Uso de Técnicas de Data Mining para Imputação de Dados
Uma Aplicação ao Censo Demográfico de 1991

Ana Cristina Pessanha Torres

IBGE
Instituto Brasileiro de Geografia e Estatística

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA – IBGE

ESCOLA NACIONAL DE CIÊNCIAS ESTATÍSTICAS – ENCE

Uso de Técnicas de Data Mining para Imputação de Dados

Uma Aplicação ao Censo Demográfico de 1991

Ana Cristina Pessanha Torres

Dissertação de mestrado em Estudos Populacionais e Pesquisas Sociais, Área de Concentração em Estudos Populacionais e Demografia, apresentada à Coordenação do Mestrado da ENCE

Orientador:

Professor Dr. Kaizô Iwakami Beltrão

Rio de Janeiro

Dezembro de 2003

A meu pai, onde estiver.

A Alexandre. Companheiro e amigo, que sem seu amor e carinho não teria conseguido seguir adiante!

A minhas filhas, que enriqueceram com seu carinho e compreensão todos os dias que se contaram em todo o mestrado.

Agradecimentos

Ao professor Kaizô Iwakami Beltrão por compartilhar comigo uma parte do seu conhecimento. E por não ter desistido desta aluna.

A Sergio Cortes que me deu o incentivo para iniciar o mestrado. E percorrer este caminho que leva a Busca do conhecimento.

Aos amigos Miriam Nahas Frazão, José de Souza Pinto Guedes e Therezinha Duarte, pela confiança e apoio incondicional. Sinto-me afortunada por sua amizade;

A Odicea Arantes e toda a sua valorosa, e afetuosa, equipe de trabalho. Sempre prestativos, foram incansáveis no atendimento às minhas idas e vindas à Biblioteca.

A professora Denise Britz do Nascimento meus sinceros agradecimentos pela assistência, paciência e carinho que sempre teve comigo nos momentos que precisei de sua ajuda.

E a todos aqueles que, direta ou indiretamente, contribuíram para a realização deste trabalho.

Resumo

Este trabalho apresenta uma proposta de metodologia de estudo para ferramentas de Mineração de Dados (Data Mining) para uma possível utilização para preencher dados omissos. Através da comparação dos modelos de Regressão Logística e Rede Neural com método de *backpropagation*, exploramos, como estudo de caso, o uso do melhor resultado para determinar a existência ou não de filhos nas mulheres com variáveis do bloco de fecundidade rejeitadas por respostas inconsistentes nos dados do Censo Demográfico de 1991. A ONU apresenta uma proposta para tratamento desta informação, o método de El Badry, que considera o agregado de mulheres e não os registros individuais como seria interessante para preservar os microdados.

Abstract

The aim of this thesis is to present a methodology that uses Data Mining´s tools to fill missing data. By the comparasing of logistic Regression model with Neural Net using backpropagation method we used as case study the applicability of the best result that could determinate the existence or not of children in women that had variables of de fertility block rejected because of inconsistent with the data from Censo Demográfico 1991. The ONU presents a proposal to treat this information that is called El-Badry method which considers the amout of women not the individual records, although it doesn´t preserve microdatas.

Índice

1. Introdução

O Crescimento da população humana ultimamente constitui um dos maiores problemas de nossa época, transformando-se em um dos fatores que podem alterar as condições ambientais do nosso mundo. As conseqüências e causas deste crescimento têm sido analisado sob vários ângulos e sua problemática perpassa pelas implicações econômicas, sociais e políticas. Pode-se afirmar, em termos gerais, que fatores demográficos responsáveis pela dinâmica populacional contemporânea estão ligados à situação de dependência econômica e política dos países subdesenvolvidos. Por outro lado, como as mudanças econômicas e sociais nos países subdesenvolvidos são muito lentas e superficiais, a falta de informações que permitam mudar o padrão tradicional de comportamento reprodutivo permanece inalterado em muitos segmentos populacionais.

Este capítulo tem por objetivo apresentar a motivação para essa dissertação, os objetivos e as barreiras encontradas para o uso da mineração de dados como uma técnica para imputação dos dados coletados em um Censo, bem como a organização dessa dissertação.

1.1 Motivação

A necessidade de conhecer as características demográficas de uma população bem como prever a sua evolução no futuro, são condicionantes que justificam a elaboração de projeções demográficas.

Estimativas e projeções de população são essenciais à tomada de decisões sendo, portanto, fundamentais na orientação de administradores públicos e privados, quando da definição de ações coerentes com as reais necessidades de uma sociedade.

Como o planejamento requer previsões não só do volume e do ritmo de crescimento de uma população total como também de suas diversas parcelas, isto implica na análise profunda das variáveis componentes da dinâmica demográfica: fecundidade, mortalidade e migração, avaliando seu comportamento no passado e no presente, e formulando hipóteses de sua atuação no futuro.

No Brasil, assim como na maioria dos países em desenvolvimento, não existe um sistema de registro de nascidos vivos completo, que permita, através das informações que coleta, sua utilização direta e atualizada nos cálculos dos parâmetros representativos das taxas de fecundidade. O esforço de analisar as tendências históricas da fecundidade no Brasil esbarra na escassez e comparabilidade de informações. Como as informações provenientes do Registro Civil [1] apresentavam até recentemente, problemas de cobertura do fenômeno a ser estudado, a maioria dos trabalhos de fecundidade baseiam-se em informações censitárias (ALTMANN; FERREIRA,1979).

O Censo Demográfico, principal fonte de informação de população no Brasil, inicia-se em 1872, quando, por ordens imperiais, determina-se o recenseamento de todos os habitantes do Império, nacionais e estrangeiros, livres e escravos, presentes ou ausentes.

A partir do Censo Demográfico de 1940, o Brasil inicia uma nova etapa histórica da Estatística populacional, sobretudo no que diz respeito à introdução de questões sobre fecundidade e mortalidade. Atendendo a padrões internacionais, visando à uniformidade de conceitos e à comparabilidade dos resultados com outras nações, o Brasil participa do

[1] Anotação oficial de todos os dados relativos aos nascimentos, casamentos, óbitos, feita por funcionário civil de um cartório. Também são feitas anotações de nulidade ou anulação do casamento, o desquite e o restabelecimento da sociedade conjugal. Dicionário Aurélio da Língua Portuguesa.

programa de Censos simultâneos proposto pelo Comitê de Censo das Américas, criado em 1946 (ALTMANN; FERREIRA,1979).

Numa pesquisa cuidadosamente planejada e executada, como é o caso do Censo Demográfico, há sempre os casos das informações prestadas estarem incompletas ou simplesmente do informante se recusar a respondê-las. Por este motivo, em qualquer pesquisa, é imprescindível fazer alguma crítica nos dados coletados, a fim de impedir que essa falta de informação afete negativamente a qualidade dos resultados.

