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p\}}{1 - \exp\{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p\}}. \tag{5.3}$$

Alguns aspectos importantes a serem considerados consistem no ajuste do modelo, na seleção de variáveis explicativas e na verificação do ajuste.

5.5.1 Ajuste do Modelo

O ajuste pode ser feito tanto através do método de mínimos quadrados quanto pelo método de máxima verossimilhança (Hosmer e Lemeshow, 1989). Sempre que possível a preferência deve ser dada ao método de máxima verossimilhança por possuir propriedades ótimas.

Dada uma amostra constituída por n pares $(y_i, \mathbf{x}_i)$, onde y_i representa a variável resposta em (5.2) para o i-ésimo indivíduo e $\mathbf{x}_i$ é o i-ésimo vetor de variáveis explicativas, a função de log-verossimilhança para os parâmetros do modelo, $\boldsymbol{\beta}' = (\beta_1, \beta_2, ..., \beta_p)$, é dada por

$$l(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \{y_i \log[p(\mathbf{x}_i)] + (1 - y_i) \log[1 - p(\mathbf{x}i)]\} \tag{5.4}$$

onde $p(\mathbf{x}_i)$ é definida em (5.3). Os estimadores de máxima verossimilhança dos parâmetros, $\widehat{\beta}_j$, podem ser obtidos via método numérico

ou através da maximização direta de (5.4). A partir dos $\widehat{\beta}_j$ encontrados podemos obter a estimativa de (5.3) como,

$$\widehat{p}(\mathbf{x}) = \frac{\exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_p x_p\}}{1 - \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + ... + \widehat{\beta}_p x_p\}}. \tag{5.5}$$

5.5.2 Seleção de Variáveis

A seleção de variáveis explicativas pode ser baseada na estatística da razão de verossomilhança dada por

$$\lambda = 2(l_c - l_s), \tag{5.6}$$

onde l_c denota a log-verossomilhança do modelo mais completo e l_s denota a log-verossimilhança do modelo menos completo. Esta estatística tem distribuição Qui-quadrado com ν graus de liberdade. O número de graus de liberdade ν é definido como sendo igual a diferença entre o número de variáveis explicativas existentes nos dois modelos.

Entretanto, dois testes são ainda comumente utilizados como alternativas ao teste da razão da verossimilhança. São eles, o teste de Wald e o teste Escore.

A estatística do teste de Wald é dada por

$$\widehat{\boldsymbol{\beta}}\ I^{-1}_{\widehat{\boldsymbol{\beta}}}\ \boldsymbol{\beta}, \tag{5.7}$$

onde $I_{\widehat{\boldsymbol{\beta}}} = [\partial^2 l(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^2]|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}$. Essa estatística possui uma distribuição Qui-quadrado com $p+1$ graus de liberdade sob a hipótese de que os $p+1$ coeficientes são iguais a zero (Garthwaite *et. al.*, 1995). Enquanto, a estatística do teste Escore é dada por,

$$u'(\boldsymbol{\beta})\ I^{-1}_{\boldsymbol{\beta}}\ u(\boldsymbol{\beta}), \tag{5.8}$$

onde $u'(\boldsymbol{\beta}) = (\partial l(\boldsymbol{\beta})/\partial \beta_1, \partial l(\boldsymbol{\beta})/\partial \beta_2, ..., \partial l(\boldsymbol{\beta})/\partial \beta_{p+1})|_{\boldsymbol{\beta} \in H_0}$ e $I_{\boldsymbol{\beta}} = [\partial^2 l(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^2]|_{\boldsymbol{\beta} \in H_0}$, onde H_0 corresponde a hipótese nula. Essa

estatística possui uma distribuição Qui-quadrado com $p+1$ graus de liberdade sob a hipótese de que os $p+1$ coeficientes são iguais a zero.

5.5.3 Verificação do Ajuste

A qualidade de um modelo ajustado via regressão logística pode ser verificada através da comparação entre os valores observados e os valores preditos para a variável resposta. De uma forma simples, pode-se construir um gráfico dos valores observados versus os valores preditos e verificar a existência de uma relação linear entre essas duas quantidades. Métodos formais da qualidade do ajuste podem ser encontrados em Hosmer e Lemeshow, 1989.

5.6 *Credit Scoring* e *Behavior Scoring*

Estas técnicas, muito difundidas no meio financeiro, basicamente consistem em aplicações do modelo de regressão logística discutido anteriormente. As probabilidades de sucesso de todos os indivíduos são calculadas e ordenadas. Um número fixo de classes é, então, determinado. A classe 1 poderia envolver, por exemplo, os indivíduos que obtiveram probabilidade menor que 0,2; a classe 2, os indivíduos que obtiveram probabilidade entre 0,2 e 0,4; e assim por diante. Com o modelo ajustado e as classes de escores determinadas, poder-se-ia, facilmente, encontrar o escore (de crédito ou de comportamento) de um novo indivíduo, ou um novo cliente, e classificá-lo à uma das classes.

A técnica *credit scoring* é utilizada, principalmente, para determinar risco de crédito. Levando em consideração um modelo de regressão logística já ajustado, a probabilidade de perda, isto é, a probabilidade do cliente não pagar o empréstimo tomado, é calculada considerando-se fatores de riscos, tais como, idade, condição

sócio-econômica, histórico de inadimplência, setor de atividade etc. e/ou fatores de riscos característicos da operação, valor total do empréstimo, prazo de pagamento, tipos de garantias.

A técnica *behavior scoring* é utilizada, entre outras, na determinação da chance do cliente abandonar uma determinada instituição financeira. Um modelo de regressão logística é construído, levando em consideração informações transacionais de clientes que permanecem e dos que já abandonaram a instituição. O escore de comportamento de um cliente (que não participou da construção do modelo) pode ser determinado através do cálculo da probabilidade de abandono dado todas as suas informações transacionais (saldo bancário, número de depósitos, quantidade de produtos, data do último depósito etc.).

Capítulo 6

Análise de Associação

6.1 Introdução

Análise de associação é o processo de interconexão de objetos na tentativa de expor características e tendências (Cabena *et.al.*,1998). Em outras palavras, a análise de associação gera redes de interações e conexões presentes nos conjuntos de dados usando as associações item a item. Onde por associação item a item entende-se que a presença de um item implica necessariamente na presença de outro item na mesma transação.

Através de uma representação explícita dos relacionamentos entre objetos é possível ganhar uma perspectiva inteiramente diferente na forma pela qual os dados podem ser analisados e os tipos de características que podem ser descobertas.

Neste capítulo são discutidas algumas técnicas de associação.