Nas críticas de consistência utilizadas, o que normalmente se faz, é separar os dados preenchidos corretamente dos que apresentam algum tipo de problema. A questão central é o que fazer com os dados do segundo grupo? Voltar ao informante tornaria uma pesquisa da complexidade e tamanho do Censo muito onerosa aos cofres públicos e ao mesmo tempo extremamente demorada. Para contornar esse problema, o IBGE, que é a agência encarregada de executar os Censos, costuma aplicar procedimentos para a crítica (detecção e identificação de problemas) e imputação (correção e substituição) dos dados, de modo a "eliminar" os erros e inconsistências encontrados, e a "completar" as lacunas das informações. Uma questão recorrente em censos demográficos no mundo é a informação sobre se a mulher já teve ou não filhos. As Nações Unidas (MANUAL X, 1986) sugerem o método de El-Badry para correção desta informação, já que parte dos dados omissos parecem ter características específicas. O método de El-Badry propõe uma imputação sistemática percentual das não-respostas, transformando-as para parturição zero de acordo com a distribuição do número de filhos segundo a idade das mulheres que responderam aos quesitos de fecundidade, isto é, estimar a proporção das mulheres com não-resposta que deveriam ter sido classificadas como sem filhos, e trata porém do agregado e não dos registros individuais.

A nossa principal motivação é estudar e apresentar métodos e modelos como alternativa para o processo de correção dos dados de fecundidade do Censo Demográfico de 1991.

1.2 Objetivo da dissertação

O principal objetivo desta dissertação é propor um método alternativo de imputação dos dados de fecundidade do Censo, e verificar qual o percentual de erro no número de mulheres que foram classificadas como não tendo filhos, usando técnicas de mineração de dados, fazendo uma comparação dos resultados de classificação das redes com a técnica estatística de regressão logística.

Para completar ou substituir os dados ausentes, ou rejeitados por valores considerados aceitáveis do ponto de vista estatístico, o método que será estudado nesse trabalho, procura ser uma nova opção para o tratamento da não-resposta parcial e para a identificação e "correção" de eventuais erros de conteúdo, buscando através das técnicas de Mineração de Dados a existência de um *Padrão* de respostas nos dados para seu uso na imputação dos dados incompletos.

Desta forma, uma das contribuições deste trabalho está na organização de definições relativas ao uso das técnicas de mineração de dados, através de uma revisão bibliográfica, apresentando uma classificação para o tema estudado e a apresentação de um estudo de caso com os dados do Censo Demográfico 1991.

1.3 Barreiras na preparação e utilização dos dados

Entre os fatores que mais tem colaborado com o crescimento da quantidade de informações disponíveis estão os avanços na coleta automática de dados, via sensores remotos e satélites; o processamento automático de transações comerciais, bancárias e científicas; todas apoiadas por instrumentação eletrônica.

Nas empresas, nas organizações governamentais e instituições de pesquisas, dados das mais diversas fontes são gerados, integrados e armazenados, passando a fazer parte de seu acervo, sendo alguns disseminados para a sociedade. Estes dados porém, se encontram espalhados por diversos sistemas de informação, tornando difícil a tarefa de integrá-los, para que possam oferecer subsídios à tomada de decisão. Isso significa que por maior que sejam os avanços das tecnologias de informação, tanto no armazenamento como na manipulação de dados, ainda se verifica uma deficiência na obtenção da informação. Preparar os dados para serem usados requer um cuidado especial, pois mesmo nas pesquisas cuidadosamente planejadas e executadas há sempre os casos de registro com dados incompletos ou totalmente errados.

Diversas técnicas computacionais foram desenvolvidas e estão em estudos de pesquisa científica para tratar e/ou analisar os dados, ou ao menos ajudar o analista a conhecê-los.

1.3.1 Barreiras no uso de Mineração de Dados

O conceito de Mineração de Dados (Data Mining) está se tornando cada vez mais usado como ferramenta de descoberta de informações, que busca revelar estruturas de conhecimento, que possam indicar decisões em condições de certeza limitada, baseado em princípios conceituais de Análise Exploratória e de Modelagem de Dados.

Com o aparecimento da Mineração de Dados, surge a comparação com os métodos tradicionais de análise estatística. Esta comparação, na verdade, apresenta uma certa dificuldade na distinção entre mineração de dados e análise estatística, pois esse novo procedimento de análise é, geralmente, utilizado em conjunto com os métodos

estatísticos, somente sendo adaptados para a análise de enormes bancos de dados, usando a automação de procedimentos e possuindo seus resultados apresentados em interfaces mais amigáveis.

Apesar de mineração de dados e análise estatística terem o mesmo objetivo: construção de modelos que incorporem as dependências entre as descrições de uma determinada situação e seus resultados, ao mesmo tempo representam dois procedimentos diferentes para a análise do dado (DINIZ; NETO, 2000).

Muitos estatísticos se preocupam com a análise primária de dados, isto é, questões são formuladas e traduzidas em forma de hipóteses a serem testadas. Por outro lado, a mineração de dados tem sido considerada e classificada como uma mistura de pesquisas em estatística, inteligência artificial e banco de dados.

Até recentemente a mineração de dados, não era reconhecido como um campo de interesse para os estatísticos, sendo mesmo considerado, nesta área, como uma área de pesquisa "pouco relevante". Devido à sua importância prática, entretanto, o campo tem emergido como uma área de crescimento acentuado e de elevada importância, destacando-se pelo surgimento de diversos congressos científicos e produtos comerciais.

Mas, nem sempre a mineração de dados pôde ajudar aos Sistemas de Apoio à Decisão – SAD, muitas barreiras ainda existem. Podemos citar que as mais importantes são: alto custo das soluções; necessidades de grandes volumes de dados armazenados; pouca amigabilidade das ferramentas de mineração de dados para pessoas que não forem altamente especializadas; dificuldade de preparar os dados para a mineração e as

dificuldades em se obter uma análise custo/benefício bem fundamentada antes do início do projeto. A seguir são detalhadas as principais barreiras:

★ *Altos Custos*

O alto custo da maioria das ferramentas dificulte a disseminação das mesmas nas instituições.

Apesar dos fornecedores dessas ferramentas procurarem introduzir no mercado produtos com custo mais baixo, o alto preço ainda continua limitando o uso em larga escala.

★ *Necessidade de grande volume de dados*

A armazenagem e a administração dos dados sempre foram os maiores obstáculos no uso destas técnicas. A maioria dos fornecedores de software insiste em afirmar que mineração de dados requer terabytes de dados e poderosos servidores, mas já existem soluções mais acessíveis no mercado.

É aceita pelo mercado a afirmação de que 80% do valor de um determinado grupo de dados pode ser encontrado em 20% dos dados que o compõem, por isso o objetivo das ferramentas é tudo fazer para achar esses 20% e minerá-los usando porção de dados em seu próprio computador, o que pode ser uma solução para contornar esses obstáculos.

★ *Pouca amigabilidade*

Mesmo com o aparecimento de novas ferramentas que permitem a mineração em menores grupos de dados, a grande maioria das ferramentas ainda continua incompreensível para os usuários comuns. Isto significa que ainda tem que ser feito por analistas de sistemas, a quem os usuários têm que submeter suas solicitações, esperar enquanto um *expert* processa os dados, para então receberem e examinarem seu resultado sumarizado. Se este não forem satisfatórios, todo o processo tem que ser recomeçado.

★ *Preparação dos dados*

A preparação dos dados envolve aproximadamente 70% do trabalho num projeto de mineração de dados. Os dados devem ser relevantes para as necessidades dos usuários, devem ficar livres de erros e consistentes.

Mesmo que haja um banco de dados com um Data Warehouse[2], onde os dados já devem estar limpos, integrados e organizados, provavelmente centralizados em um único local, continua havendo a necessidade de prepará-los para a escolha dos dados certos a serem usados no processo.

[2] Data Warehouse é um banco de dados para apoio ao processo de tomada de decisão com dados que possuem as seguintes características: Orientado a assunto, Integrado, Variante no tempo, Não volátil (Não permite atualização na sua base de dados, apenas carga e recarga de dados).

* *Análise custo/benefício*

Estimar a taxa de retorno do investimento de um projeto de mineração de dados é complicado devido ao fato que, como o objetivo é descobrir tendências (em dados) que não seriam visíveis de outra maneira, torna-se virtualmente impossível estimar tal taxa a partir de algo que é desconhecido. Visto que normalmente um projeto de Mineração de Dados é razoavelmente caro, pode ser um tanto arriscado se decidir por um projeto desse tipo.

1.3.2 Barreiras na coleta de dados

Mesmo em pesquisas cuidadosamente planejadas, há sempre casos de registros com dados incompletos, inconsistentes ou totalmente errados, já que grande parte das informações é fornecida por informantes baseados na memória individual, e não em documentos legais.

Por este motivo, em qualquer pesquisa é imprescindível fazer alguma crítica dos dados, a fim de impedir que os erros ou omissões afetem negativamente a qualidade dos resultados, ou que tornem muito complexos e caros os processos de elaboração dos resultados.

As pesquisas por amostragem estão sujeitas aos chamados erros amostrais, pela simples razão de se pesquisar apenas uma parte da população. Os demais erros que podem ocorrer são os chamados de erros não amostrais que surgem, geralmente, durante a

execução da pesquisa, tanto numa pesquisa por amostragem como num Censo Demográfico.

Os erros amostrais podem, em geral, ser controlados ou especificados. No caso, a especificação do desenho amostral e a fixação do tamanho da amostra a ser pesquisada são as ferramentas do estatístico para o exercício desse controle sobre os erros amostrais. Já os erros não-amostrais são mais difíceis de controlar, e mesmo sua mensuração pode apresentar grandes dificuldades. No entanto, podem ser mais graves ou maiores que os erros amostrais, dependendo do tipo ou das condições de realização da pesquisa (SILVA, 1989).

Muitos esforços de pesquisa têm sido feitos no sentido de propor métodos estatísticos eficazes para resolver os problemas causados pelos erros alheios à amostragem nas pesquisas. Segundo Kalton & Kasprzyk (1982), os principais erros não-amostrais que podem ocorrer numa pesquisa são causados por :

- *erros de cobertura* – ocorrem quando o sistema de referência (cadastro) da pesquisa possui falhas, omissões ou duplicações de registros;

- *erros de conteúdo ou de resposta* - ocorrem quando há inconsistência ou erros nas informações fornecidas pelos informantes; podem ser também introduzidos durante o processo de apuração dos dados, etc; e

- *erros de não resposta* – ocorrem quando há recusa do informante, presunção do recenseador ou impossibilidade de obter dados (no todo ou apenas parte) referentes a algumas unidades da população incluídas na pesquisa.

Esses erros e suas inconsistências podem afetar de forma considerável os resultados obtidos de uma pesquisa, ou tornar a tarefa de estimação bastante difícil e cara.

1.4 Organização desta dissertação

Esta dissertação está estruturada em sete capítulos. No Capítulo 1 é apresentada uma introdução, a motivação para o desenvolvimento desta dissertação, seus objetivos e as principais barreiras para preparação e utilização dos dados. No Capítulo 2 são apresentados os principais conceitos utilizadas para o desenvolvimento deste trabalho. São introduzidos os principais conceitos dos métodos utilizados para o tratamento de dados ausentes, da metodologia aplicada no Censo Demográfico de 1991 e finalmente os conceitos ligados à área de Descoberta de Conhecimento em Banco de Dados, amplamente conhecida como *Knowledge Discovery in Databases* (KDD). As abordagens da mineração de dados ou metodologias de aplicação são apresentadas no Capítulo 3, também neste capítulo estaremos abordando o uso de um valor padrão para a imputação de dados. No Capítulo 4 é feito um estudo de caso onde é realizada uma comparação entre três métodos apresentados e os resultados obtidos. O Capítulo 5 apresenta as principais conclusões e são feitas algumas recomendações e extensões para eventuais trabalhos futuros a esta dissertação, baseados nas avaliações dos resultados encontrados e nas considerações finais. No Capítulo 6 são apresentadas as referências bibliográficas utilizadas como estudo para realização desta dissertação. Por fim em

anexo, tabela com a descrição de software de mineração de dados e áreas de aplicação, e o questionário usado no Censo Demográfico de 1991.

1.5 Conclusão

Através das metodologias apresentadas e do mapeamento de diversos exemplos constantes no anexo , o analista terá em suas mãos uma ferramenta auxiliar para guiar a decisão com relação à abordagem a ser utilizada em seu projeto de Busca do Conhecimento e contará também com uma opção de método para imputação dos dados de fecundidade utilizando um valor padrão. No próximo capítulo apresentaremos os principais conceitos a serem implementados neste estudo.

2.Conceitos

Neste capítulo são apresentados os principais termos referentes ao universo de Imputação de Dados, da Busca do Conhecimento e da Mineração de Dados, disponibilizando um conjunto de conceitos e definições que serão utilizadas ao longo deste trabalho.

2.1 O tratamento de dados ausentes via imputação de dados

Segundo Albieri (1992,p.138), entende-se por método de imputação todo e qualquer procedimento de tratamento de dados ausentes que atribui um valor específico individual ao dado ausente, seja ele do tipo item perdido ou unidade perdida. Ou seja, a imputação é um processo de estimação de valores individuais em um conjunto de dados.