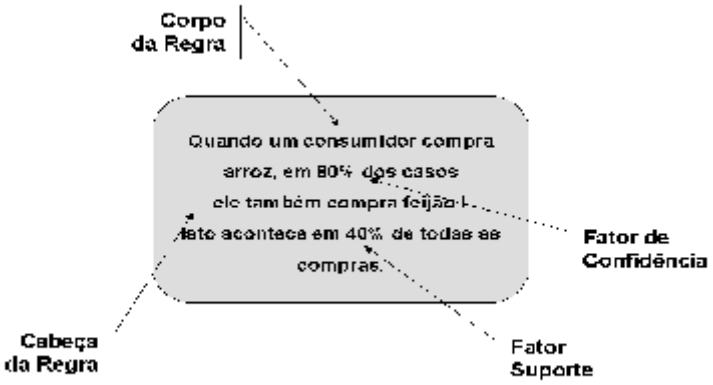

Figura 6.1: Exemplo de uma regra de associação.

6.2 Regras de Associação

Sistemas de análise de associação foram utilizados inicialmente em investigação criminal. Recentemente, esses sistemas fizeram significantes incursões em uma grande variedade de aplicações comerciais. Por exemplo, considere um banco de dados de compras, onde cada compra (transação) consiste de vários artigos (itens) comprados por um consumidor. A aplicação de técnicas de análise de associação neste conjunto de transações pode revelar afinidades entre uma coleção de itens. Estas afinidades entre itens são representadas por regras de associação. Uma regra expõe, em uma forma textual, quais itens implicam a presença de outros itens. A Figura 6.1 ilustra uma regra derivada de uma análise de compras.

Em geral a regra tem a forma "Se X, então Y", onde X é chamado

de *corpo da regra* e Y de *cabeça da regra*. Algoritmos de associação são bastante eficientes na derivação de regras, deixando para o analista o desafio de fazer um julgamento sobre a validade e a importância das regras derivadas. Para isto dois parâmetros são importantes: o *fator de suporte* e o *fator de confidência*.

O fator de suporte indica a ocorrência relativa da regra de associação detectada dentro do conjunto de dados de transações. Por ser determinado pela razão entre o número de transações que sustentam a regra e o número total de transações, este fator é uma medida relativa. Uma transação sustenta a regra "Quando X, então Y" se itens X e Y na regra também ocorrem na transação. No exemplo ilustrado na Figura 6.1, a regra é sustentada por cerca de 40% dos registros do banco de dados.

O fator de confidência de uma regra de associação é o grau com o qual a regra é verdadeira entre os registros individuais, e é calculado pela razão do número de transações sustentando a regra pelo número de transações sustentando somente o corpo da regra. No exemplo, o fator de confidência é 80%.

No exemplo acima, as associações são baseadas na contagem de ocorrências de todas as possíveis combinações de itens. Primeiro, antes da mineração para as associações, as identificações para as transações são ordenadas. Em seguida, a técnica conta as ocorrências de todas as transações com relação a cada item e cria um vetor, onde cada posição (cela) contém o número de ocorrências de cada item. As celas onde a contagem está abaixo de um deteminado nível de sustentação são suprimidas. A seguir, uma matriz com duas colunas é formada para estocar as contagens de ocorrências de cada item com um dos outros possíveis itens e, novamente, as celas são filtradas considerando um valor de corte. Quando a contagem é feita para um possível terceiro item dentro da mesma transação, uma matriz com três colunas é criada e o processo é repetido. Assim, a técnica

envolve a leitura de um conjunto de dados, sequencialmente, do alto a baixo cada vez que uma nova dimensão é adicionada e é feita uma simples contagem de ocorrências.

O desempenho computacional da técnica é afetado pelo número médio de itens por transação. Por exemplo, examinar regras com apenas um item no corpo da regra e um item na cabeça da regra exige um tempo computacional muito menor do que, examinar, digamos, oito itens no corpo e dois na cabeça.

O resultado da aplicação do algoritmo, levando-se em consideração um conjunto de transações, é uma lista de características que definem afinidades entre os itens. Estes resultados são apresentados frequentemente no formato da Figura 6.1.

Uma vez que para cada regra o fator de suporte e o fator de confidência são determinados, estatísticas adicionais podem ser calculadas, como por exemplo, o quociente entre o fator de confidência e o fator de suporte, denominado índice de associação. No exemplo acima, se o fator de sustentação para feijão no conjunto geral de transações é de 40% (isto é, em 40% das transações são comprados feijões) e o fator de confidência para a associação entre arroz e feijão é 80%, o índice da associação é 2,0. Indicando que a ocorrência esperada de feijão em uma transação é o dobro se ela ocorre em uma transação onde arroz é comprado (em uma linguagem estatística, o índice de associação é conhecido como *odds ratio*).

Claramente, grandes fatores de suporte e confidência representam um grau de relevância maior que baixos fatores. Entretanto, a implicação de baixos fatores de sustentação e confidência é a existência de várias possíveis associações de produtos que não são, na realidade, o que se busca. Enquanto que fatores de suporte e confidência muito elevados implicam na não descoberta de regras de associação.

Tabela 6.1 Banco de dados de transações comerciais.

Consumidor	Período da Transação	Itens Comprados
José Oliveira	25 de fevereiro de 2000, 16:26 hs.	Cervejas
José Oliveira	26 de fevereiro de 2000, 10:35 hs.	Vodka
João Soares	25 de fevereiro de 2000, 14:18 hs.	Guaraná, Suco
João Soares	25 de fevereiro de 2000, 15:48 hs.	Cerveja
João Soares	26 de fevereiro de 2000, 9:29 hs.	Água, Licor, Vinho
João Soares	26 de fevereiro de 2000, 15:11 hs	Gin, Licor
Pedro Tenório	25 de fevereiro de 2000, 11:06 hs.	Cerveja
Pedro Tenório	26 de fevereiro de 2000, 17:45 hs.	Água, Gin, Vinho
Pedro Tenório	27 de fevereiro de 2000, 18:04 hs.	Vodka, Soda
José Zappa	25 de fevereiro de 2000, 08:55 hs.	Guaraná, Vodka

A chance do aparecimento de uma correlação ilegítima cresce exponencialmente com o tamanho do conjunto de dados. O que nos leva a dizer que qualquer grande conjunto de dados certamente conterá alguma correlação.

Uma vantagem da análise de associação é a sua simplicidade. Entretanto, a não existência de uma simples maneira de se considerar o valor comercial da associação pode ser vista como uma desvantagem

da técnica. Em se tratando de regras de associação, a venda de um produto caro conta tanto quanto a venda de um produto barato (Cabena, *et.al.*, 1998).

6.3 Característica Sequencial

O conjunto de dados da Tabela 6.1 mostra algumas transações em detalhes, incluindo nome do consumidor, data e hora da venda e itens comprados, ocorridas em uma loja de bebidas. O conjunto de dados está ordenado pelo sobrenome do consumidor e pelo período da transação. Por exemplo, José Oliveira visitou a loja em dois consecutivos dias. Ele comprou cerveja no primeiro dia, vodka no segundo dia.

A Tabela 6.2 mostra as sequências de transações dos consumidores organizadas segundo o tempo. Cada conjunto de parênteses indica uma transação que inclui um ou mais itens.

Tabela 6.2 Sequências de transações dos consumidores.