A primeira idéia que pode surgir para corrigir a não-resposta numa pesquisa igual ao Censo Demográfico seria a volta ao informante na tentativa de se obter o dado completo. Embora isso possa parecer uma solução, tal procedimento apresenta desvantagens, pois implica em atraso da apuração, aumento de custo e, mesmo assim, não garante a obtenção do dado, pois a recusa ou dificuldade de obtenção do dado correto pode persistir.

Segundo Little & Rubin (1987), outras formas de tratamento da não-resposta que podem ser usadas, são elas:

- ✓ ***Procedimento baseado somente nas unidades com informações completas***: Os casos com algum dado ausente são desprezados e a análise passa a se basear nos

casos completos. Este procedimento é de fácil aplicação, sendo que diversos software comerciais dispõem dessa opção já implementada. Este método pode ser satisfatório e fornecer estimativas não viciadas, desde que a ocorrência de dados ausentes ocorra totalmente ao acaso. Nesta hipótese, seu único efeito na estimação de médias é aumentar a variância (imprecisão) devido à redução da amostra. Por outro lado o método pode dar bons resultados, ainda que a ocorrência não seja totalmente aleatória, mas o número de casos desprezados seja pequeno, o que é pouco provável à medida que o número de variáveis pesquisadas aumenta.

- ✓ ***Procedimento de ponderação***: nas pesquisas amostrais, estes procedimentos modificam os pesos das unidades decorrentes do desenho amostral adotado para considerar a não-resposta, por meio de uma estimativa da probabilidade de resposta. Usualmente a probabilidade de resposta é estimada pela proporção de unidades respondentes em uma subclasse da amostra, chamada de classe de ponderação. Os pesos são definidos para as unidades amostrais e, portanto, estes procedimentos são comumente aplicados quando todos os dados de uma unidade estão ausentes, isto é, quando ocorre não resposta total. As informações disponíveis para a formação das classes de ponderação são as variáveis de planejamento da pesquisa.

- ✓ ***Procedimento de imputação***: estes procedimentos se preocupam em substituir os valores ausentes de uma unidade por estimativas dos mesmos. A partir do conjunto de dados completo, isto é, sem erro, gerado com a imputação pode-se utilizar os métodos clássicos de análise de dados tais como análise de regressão, análise fatorial, entre outros. Pode-se entender como *método de imputação*, qualquer procedimento de

tratamento de dados que substitui por um valor específico individual o dado ausente, seja ele do tipo item perdido ou unidade perdida.

- ✓ ***Procedimento de imputação baseado em modelos***: estes procedimentos têm como característica comum o fato de admitirem a hipótese de que os dados ausentes se distribuem segundo um determinado modelo. Isso permite a estimação dos parâmetros do modelo por métodos conhecidos como, por exemplo, o método de máxima verossimilhança, além de permitir uma avaliação da precisão dessas estimativas.

2.2 Métodos para Imputação de Dados

A seguir são apresentados alguns métodos usuais para a imputação de dados ausentes (DIAS; ALBIERI, 1992):

2.2.1 O caso *univariado* – caso em que numa pesquisa, exista um subconjunto de variáveis com ocorrência de dados ausentes e que para as demais variáveis todos os dados estão completos. Neste caso, para cada uma das variáveis com dados ausentes, pode-se aplicar um método de imputação. Outra suposição que se faz é que a ocorrência de não resposta é completamente aleatória.

- ✓ ***Imputação dedutiva***: aplicada quando se é possível deduzir, com grande margem de segurança, o valor de um dado ausente. Por exemplo: se um registro contem parcelas e seu total está ausente, seu valor pode ser deduzido somando-se as parcelas; a ausência de informação para a variável sexo, pode ser solucionada por informações adicionais como número de filhos tidos, informações de serviço militar,

entre outras. Esse tipo de tratamento também é denominado ***imputação determinística***.

- ✓ ***Imputação pela média***: aplicada quando é possível substituir pela média dos valores das variáveis, os valores ausentes desta variável. Uma desvantagem deste método é que acaba diminuindo a variabilidade dos dados e cria um "pico" na distribuição das respostas.

- ✓ ***Imputação da média dentro de classes***: aplicada quando é possível utilizar variáveis auxiliares (com dados completos) para a criação de classes de imputação. O valor ausente de cada classe é imputado pela média dos valores presentes na classe, para a variável que estiver sendo imputada. Este método também apresenta o problema da diminuição da variabilidade pela concentração em torno da média, mas, geralmente, o resultado é melhor que no método de imputação pela média.

- ✓ ***Imputação aleatória***: aplicada quando é possível atribuir para cada valor ausente o valor da mesma variável escolhida de forma aleatória dentre todos os registros de valores presentes. Este método não apresenta distorções de variabilidade, pois produz estimadores não viciados da média e da variância da distribuição da variável que está sendo imputada. O valor imputado é um valor observado da unidade amostral selecionada.

- ✓ ***Imputação aleatória dentro de classes***: aplicada quando é possível criar classes de imputação e os valores ausentes são imputados de forma aleatória a partir de

registros doadores sorteados dentro de cada classe de imputação independentemente. Este método pode apresentar um grande número de variações para a seleção dos doadores.

- Seleção sem reposição – os valores imputados são provenientes de unidades distintas.
- Seleção com reposição – os valores imputados para diferentes unidades podem vir de um mesmo doador.
- Seleção onde as unidades imputadas anteriormente, passam a fazer parte do conjunto de possíveis doadores.

Este método também é chamado de *hot deck* (LITLE; RUBIN, 1987).

✓ ***Imputação "hot deck" seqüencial***: aplicada quando é possível criar classes de imputação e atribuir um valor inicial a ser usado no primeiro valor ausente detectado. A escolha desse valor pode ser realizada de várias maneiras, por exemplo, a partir de pesquisas anteriores. Uma vantagem deste método é que o arquivo de dados só precisa ser lido uma única vez, trazendo economia computacional. Uma desvantagem se dá quando existem unidades da mesma classe, com valores ausentes agrupadas seqüencialmente no arquivo, fazendo com que todos os valores ausentes sejam imputados de um mesmo registro doador, ocasionado uma diminuição da variabilidade dos dados.

✓ ***Imputação por função distância***: aplicada quando o doador é escolhido na unidade mais próxima à unidade a ser imputada. Esta proximidade é medida através de uma função de distância definida sobre um conjunto de variáveis auxiliares cujos valores estão presentes para todas as unidades.

- ✓ ***Imputação por regressão***: aplicada quando, a partir de um sub-conjunto de unidades com dados completos, se ajusta uma regressão linear da variável a ser imputada como função de variáveis auxiliares. Em seguida, a função de regressão é aplicada a cada unidade com valor ausente para estimar o valor a ser imputado.