Consumidor	Sequências dos consumidores
José Oliveira	(Cervejas)(Vodka)
João Soares	(Guaraná, Suco)(Cerveja)(Água, Licor, Vinho)
	(Gin, Licor)
Pedro Tenório	(Cerveja)(Água, Gin, Vinho)(Vodka, Soda)
José Zappa	(Vodka)

Técnicas de busca de característica sequencial detectam características entre transações de tal forma que a presença de um conjunto de itens é seguido por um outro conjunto de itens em um

banco de dados de transações em um período de tempo. A técnica determina a frequência de cada combinação de transações que pode ser produzida nas sequências de consumidores e disponibiliza as características sequenciais cujas ocorrências relativas são maiores que um nível de suporte mínimo requerido. A Tabela 6.3 apresenta as características sequenciais com suporte maior que 40%. A característica sequencial "cerveja é comprada em uma transação anterior a transação em que vodka é comprada" ocorre em dois dos quatro consumidores.

Tabela 6.3 Características Sequenciais com Suporte > 40%.

Características Sequenciais com Fator de Sustentação > 40%	Consumidores de Apoio
(Cervejas)(Vodka)	José Oliveira, Pedro Tenório
(Cerveja)(Vinho, Água)	João Soares, Pedro Tenório

As vantagens e desvantagens da análise de associação discutidas anteriormente também se aplicam em características sequenciais. Além disso, alguns pontos adicionais devem ser ressaltados. Primeiro, o fator de sustentação deve ser especificado. Segundo, um grande número de registros é necessário para assegurar uma representatividade do número de transações por consumidor. Terceiro, é requerido um novo campo no banco de dados para representar a identificação do consumidor; nem toda empresa, principalmente as pequenas, tem este campo armazenado no banco de dados de transações.

Capítulo 7

Análise de *Cluster*

7.1 Introdução

A partição de uma dada população em grupos de similares é exigida em várias aplicações. O objetivo básico desse procedimento está relacionado, entre outros, ao desenvolvimento de alguma estratégia de negócio idealizada para grupo de clientes específicos, para um possível aumento na eficiência comercial e no relacionamento venda/consumo.

Análise de *cluster* é uma técnica que visa detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso da existência, determinar estes grupos. Uma revisão detalhada do tópico pode ser encontrada em Everitt (1993).

7.2 Partição

A técnica é baseada na obtenção de uma determinada partição que otimiza alguma função objetivo, conhecida como critério de partição (Michaud, 1997).

Por definição, uma partição ou *cluster* é um subconjunto de todos os possíveis subconjuntos mutuamente excludentes e disjuntos de uma população. A população toda forma um *cluster*, que pode ser dividido em dois ou mais *clusters*, que podem ser novamente divididos em dois ou mais *clusters* e, assim por diante, incluindo até a partição onde cada elemento é o único elemento do *cluster*. Pesquisar todas essas possíveis partições é, usualmente, impraticável. A técnica análise de *cluster* utiliza várias estratégias de busca para obtenção de uma solução aproximadamente ótima. Duas questões devem ser consideradas neste contexto. Que critério de partição adotar? Que medida de similaridade utilizar? Estas questões são consideradas na seção seguinte.

7.2.1 Critérios da Partição

Considere uma população de n elementos descritos por m atributos. Intuitivamente, de acordo com o espaço m-dimensional dos atributos, a técnica de *cluster* baseia-se no fato de que o critério de partição deve promover a menor distância entre os elementos de um mesmo *cluster* e a maior distância entre os elementos de *clusters* diferentes. Dessa forma, o critério de partição escolhido, denotado como função objetivo, $F(P)$, para uma dada partição P**,** é a implementação numérica dessa noção intuitiva (Michaud, 1997). Existem vários critérios de partição que podem ser encontrados em detalhes em Bussab *et.al.* (1990).

Um exemplo de critério de partição é baseado na distância Euclidiana quadrática média entre os n elementos (Johnson e Wichern, 1982). Considere um vetor $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{im})'$ de m atributos do elemento i. O vetor de médias da k-ésima classe, C_k, a qual contém n_k elementos, é dada por

$$\overline{\mathbf{x}}_k = (\overline{x}_{k1}, \overline{x}_{k2}, ..., \overline{x}_{km})',$$

onde $\overline{x}_{ka} = (1/n_k) \sum_{i \in C_k} x_{ia}$, para $a = 1, ..., m$.

A distância Euclidiana quadrática média entre cada elemento e a sua classe média é definida como (Michaud, 1997)

$$F(P) = \frac{1}{n} \sum_{k=1}^{P} \sum_{i \in C_k} \sum_{a=1}^{m} (x_{ia} - \overline{x}_{ka})^2.$$

O objetivo básico da análise de *cluster* consiste na obtenção de uma partição que minimize a expressão acima. Na literatura estatística $nF(P)$ é conhecida como soma de quadrados dentre grupos.

7.2.2 Técnicas de Partição

A procura por uma partição ótima pode muitas vezes ser impraticável devido a enorme quantidade de partições possíveis. Procedimentos que minimizam este problema podem comumente ser divididos em dois grupos: métodos hierárquicos e não hierárquicos.

Métodos Hierárquicos

Esses métodos englobam técnicas que buscam hierarquicamente os grupos. Uma possibilidade é considerar o número máximo n de *clusters*, onde cada *cluster* é composto por um único elemento e, interativamente, agrupar cada par de *clusters* em um novo, decrescendo o número de *clusters* na ordem de um. Outra possibilidade consiste em partir de um único *cluster*, englobando todos os elementos, e iniciar um processo de subdivisões sucessivas. Esses métodos são conhecidos como hierárquicos aglomerativos e hierárquicos divisivos, respectivamente.

A escolha do par de *clusters* que deve ser agrupado (nos aglomerativos) ou do *cluster* que deve ser dividido (nos divisivos) é feita pelo valor da função objetivo obtida pelo agrupamento ou pela divisão.

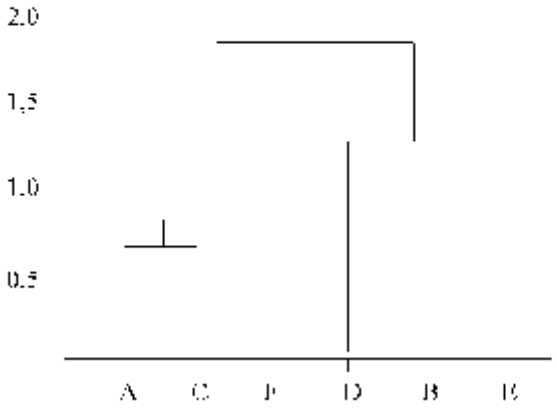

Figura 7.1: Dendograma com quatro clusters para um exemplo hipotético.

Existem vários procedimentos hierárquicos que diferem somente na escolha do critério de partição.

O processo de fusão ou divisão iterativa cria uma hierarquia dos *clusters*. O número de *clusters* pode sugerir o critério de parada do processo tanto de fusão quanto de divisão.