2.2.2 O caso *multivariado* - no caso multivariado, diferentemente do univariado, é que para cada unidade, as imputações das variáveis são feitas em um mesmo momento e não independentemente para cada variável, deste modo consegue-se manter a consistência dos dados sem gerar novos erros.

- ✓ ***Imputação baseada no algoritmo EM*** [3] ***:*** aplicada quando se supõe que os dados tem distribuição normal multivariada com vetor de médias μ e matriz de covariâncias Σ. Para a aplicação desse método deve-se considerar a retirada dos outliers presentes no conjunto de dados. A idéia básica do modelo pode ser descrita como:
  - ♦ Estimar os parâmetros da distribuição normal multivariada, como se os dados fossem completos;
  - ♦ Reestimar os valores ausentes com base nos parâmetros obtidos nos resultados do item anterior; e
  - ♦ Repetir os passos anteriores até a convergência do método.

As estimativas dos dados ausentes são obtidas através de uma regressão linear multivariada das variáveis com dados ausentes como função de variáveis auxiliares com dados completos.

[3] Algoritmo EM – Foi denominado "Algoritmo EM" por Dempster, Laird & Rubin (1977)

- ✓ ***Imputação baseada na metodologia de Fellegi & Holt***: aplicada quando só é necessário especificar os códigos válidos para um conjunto restrito de variáveis, além das regras de incompatibilidade entre respostas dentro da mesma unidade. É usado basicamente para o estudo de dados qualitativos. Os princípios fundamentais da metodologia, segundo os autores são (FELLIGI; HOLT, 1976):
  - ♦ Alterar, em cada unidade, o menor número possível de respostas, de tal maneira que todas as regras de crítica sejam respeitadas;
  - ♦ As regras de crítica devem gerar automaticamente as regras de imputação; e
  - ♦ Imputar os valores ausentes ou inconsistentes preservando ao máximo as distribuições de freqüência dos dados bons observados.

Esta seção teve como objetivo apresentar os principais conceitos e métodos para o tratamento de dados ausentes, mas ressalta-se que na prática, em muitos casos o tratamento da questão da imputação é ignorado. A presença de respostas como "sem declaração" ou "ignorado" em tabelas de divulgação e resultados de pesquisas é uma prova da não utilização de métodos de imputação.

Outro problema a ser observado é que, em algumas pesquisas, não se consegue diferenciar o valor de resposta zero, para o caso da falta de informação, fatos que do ponto de vista da análise final da informação têm significado bem distinto.

2.3 Busca de Conhecimento em Banco de Dados - (*KDD – Knowledge Discovery in Database)*

A automatização dos processos de análise de dados, com a utilização de software ligados diretamente a massa de informações tornou-se uma necessidade, já que conhecer profundamente o dado, tem o potencial de transformá-lo em conhecimento.

Utilizando o processo de extração em um banco de dados para encontrar padrões escondidos sem uma idéia ou hipótese pré-determinada sobre o que são esses padrões, nos permite detectar os dados que seguem uma norma ou tendências, nos trazendo uma riqueza de padrões que podem ser expressados e descobertos determinando a força e a utilidade da técnica de descoberta. Isto permite ao usuário aplicar os padrões encontrados para estimar valores nos novos itens ou nos itens incompletos.

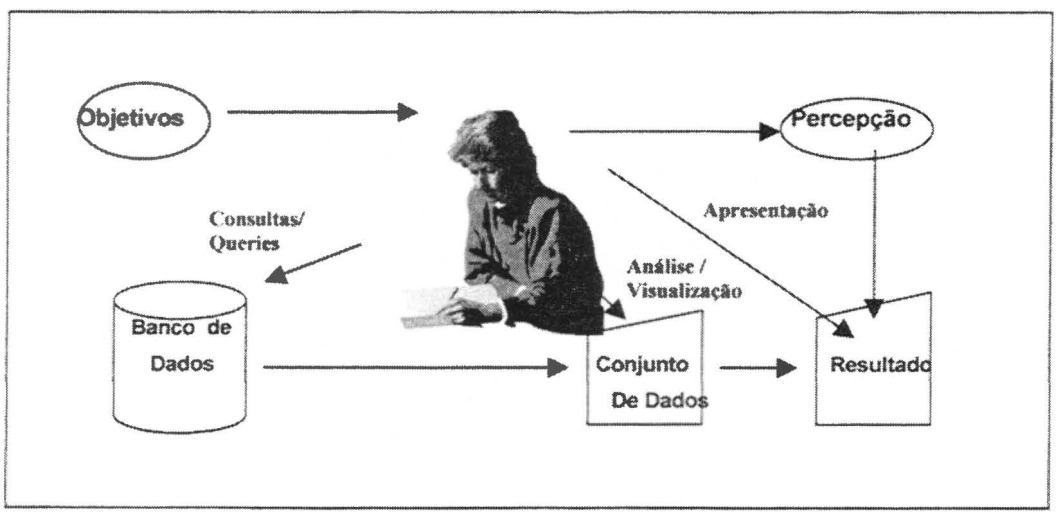


Figura 3: Tarefas em um Processo de KDD

O processo de Busca de Conhecimento é geralmente feito em etapas, que envolvem a preparação dos dados, a procura de padrões, o teste do conhecimento e o refino do modelo. O processo é caracterizado por não ser trivial, ou seja, por possuir um grau de

autonomia na procura pelo conhecimento. O processo de KDD é iterativo, envolvendo inúmeras tarefas com muitas decisões tomadas pelo usuário. A figura acima, mostra essas tarefas que um usuário de KDD precisa enfrentar, descrita por Brachmam & Anand (1996, p. 37-57).

O analista envolvido em um processo de busca de conhecimento, como podemos acompanhar na Figura 3, em resposta a um determinado objetivo, extrai de um banco de dados, através de uma consulta, um conjunto de dados para sua análise. Após a geração desse conjunto de dados, ferramentas de análise e visualização são utilizadas.

Essas análises levam ao analista algumas informações preliminares sobre as questões relacionadas com o objetivo. A figura abaixo mostra todos os passos que devem ser executados para identificar um processo KDD.