Uma desvantagem desses métodos hierárquicos refere-se à possibilidade de serem impraticáveis para grandes conjuntos de dados, devido a complexidade computacional (Michaud, 1997).

Métodos Não Hierárquicos

Esses métodos procuram de forma direta por uma partição aproximadamente ótima dos n elementos sem a necessidadade de associações hierárquicas. Primeiramente, uma partição inicial com um determinado número k de *clusters* deve ser considerada. A seguir, seleciona-se, então, uma partição dos n elementos em k *clusters* que otimize algum critério. A seleção da melhor partição deve ser feita através de algum procedimento específico, uma vez que pesquisar todas as possíveis partições tornaria o problema de difícil solução com k^{n-1} partições a serem pesquisadas (Bussab *et.al.*, 1990). Uma das técnicas mais conhecidas entre os métodos não-hieráquicos é a técnica das k-médias. A base desta técnica é o critério descrito na seção anterior.

7.3 Apresentação dos *Clusters*

Um aspecto importante na análise de *cluster* é a apresentação dos grupos obtidos na etapa intermediária entre a escolha da função objetivo a ser minimizada e a obtenção da partição que minimize tal função. A Figura 7.1 ilustra a descoberta de 4 grupos em um problema hipotético baseado em Bussab et.al. (1990). Esse gráfico é comumente conhecido na literatura estatística como *dendograma* e é a forma gráfica mais comum de representação dos *clusters*. Outras representações podem ser encontradas em Johnson e Wichern (1982).

Capítulo 8

Redução de Dimensionalidade

8.1 Introdução

Um problema comum que surge em *data mining* é a necessidade de redução da dimensionalidade do vetor de variáveis originais ou a criação de um novo vetor de menor dimensão cujos componentes são combinações lineares das variáveis originais. Esta necessidade é obvia principalmente quando um grande número de variáveis está disponível para análise.

A redução da dimensionalidade do vetor de variáveis originais pode ser feita segundo a opinião de especialista (critério informal de redução muitas vezes utilizado no dia-a-dia), ou através de critérios estatísticos, conhecidos como técnicas de seleção de variáveis (Draper e Smith, 1998).

A criação de um novo vetor de variáveis de menor dimensão, cujos componentes são combinações lineares das variáveis originais, pode ser conduzida via uma técnica de análise multivariada con-

Figura 8.1: Formas de redução.

hecida como Análise de Componentes Principais (Anderson, 1984; Johnson e Wichern, 1982).

A Figura 8.1 resume as diferentes formas de redução. Entre as técnicas que selecionam algumas das variáveis originais pode-se citar i) todas as possíveis regressões; ii) melhor sub-conjunto de regressão; iii) eliminação backward; iv) regressão stepwise, entre outras. Estas técnicas podem ser encontradas no livro de Draper e Smith (1998). A técnica componentes principais, entretanto, será discutida a seguir.

8.2 Análise de Componentes Principais

Os componentes principais são determinados através de combinações lineares padronizadas de $\mathbf{x}$, onde $\mathbf{x}$ é o vetor das variáveis originais. Uma combinação do tipo $\mathbf{l}'\mathbf{x}$, onde $\mathbf{l} = (l_1, \ldots, l_p)'$ é um vetor de constantes, é denotada combinação linear padronizada (CLP) se $\sum l_i^2 = 1$.

Um dos objetivos da técnica análise de componentes principais é encontrar uma CLP das variáveis originais de tal sorte que a va-

riância seja maximizada. Especificamente, a análise de componentes principais busca algumas combinações que possam sumarizar os dados, perdendo o mínimo possível de informação. Ou seja, se o vetor $\mathbf{x}$ contém p componentes, e estes componentes são necessários para reproduzir a variabilidade total do sistema, grande parte desta variabilidade poderá ser explicada por um número pequeno k de combinações denominadas componentes principais. Estes k componentes principais podem, então, substituir as p variáveis originais e o conjunto de dados passa a ter então, n medidas em k componentes principais. Esta tentativa de reduzir a dimensionalidade pode ser considerada como uma *sumarização parcimoniosa* dos dados.

8.2.1 Definição

Suponha que um vetor $\mathbf{x}$ de p componentes tenha matriz de variâncias-covariâncias $\mathbf{\Sigma}$ com autovalores $\lambda_1, \ldots, \lambda_p$ e vetor de médias $\mathbf{0}$, onde

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix}.$$

O elemento σ_{ij} é referido como a covariância entre a i-ésima e j-ésima variáveis e o elemento σ_{ii} é referido como a variância da i-ésima variável. A distribuição estatística de $\mathbf{x}$ será irrelevante no desenvolvimento das idéias; todavia, se $\mathbf{x}$ seguir distribuição normal, alguns significados podem ser atribuídos aos componentes principais.

Considere as CLP's

$$\begin{aligned} Y_1 &= \mathbf{a}_1^{'}\mathbf{x} = a_{11}x_1 + \ldots + a_{1p}x_p, \\ Y_2 &= \mathbf{a}_2^{'}\mathbf{x} = a_{21}x_1 + \ldots + a_{2p}x_p, \\ &\vdots \\ Y_p &= \mathbf{a}_p^{'}\mathbf{x} = a_{p1}x_1 + \ldots + a_{pp}x_p. \end{aligned}$$

As variâncias dos Y_i's e as covariâncias entre os Y_i's e os Y_k's, $i, k = 1, 2, \ldots, p$, são dadas por (Anderson, 1984),

$$\begin{aligned} Var(Y_i) &= \mathbf{a}_i^{'}\mathbf{\Sigma}\mathbf{a}_i \qquad i = 1, 2, \ldots, p \\ Cov(Y_i, Y_k) &= \mathbf{a}_i^{'}\mathbf{\Sigma}\mathbf{a}_k \qquad i, k = 1, 2, \ldots, p. \end{aligned}$$

As CLP's não correlacionadas, cujas variâncias são as maiores possíveis, determinam os componentes principais. Assim, o primeiro componente principal é a CLP cuja variância é máxima e o j-ésimo componente principal é a CLP de $\mathbf{x}$ não correlacionada com os $(j-1)$-ésimos componentes principais e tem a maior variância entre todas as CLP's restantes.

Resultados de análise de componentes principais mostram que, se $\mathbf{x}$ tiver matriz de variâncias-covariâncias $\mathbf{\Sigma}$, com pares de autovalor-autovetor dados por $(\lambda_1, \mathbf{e}_1), \ldots, (\lambda_p, \mathbf{e}_p)$, tal que $\lambda_1 \geq \ldots \geq \lambda_p$, então a variância do i-ésimo componente principal é dado por λ_i e o i-ésimo componente principal é dado por $Y_i = \mathbf{e}_i^{'}\mathbf{x}$. A variância total da população $\sigma_{11} + \sigma_{22} + \ldots + \sigma_{pp}$ será igual a soma dos autovalores $\lambda_1 + \lambda_2 + \ldots + \lambda_p$ e, consequentemente, a proporção da variância total explicada pelo k-ésimo componente principal é dada por,

$$\left(\begin{array}{c} \text{Proporção da variância} \\ \text{total da população, explicada} \\ \text{pelo } k\text{-ésimo componente principal} \end{array} \right) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p},$$

$k = 1, 2, \ldots, p$.

8.2.2 Número de Componentes Principais

Uma regra prática de determinar o número de componentes principais é a de incluir somente os componentes que explicam, digamos, entre 80 a 90% da variância total. Uma outra regra (Kaiser, 1958) exclui aqueles componentes principais cujos autovalores são menores que a média. Um critério visual útil para determinar o número de componentes principais apropriado é o *scree plot.* Um *scree plot* é um gráfico do i-ésimo autovalor λ_i contra i, onde os autovalores devem estar ordenados do maior para o menor. O número de componentes apropriados é tomado como sendo o ponto a partir do qual todos os autovalores restantes são relativamente pequenos, com a mesma magnitude.

8.2.3 Componentes Principais Via Matriz de Correlação

Os componentes principais não são invariantes em escala. Se as variáveis estão em escalas diferentes, como, por exemplo, peso em kilogramas, altura em centímetros e idade em anos, então, para assegurar que todas as variáveis medidas tenham o mesmo peso, padroniza-se todas as variáveis para que as mesmas tenham variância unitária e os componentes principais são encontrados então, através da matriz de correlação, ao invés da matriz de variâncias-covariâncias.

Suponha que um vetor $\mathbf{x}$ de p componentes tenha vetor de médias $\boldsymbol{\mu}$, matriz de variâncias-covariâncias $\boldsymbol{\Sigma}$, como dada anteriormente, e

matriz de correlação $\boldsymbol{\rho}$, onde

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \qquad \text{e} \qquad \boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}.$$

Padronizando as variáveis do vetor $\mathbf{x}$ na forma

$$\mathbf{z} = (\mathbf{V}^{1/2})^{-1}(\mathbf{x} - \boldsymbol{\mu}),$$

onde

$$\mathbf{V} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix},$$

o vetor $\mathbf{z}$ tem vetor de médias $\mathbf{0}$ e matriz de variâncias-covariâncias igual a matriz de correlação de $\mathbf{x}$. Assim, os componentes principais de $\mathbf{z}$ podem ser obtidos através dos autovetores da matriz de correlação $\boldsymbol{\rho}$. Se $(\lambda_1^*, \mathbf{e}_1^*), \ldots, (\lambda_p^*, \mathbf{e}_p^*)$, representam os pares de autovalor-autovetor da matriz de correlação $\boldsymbol{\rho}$, o i-ésimo componente principal do vetor de variáveis padronizadas $\mathbf{z}$ é dado por $Y_i = \mathbf{e}_i^{*'}\mathbf{z}$, $i = 1, 2, \ldots, p$. A variância total da população será igual a p e, consequentemente, a proporção da variância total explicada pelo k-ésimo componente principal é dado por,

$$\left(\begin{array}{c}\text{Proporção da variância}\\ \text{total da população (padronizada) explicada}\\ \text{pelo k-ésimo componente principal}\end{array}\right) = \frac{\lambda_k}{p},$$

$k = 1, 2, \ldots, p$.

8.2.4 Componentes Principais Amostrais

A contra partida amostral das seções anteriores é agora discutida. Seja $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ uma amostra de tamanho n de uma população p-dimensional com vetor de médias $\boldsymbol{\mu}$ e matriz de variâncias-covariâncias $\boldsymbol{\Sigma}$. Seja $\overline{\mathbf{x}}$ o vetor de médias amostrais, $\mathbf{S}$ a matriz de variâncias-covariâncias amostrais e $\mathbf{R}$ a matriz de correlação amostral. Sejam $(\widehat{\lambda}_1, \widehat{\mathbf{e}}_1), \ldots, (\widehat{\lambda}_p, \widehat{\mathbf{e}}_p)$ pares de autovalor-autovetor da matriz $\mathbf{S}$. O i-ésimo componente principal amostral é dado por $\widehat{Y}_i = \widehat{\mathbf{e}}_i'\mathbf{x}$, $i = 1, 2, \ldots, p$, onde $\widehat{\lambda}_1 \geq \ldots \geq \widehat{\lambda}_p$.

Para variáveis padronizadas os componentes principais são obtidos considerando a matriz $\mathbf{R}$. Sejam $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n$ observações padronizadas com matriz de variâncias-covariâncias amostrais $\mathbf{R}$, e sejam $(\widehat{\lambda}_1^*, \widehat{\mathbf{e}}_1^*)$, $\ldots$, $(\widehat{\lambda}_p^*, \widehat{\mathbf{e}}_p^*)$ pares de autovalor-autovetor dessa matriz. O i-ésimo componente principal amostral é dado por $\widehat{Y}_i = \widehat{\mathbf{e}}_i'\mathbf{z}$, $i = 1, 2, \ldots, p$, onde $\widehat{\lambda}_1^* \geq \ldots \geq \widehat{\lambda}_p^*$.

8.2.5 Uma Aplicação

A aplicação desta técnica de redução de dimensionalidade é comum nas situações onde o número de variáveis levantadas é muito grande. Como exemplo (ver detalhes no Capítulo 12), considere um companhia de seguros com 280.000 clientes e com 250 variáveis sendo medida para cada caso. É do interesse da empresa predizer quem

cancelaria as suas apólices. Para a construção do modelo, e mesmo para facilitar a interpretabilidade, é extremamente importante que o vetor de variáveis seja reduzida o máximo possível.

8.3 Comentários Finais

A técnica análise de componentes principais discutida neste capítulo não leva em consideração a distribuição estatística de $\mathbf{x}$ no desenvolvimento das idéias. Se $\mathbf{x}$ seguir distribuição normal multivariada, alguns significados podem ser atribuídos aos componentes. Entre eles, o fato do vetor de componentes principais também seguir distribuição normal multivariada, o que é uma informação relevante para inferência.

Gráficos dos pares de componentes principais são úteis para verificar a suposição de normalidade, bem como para detectar observações suspeitas. Para auxiliar na verificação de normalidade analisase os diagramas de dispersão de pares dos primeiros componentes principais. Estes diagramas devem apresentar formas elípticas. Para auxiliar na detecção de observações suspeitas analisa-se os diagramas de dispersão de pares dos últimos componentes principais. Pontos extremos serão considerados suspeitos.

Capítulo 9

Exemplos de Aplicação

Aplicações de *data mining* têm sido observadas em várias áreas do conhecimento, entre elas, finanças, saúde, criminologia, sociologia, ecologia, saneamento básico, climatologia, atuária, manufatura, controle de qualidade, *marketing* e medicina.

Este capítulo apresenta algumas situações reais, envolvendo volumes grandes de dados, em quatro grandes áreas de aplicação: comércio, finanças, seguros e saúde, que se utilizaram de *data mining* para encontrar soluções para os problemas levantados.