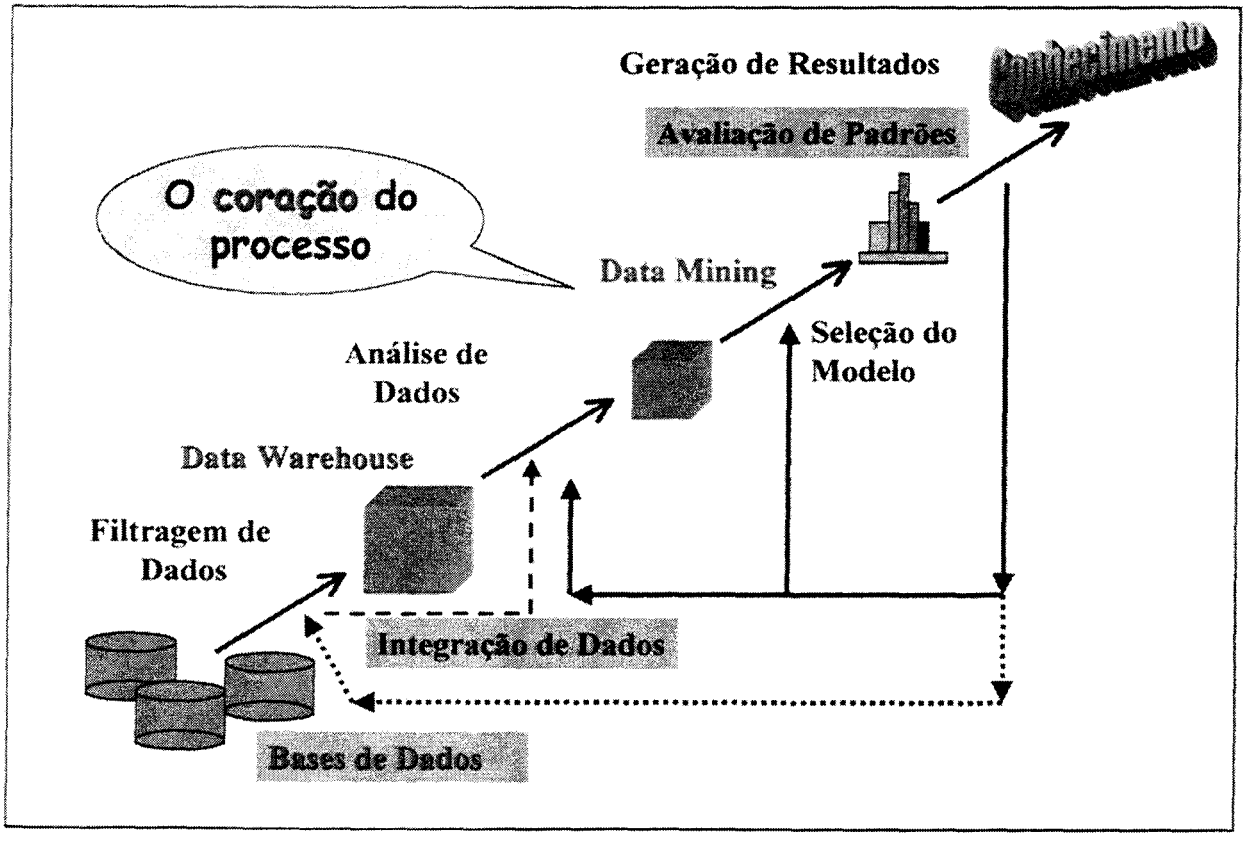


Figura 4: Etapas do processo de KDD

Filtragem dos Dados

O objetivo desta etapa é identificar a origem das informações e selecionar as variáveis de interesse para a aplicação das técnicas. As variáveis selecionadas podem ser do tipo categórica ou quantitativa.

O método mais comum utilizado para verificar a consistência de um conjunto de dados é selecionar o mesmo dado de fontes múltiplas e comparar seus resultados. No entanto, é uma tarefa que exige um conhecimento muito grande do analista, que deve discernir entre um dado realmente incorreto e uma exceção (*outlier)* à regularidade do conjunto de dados que deve ser incluída na amostra. Quando temos um grande volume de dados selecionados, podemos considerar apenas uma amostra representativa dos mesmos no processo de revisão.

As variáveis categóricas assumem valores num conjunto finito, podem ser nominais ou ordinais. Exemplos de variáveis nominais são estado civil, sexo. Exemplos de variáveis ordinais são graus de instrução, escore de crédito pessoal. Ao contrário da variável nominal, existe uma ordem entre os valores de uma variável categórica ordinal.

As variáveis quantitativas assumem valores numéricos e podem ser contínuas ou discretas. Exemplo de variáveis contínuas são salário e renda. Exemplos de variáveis discretas são número de filhos, número de empregados.

Uma maneira simples de entender o conteúdo dos dados, no caso de variáveis categóricas, é através da distribuição de freqüência dos valores e/ou através de ferramentas gráficas tais como histogramas e diagramas de setores (DINIZ; NETO, 2000).

No caso de variáveis quantitativas, uma forma de determinar a presença de valores inválidos é através do cálculo de medidas estatísticas tais como o valor mínimo e máximo, média aritmética, moda, mediana e desvio padrão amostral.

Analisando os valores mínimos e máximo pode-se detectar rapidamente valores suspeitos nos dados e através das diferentes medidas de tendência central e de dispersão, pode-se verificar o grau de ruído nos dados. O diagrama de caixas (boxplot) pode ser utilizado para comparação da distribuição da média ou do desvio de duas ou mais variáveis e o diagrama de dispersão é um gráfico simples bi-dimensional que representa a relação entre duas variáveis contínuas.

Valores que estão fora do esperado para uma ou mais variáveis podem ser devido a erros ou *outliers*. Os *outliers* podem identificar uma nova tendência de resultados para as variáveis em questão ou podem ser também dados inválidos. Nesses dois casos, os diferentes tipos de *outliers* devem ser tratados de forma diferente. Um tipo comum de erro, pode ser gerado por uma falha humana, como o registro de uma pessoa com idade superior a 200 anos.