9.1 *Data Mining* em Comércio

As entidades comerciais, em geral, retêm grande quantidade de informações sobre seus clientes. Nesta seção são descritas duas aplicações de *data mining* neste setor. O primeiro descreve a aplicação de técnicas de *data mining* para alocação de novas filiais de uma grande cadeia de lojas e o segundo para a identificação de clientes leais.

9.1.1 Alocação de Novas Filiais

Nos últimos anos o comércio tradicional, geralmente localizado em ruas, tem sofrido com a crescente preferência dos consumidores por *shopping center* e *superstores*.

Uma cadeia de lojas iniciou um ambicioso projeto de aumento de suas filiais, e uma de suas necessidades básicas era a determinação de locais que seriam promissores do ponto de vista de vendas.

O sucesso de uma determinada filial pode ser consequência da interferência de muitos fatores, como por exemplo, localização da loja, histórico das vendas, número de competidores e suas localizações, habilidades dos vendedores e executivos na administração, entre outros. Tanto os fatores quanto seus efeitos devem ser quantificados para que se possa predizer o possível sucesso de uma filial hipotética, com uma confiança suficiente.

No projeto foram utilizados dados de venda, ao longo de 10 anos, de mais de 3000 produtos de 200 filiais, dados de um questionário sobre a qualidade de várias possíveis localizações e dados demográficos relacionados a área de atuação de cada filial. Criando, assim, um grande banco de dados. Algumas variáveis foram descartadas no passo seleção e outras no preprocessamento dos dados por não apresentarem informações relevantes. Transformações, tais como, venda mensal de produto e venda por setor, foram consideradas. Esta parte do processo (criação do banco de dados, seleção, preprocessamento e transformação de variáveis) consumiu cerca de 70% do tempo total gasto no projeto. A seguir, informações do banco de dados final foram extraídas através de um modelo de classificação, desenvolvido para predizer o retorno financeiro das lojas. Concluindo o processo KDD com os passos de assimilações, interpretações e avaliações, o modelo foi testado nas filiais existentes e está sendo utilizado para auxiliar nas decisões de novos investimentos em futuras filiais.

9.1.2 Lealdade do Cliente

Uma cadeia de supermercados, que atende a mais de 2 milhões de pessoas por semana, preocupada com a manutenção e possível aumento do número de clientes, tinha interesse em estudar o comportamento de compra dos seus consumidores.

Em uma primeira etapa, a cadeia disponibilizou aos clientes cartões de lealdade, cartões tipo *credit card*. Esses cartões são, em geral, distribuídos a custo zero. O cliente para recebê-lo deve fornecer algumas informações pessoais, tais como, nome, sexo, idade, endereço, situação matrimonial, números de filhos. Em cada transação rea- lizada, o cartão é registrado juntamente com a compra no sentido de estabelecer uma ligação entre compra e cliente. Criando, desta forma, um banco de dados constituído de informações do tipo: quando o cliente utilizou os serviços da cadeia, em qual filial, que produtos comprou e qual foi a quantidade. Ao mesmo tempo eram incorporados ao banco informações de promoções e demostrações de produtos.

O fato de considerar apenas parte do total de clientes, somente os que possuem cartão, já se configura em uma seleção (o segundo passo do processo KDD). Outra seleção utilizada foi com relação aos tipos de produtos. Apenas aqueles que estavam disponibilizados para compra em todos os supermercados da cadeia fizeram parte do banco de dados. No passo preprocessamento dos dados foram descartados alguns produtos que deixaram de ser comercializados e clientes inadimplentes. Transformações, tais como, compra mensal por cliente, venda mensal de produto e venda por setor, também foram consideradas. Através da aplicação de técnicas de *data mining* foi possível observar a periodicidade de compra do cliente nos supermercados da cadeia, a preferência de compra e a quantidade de cada produto. Além disso, foi possível descobrir, utilizando análise

de associação, quais produtos são comprados em conjunto e o efeito da promoção de alguns produtos na venda de outros.

9.2 *Data Mining* em Finanças

Bancos, entre outras instituições financeiras, retêm grande quantidade de informações pessoais de seus clientes e as transações efetuadas por estes. Estes dados, se analisados propriamente, podem apresentar conhecimento de grande valia para tomadas de decisões futuras.

Nesta seção é apresentado um estudo de avaliação de empréstimo bancário baseado em Feelders *at. al.*,1996. Dados históricos dos clientes e seus comportamentos nos retornos dos pagamentos são utilizados para predizer os resultados de possíveis inadimplências.

9.2.1 Empréstimo Pessoal

Para obtenção de um empréstimo pessoal, as instituições de créditos exigem alguns requisitos mínimos que o aplicante deve preencher. Estes requisitos representam a garantia da instituição de que o empréstimo será pago. Neste contexto, um sistema de avaliação de crédito torna-se imprescindível.

Um sistema de avaliação de crédito tem como objetivo principal premiar (pontuar), quantitativamente, atributos que evidenciem indivíduos que tenham maior ou menor chance de pagar o empréstimo tomado. Se a pontuação atingir um determinado limite, o crédito não é negado. Existem várias formas de avaliar o tomador de empréstimo, entre elas, as técnicas discutidas no Capítulo 5.

Alguns atributos, com os respectivos graus de importância para o empréstimo, usualmente considerados, estão listados na Tabela 9.1.

O interesse principal desse tipo de estudo concentra-se em predi-

zer qual tipo de aplicante tornar-se-a um inadimplente. Uma maneira de obter tal informação é estudar em detalhes uma amostra aleatória da população de aplicantes e verificar se os mesmos foram ou não inadimplentes. Entretanto, duas possibilidade são comuns. Aplicantes com um escore de avaliação baixo têm chance zero de pertencer a amostra, uma vez que não obtiveram o crédito e não estão presentes no banco de dados e, portanto, não existe a possibilidade de saber se eles se tornariam ou não inadimplentes.

Geralmente, o aprendizado sobre o problema se restringe ao conhecimento da diferença entre pagadores e inadimplentes dentro da população de aplicantes que seriam aceitos para obterem crédito com base nas suas avaliações.

A aplicação considerada por Feelders *et. al.*, 1996, envolvia um banco de dados de 52.000 registros de empréstimos concedidos por um determinado Banco, com um total de 38 atributos. Vários problemas associados ao banco de dados foram reportados: dados incompletos, respostas ambíguas, valores fixados em zero sem um sentido comum a todos os analistas, entre outros. Os passos de seleção e preprocessamento de dados foram fortemente empregados na tentativa de eliminar os problemas mencionados.

Para a obtenção de resultados foi utilizado o método de classificação C4.5, discutido no Capítulo 5. A Figura 9.1 mostra a árvore gerada através do uso desta técnica, para um determinado tipo de crédito. A letra **I** na Figura indica inadimplente enquanto a letra **A** indica cliente apto ao crédito.