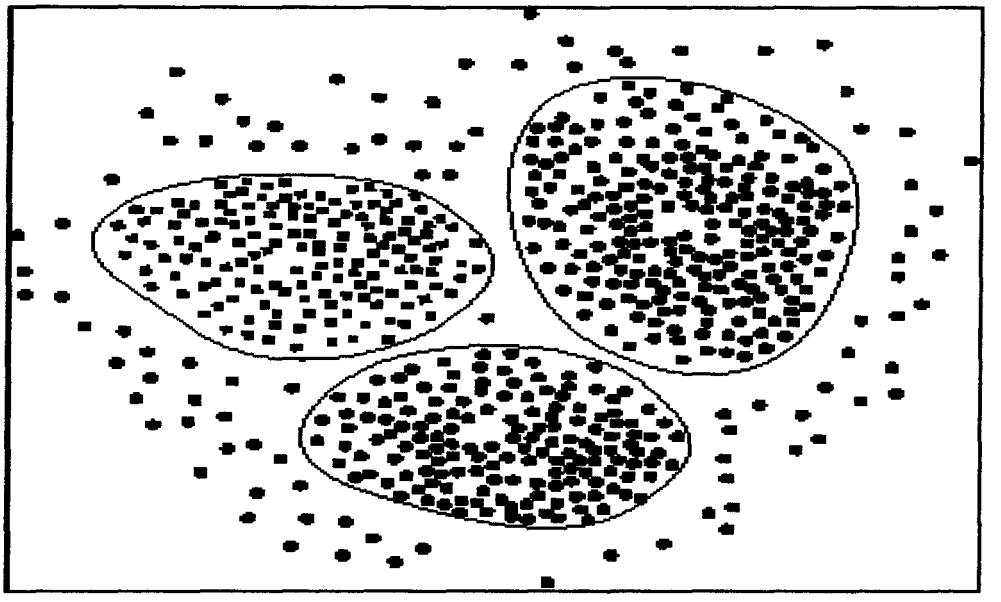


Figura 5: Detecção de *outliers* utilizando uma abordagem visual

Outro problema encontrado também nos dados, diz respeito aos valores *missing* , que são os dados que não estão presentes no conjunto selecionado e os valores inválidos que foram eliminados durante a limpeza dos dados. Os valores *missing* podem ocorrer por falha humana ou porque a informação não está disponível no momento do levantamento desses dados.

Uma forma de tratamento de dados *missing* é simplesmente eliminar todo o registro que contenha esses valores faltantes. Este procedimento seria simples, mas acarreta a perda de informações. Embora esta perda possa não ser um problema quando temos um grande volume de dados, mas com certeza terá algum efeito nos casos onde o volume de dados for pequeno ou onde os objetivos se direcionem para o controle de qualidade ou fraudes.

Portanto, a decisão de eliminar observações ou variáveis não é fácil e não se pode predizer as conseqüências de tais atitudes. Felizmente existem técnicas que permitem substituir os dados faltantes. Para variáveis quantitativas, a mais simples é o uso da média ou da moda amostral. Para as variáveis categóricas, pode-se utilizar a moda ou um novo atributo criado para a variável, como por exemplo, usar "99" para os valores desconhecidos. Outras técnicas mais avançadas como modelos de predição e técnicas de imputação também podem ser usadas para os dois modelos de variáveis (HAIR et al, 1998).

Um fator que temos que ter sempre em mente, é de não desprezar nenhum dado, e se necessário for, usar a transformação de dados para converter o conjunto de dados a ser utilizado para uma forma padrão de uso. Técnicas como discretização[4] , 1 a N [5] e técnicas de redução de dimensionalidade[6] são comumente usadas.

Uma técnica que pode ser aplicada para obtenção de uma representação reduzida (*compactada*) de um conjunto de dados é a construção de cubos de dados [7] . A figura 6 a seguir, extraída de (HAN; KAMBER, 2001), mostra a transformação de dados relacionais em multidimensionais.

[4] Discretização, é a técnica de converter variáveis contínuas em categóricas.
[5] 1 a N é usada para converter variáveis categóricas em uma representação numérica.
[6] Redução de dimensionalidade é usada para combinar várias variáveis em uma única variável.
[7] Cubo de dados – Estrutura multidimensional para análise de dados.

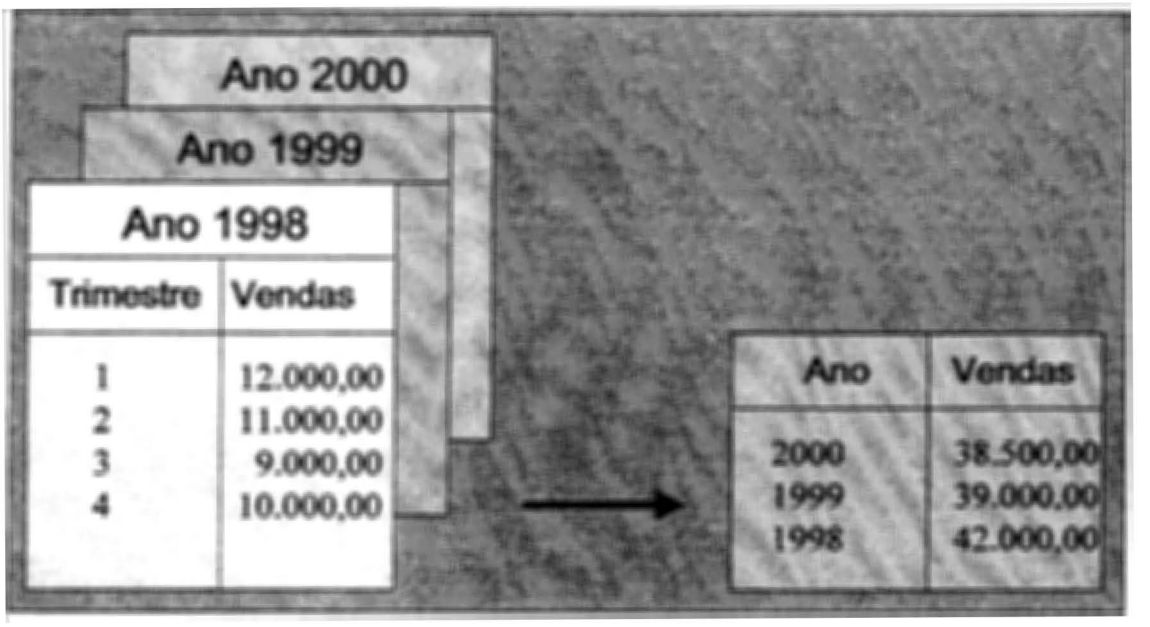


Figura 6: Agregação de dados em forma multidimensional

A figura 7 abaixo, mostra uma forma de visualização e interpretação dos dados no modelo multidimensional.