9.3 *Data Mining* em Seguros

Uma preocupação presente nas companhias de seguros diz respeito às perdas obtidas devido ao cancelamento de apólices e ao custo para

obtenção de novos clientes.

Esta seção apresenta uma aplicação de *data mining* na previsão de cancelamentos de apólices de seguro.

9.3.1 Cancelamento de Apólice de Seguro

Uma companhia internacional de seguros com mais de um milhão de clientes, somente em seu país de origem, tem cerca de 100.000 cancelamentos de apólices a cada ano. Por esta razão, a administração da companhia tinha interesse na construção de um modelo que fizesse a previsão dos clientes mais propícios ao cancelamento.

O projeto inicial enfocou apólice de seguros de carros de clientes de um país. O conjunto de dados inicial era composto de mais de 280.000 clientes com 250 variáveis para cada caso.

Técnicas de visualização de dados, acrescidas de técnicas multivariadas foram utilizadas para a redução da dimensionalidade. Permaneceram as 30 variáveis mais significantes para a análise.

O modelo proposto foi baseado em uma rede neural. Este modelo prediz corretamente quem cancelaria as suas apólices para 90% dos casos incluídos no conjunto de dados de teste. Para cada cliente foi dado um escore, o qual indica a probabilidade do mesmo vir a cancelar sua apólice.

Com base nos resultados obtidos, os administradores da companhia deflagaram campanhas publicitárias direcionadas aos clientes propícios ao cancelamento, reduzindo drasticamente a porcentagem de cancelamento de apólice de seguros de carros.

9.4 *Data Mining* em Medicina

É comum conjuntos de dados provenientes de estudos médicos serem compostos de uma grande massa de dados.

Nesta seção três aplicações de *data mining* a dados médicos são descritas. O primeiro estudo relaciona-se a transplante de rim, o segundo é voltado a um estudo em oncologia e o último diz respeito a modelagem da corrosividade da pele.

Tabela 9.1. Alguns atributos importantes.

Atributo	Valores	Importância
Idade do Aplicante	< 26	negativa
	≥ 45	positiva
Telefone	não fornecido	negativa
Casa Própria?	sim	positiva
Emprego	tempo parcial	negativa
	desempregado	negativa
Tempo no Emprego	mais de 4 anos	positiva
É Cliente do Banco?	não	negativa
Número de Empréstimos Atuais	1 ou mais	negativa
Número de Faltas de Pagamento	1 ou mais	negativa

9.4.1 Sobrevivências de Pacientes Transplantados

Em um estudo realizado na unidade de transplante de rim de um centro médico, mais de 150 diferentes medidas foram feitas para cada paciente transplantado, durante um período de 20 anos. O inte- resse dos pesquisadores era descobrir os fatores que afetavam as taxas de sobrevivência de pacientes transplantados. Após a seleção e preprocessamento dos dados, foi utilizado um método de redução de

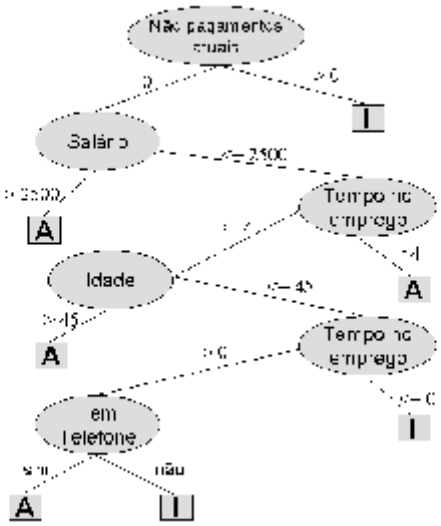

Figura 9.1: Árvore de classificação C4.5.

dimensionalidade na tentativa de diminuir o, ainda, elevado número de variáveis. Em seguida, técnicas de *data mining* em conjunto com métodos estatísticos vinculados à análise de sobrevivência (Cox e Oakes, 1984) foram empregados.

Com isto, foi identificada uma correlação positiva, previamente desconhecida, entre o tempo de espera até o transplante e as taxas de sobrevivência após cinco anos da operação. Foi, também, observado que as taxas de sobrevivências de longo tempo são diretamente influenciadas pelo período de tempo que o rim a ser utilizado no transplante ficou armazenado.

9.4.2 Redução do Efeito Colateral de Quimioterapia

Um estudo sobre o efeito da ansiedade do paciente sobre a náusea relacionada a quimioterapia, realizado em um centro americano de pesquisas oncológicas, considerou quatro grupos de pacientes. Para cada grupo foi ministrada uma dose diferente de uma droga redutora da ansiedade e observado os resultantes níveis de náuseas em conjunto com outras variáveis. O estudo prolongou-se por um período de vários anos com muitos pacientes.

Técnicas de visualização de dados e análise discriminante não revelaram diferenças entre os grupos previamente existentes, mas sim entre subgrupos de indivíduos para os quais o nível pessoal de ansiedade variou. Esta informação indicou para os pesquisadores que a intervenção, realizada através da droga redutora da ansiedade, não foi eficaz. Os pesquisadores também notaram a necessidade de uma forma de intervenção alternativa para redução do efeito colateral da quimioterapia e a necessidade da identificação de subgrupos de pacientes para os quais o tratamento seria mais eficiente.

9.4.3 Modelando a Corrosividade da Pele

A corrosividade da pele pode ser causada por vários mecanismos e está diretamente ligada a estrutura e propriedade dos elementos químicos utilizados. Por definição, todo componente que causa dano irreversível em quatro horas de contato com a pele é classificado como corrosivo.

Quando um novo produto está para ser lançado no mercado deve-se verificar e aprovar sua segurança. Neste contexto, novos procedimentos devem ser propostos para redução da necessidade por testes em animais. A idéia relaciona-se ao estabelecimento dos efeitos que as substâncias potencialmente corrosivas provocam na pele.

O objetivo do estudo é classificar os diversos tipos de substâncias

através de vários atributos descritivos, como por exemplo, volume molecular e ponto de fusão.

Técnicas de *data mining* foram utilizadas para modelar a corrosividade de bases, ácidos orgânicos e fenóis. Redes neurais foram treinadas para julgar a corrosividade de substâncias com classificação conhecida utilizando-se atributos descritivos.A classificação era dada em uma escala de 0 a 1. Substâncias altamente corrosivas têm classificação próxima de 1. Os resultados corroboram com classificações de substâncias conhecidas e substâncias com propriedades corrosivas de fronteira são indicadas com facilidade.

Esse procedimento de modelagem reduz o custo e o tempo de desenvolvimento de novos produtos, minimizando a utilização de animais para teste.

Bibliografia

[1] Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*. Nova Iorque: Wiley.

[2] Apté, C. e Weiss, S. (1997). Data mining with decision trees and decision rules. *Future Generation Computer System.* 13, 197–210.

[3] Box, G.E.P. and Tiao, G.C. (1973). *Bayesian Inference in Statistical Analysis*. Nova Iorque: Wiley.

[4] Breiman, L.L., Friedman, J., Olshen, R. e Stone, C. (1984). *Classification and Regression Trees.* Monterrey: Wadsworth.

[5] Bussab, W.O., Miazaki, E.S. e Andrade, D.F. (1990). *Introdução à Análise de Agrupamentos*. IX Simpósio Brasileiro de Probabilidadde e Estatística. IME-USP, São Paulo.

[6] Bussab, W.O. e Morettin, P.A. (1985). *Estatística Básica*, 4^a Edição. Editora Atual: São Paulo. .

[7] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. e Zanasi, A. (1998). *Discovering Data Mining — from Concept to Implementation*. Upper Saddle River: Prentice Hall.

[8] Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society A*, 158, 419–466.

[9] Clark, L.A. e Pregibon, D. (1992) Tree-based Models. In Statistical Models in S (Eds. Chambers , J.M e Hastie, T.J.), 317–419. Pacific Grove: Wadsworth.

[10] Clementine User Guide, a data mining toolkit (2000). Internet: http://www.spss.com/clementine/.

[11] Cox, D.R. e Oakes, D. (1984). *Analysis of Survival Data*. Londres: Chapman and Hall.

[12] Craven, M.W. e Shalvlik, J.W. (1997). Using neural networks for data mining. *Future Generation Computer Systems*, 13, 211–229.

[13] Dilly, R. (1999). *Data Mining — An Introduction.* Belfast: Parallel Computer Centre, Queens University.

[14] Draper, N.R. e Smith, H. (1981). *Applied Regression Analysis*. 2^a Edição. Nova Iorque: Wiley.

[15] Everitt, B.S. (1993). *Clustering Analysis*. Sevenoaks: Edward Arnold.

[16] Faraggi, D. e Simon, R. (1995). A neural network model for survival data. *Statistics in Medicine*, 14, 73–82.

[17] Fayyad, U.M., Djorgovski, S.G. e Weir, N. (1996). Automating the analysis and cataloging of sky surveys, In *Advances in Knowledge Discovery and Data Mining*, Eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, Menlo Park, CA: AAAI Press, 471–493.

[18] Ferreira. C. A (1999). *Comparação da capacidade preditiva da regressão logística, CART e redes neurais*. Tese de Mestrado, DEs/UFMG.

[19] Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics,* 8**,** 376–388.

[20] Fletcher, R. (1987). *Practical Methods of Optimization*. Nova Iorque: Wiley.

[21] Garthwaithe, P.H. et al. (1995). *Statistical Inference*. Hemel Hempstead, Hertfordshire: Prentice Hall International.

[22] Glymour,C. Madigan, D., Pregibon, D., Smyth, P. (1996), Statistical inference and data mining. *Comunication of the ACM*. 39, 35–41.

[23] Hand, D.J. (1998). Data mining: statistics and more? *The American Statistician*, 52, 112–118.

[24] Hair, J. F. Anderson, R. E., Tatham, R. L. e Black, W. C. (1998).*Multivariate Data Analysis*. Upper Saddle River, New Jersey: Prentice Hall.

[25] Hosmer, D.W. e Lemeshow, S. (1989). *Applied Logistic Regression*. Nova York: Wiley.

[26] Hong, S. J. (1997). R-MINI: An interative approach for generating minimal rules from examples. *IEEE Trans. Knowledge and Data Engrg.*

[27] Huber, P.J. (1997). From large to huge: a statistician reaction to KDD and DM. *Proceedings da III International Conference on Knowledge Discovery and Data Mining*, AAAI Press.

[28] Inmon, W. H. (1993). *Building the Data Warehouse.* Nova Iorque: Wiley.

[29] Inmon, W.H. (1996) The data warehouse and data mining. *Communication of the ACM*, 39, 49–50.

[30] Johnson, R.A. e Wichern, D.W. (1982). *Applied Multivariate Statistical Analysis*. Londres: Prentice-Hall.

[31] Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23, 187–200.

[32] KnowledgeSEEKER user guide (2000). Internet: http://www.angoss.com/.

[33] Lael, A.O.(1996). Comparação de dois indicadores da desnutrição materna usando regressão e classificação por árvore e logística multinomial. Tese de Doutorado Faculdade de Saúde Pública, USP

[34] Lovell, M.C. (1983). Data Mining. *The review of Economics and Statistics*. 65, 1.

[35] Mardia, K.V., Kent, J. T. e Bibby, J. M. (1989). *Multivariate Analysis*. Londres: Academic Press.

[36] McCullagh, P. e Nelder, J.A. (1989). *Generalized Linear Models*. Londres: Chapman and Hall.

[37] Michalski, R., Mozetic, I.; Hong, S. J. e Lavrac, N. (1986). The multi-purpose incremental learning system AQ15 and its testing application to three medical domains. In *Proceedings AAAI-86,* 1041–1045.

[38] Michaud, P.(1997) Clustering techniques. *Future Generation Computer Systems*, 13, 135–147.

[39] Parsaye, K. e Chignell, M. (1993). *Intelligent Database Tools & Applications.* Nova Iorque: Wiley.

[40] Pereira, B.B. e Rodrigues, C.V.S. (1998). *Redes Neurais em Estatística.* XIII Simpósio Brasileiro de Probabilidadde e Estatística. Caxambu.

[41] Potts, W.J.E. (1998). *Data Mining Primer: Overview of Applications and Methods.* SAS Institute Inc.

[42] Quinlan, J. R.(1993). *C4.5: Programs for machine learning.* Los Altos, CA: Morgan Kaufmann.

[43] Ripley, B. (1993). Statistical aspects of neural networks. In *Network and Chaos — Statistical and Probabilistics Aspects*, Eds. O.E. Bradorff-Nielsen, J.L. Jensen e W.S. Kendall. Londres: Chapman and Hall.

[44] Segal, M.R. (1992). Tree-Structured Methods for Longitudinal Data. *Journal of the American Statistical Association*, 87, 407–418.

[45] Segal, M.R. (1988). Regression Trees for Censored Data. *Biometrics*, 44, 35–47.

[46] Taylor, P.C e Jones, M.C. (1996) Splitting Criteria For Regression Trees. *Journal of Statistical Computation and Simulation*, 55, 267-285.

[47] Taylor, C.C., Nakhaeizadeh, G. e Kunisch, G. (1997). Statistical aspects of classification in drifting populations. In *Preliminary Papers of the 6th International Workshop on Artificial Inteligence and Statistics*, Fort Lauderdale, FL, 521–528.

[48] Tkach, D. (1998). *Text Mining Technology — Turning Information Into Knowledge.* A White paper from IBM.

[49] Tukey, J. (1973). *Exploratory Data Analysis*. Nova Iorque: McMillan.

[50] Wasserman, P.D. (1989). *Neural Computiong Theory and Practice*. Nova Iorque: Van Nostrand Reinhold.

[51] Zhang, H. et al. (1996). A Tree-Based Method of Analysis for Prospective Studies. *Statistics in Medicine*, 15, 37–49.