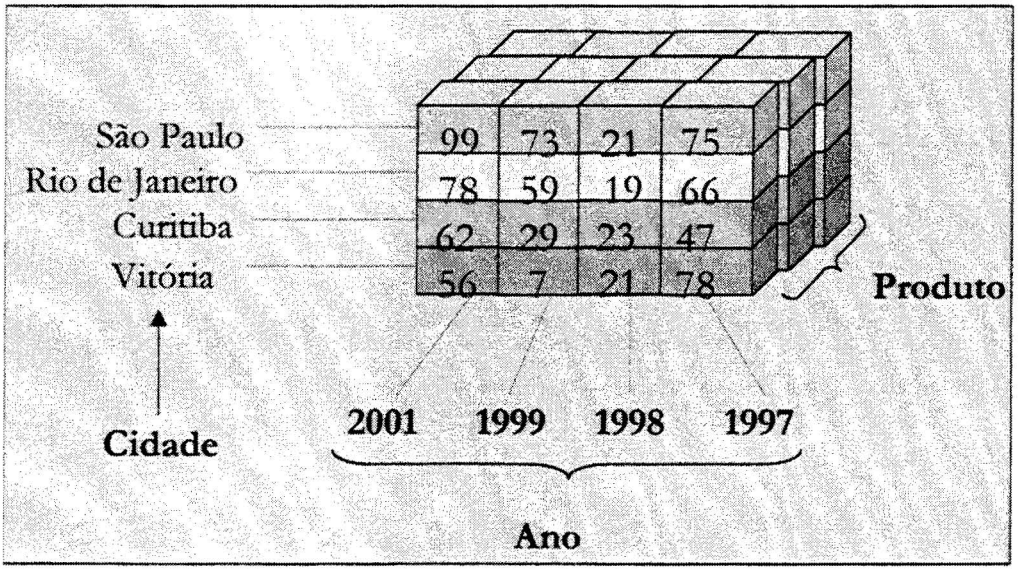


Figura 7 : Representação de dados no modelo multidimensional

Existem vários tipos de processos de limpeza que podem ser aplicados inicialmente, outros, no entanto, podem ser aplicados somente após a detecção de algum tipo de problema nas etapas subseqüentes do KDD (ADRIAANS; ZANTINGE, 1996).

Criando um Data Warehouse

Todos os dias, dados das mais diversas fontes dentro de uma instituição são gerados, armazenados e disseminados, passando a fazer parte do seu acervo. Estes dados, porém, se encontram espalhados por diversos sistemas de informação, tornando difícil a tarefa de integrá-los. Os bancos de dados que armazenam as operações diárias foram feitos para responder a questões simples, como totalizações, somatórios e revelam enorme dificuldade para responder a pesquisas que necessitam relacionar dados de diversos armazenamentos.

A tecnologia de *Data Warehouse* veio resolver essa dificuldade. Um sistema de data warehouse é composto, entre outras ferramentas, de um banco de dados para onde somente dados necessários para a tomada de decisão são carregados, vindos dos bancos de dados operacionais. Assim, as consultas e pesquisas feitas sobre ele são mais rápidas e podem responder a questões complexas. William H. Inmon marcou o termo *Data Warehouse* em 1990. Segundo Inmon (1997,p.388), um *Data Warehouse* é um depósito integrado de informações coletadas de fontes heterogêneas de dados operacionais e reunidos em um banco de dados, centralizando informações e possibilitando que elas sejam compartilhadas.

O processo de *Data Warehousing* é iterativo, ou seja, está em constante evolução. Segundo Inmon, "o *Data Warehouse* é construído e implementado de uma maneira

evolucionária, passo a passo, organizando e armazenando os dados necessários para a análise informacional e o processamento analítico sob uma perspectiva de longo prazo"

Entre os benefícios do *Data Warehouse*, podem-se destacar o monitoramento e a comparação de operações atuais com as passadas, bem como prever situações futuras. De acordo com W. H. Inmon, arquiteto principal na construção de sistemas de armazenamento de dados, *Data Warehouse* é um banco de dados para apoio ao processo de tomada de decisão com dados que possuem as seguintes características:

- Orientado a assunto (tópico)
- Integrado
- Variante no tempo
- Não volátil

Abaixo são descritas as características de um *Data Warehouse*:

Orientado a assunto

O banco de dados é orientado a assunto quando os dados deixam de ser orientados para as aplicações do dia a dia que são projetados para apoiar procedimentos operacionais, para terem como alvo uma visão simples e concisa sobre determinado assunto, excluindo dados que não são úteis ao processo de apoio a decisão. Por exemplo, os dados orientados para aplicação contém uso de um determinado produto e seus consumidores. Por outro lado, dados para apoio a decisão contém histórico de uso em um determinado período de tempo.

A característica marcante de um Data Warehouse é que desde sua concepção até o projeto de sua estrutura, o objetivo é focar o apoio no gerenciamento do processo de tomada de decisão. Deste modo, todos os envolvidos nesse processo são considerados e dados não relevantes são eliminados do Banco de Dados.

Integrado

Um dos mais importantes aspectos de um Data Warehouse é sua integração entre diferentes bancos de dados e plataformas, garantindo a consistência na informação.

Esta consistência significa padronização nos nomes ou convenções dadas, nas medidas das variáveis, na codificações de estruturas, etc., ou seja, um dado que poderia estar repetido em várias tabelas em diferentes bancos de dados, são unificados no Data Warehouse. Por exemplo, suponhamos que uma informação relevante para o estudo seja identificar o sexo do informante.

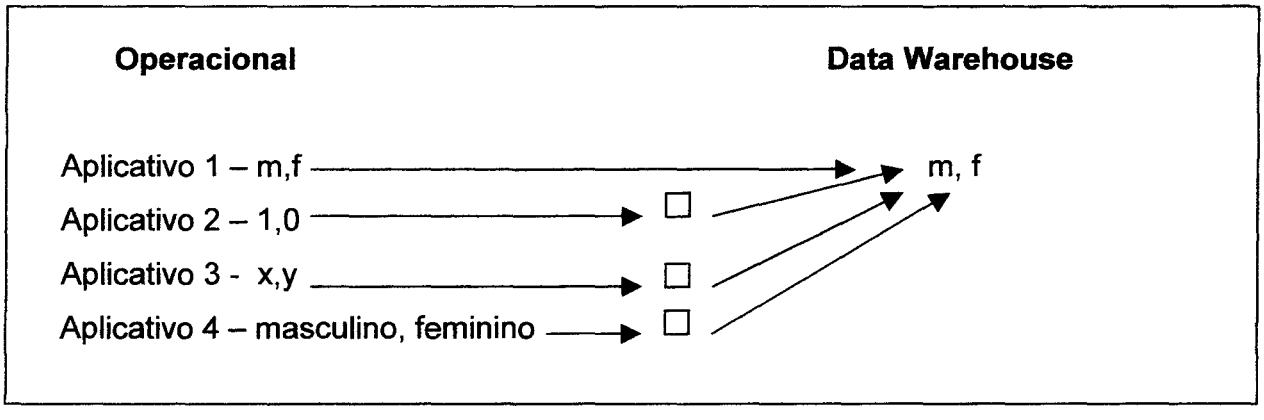


Figura 1: Data Warehouse: Integração de dados (INMON, 1997).

Em quatro aplicativos diferentes, como mostra a Figura 1, um mesmo atributo, no caso o sexo do informante, é representado de quatro formas diferentes. Num Data Warehouse,

quando esses quatro aplicativos forem reunidos, os dados serão integrados antes de sua migração.

Deste modo, a utilização de um Data Warehouse gera como primeiro resultado a *padronização* de estrutura de dados e a *eliminação de inconsistências* nos conjuntos de dados.

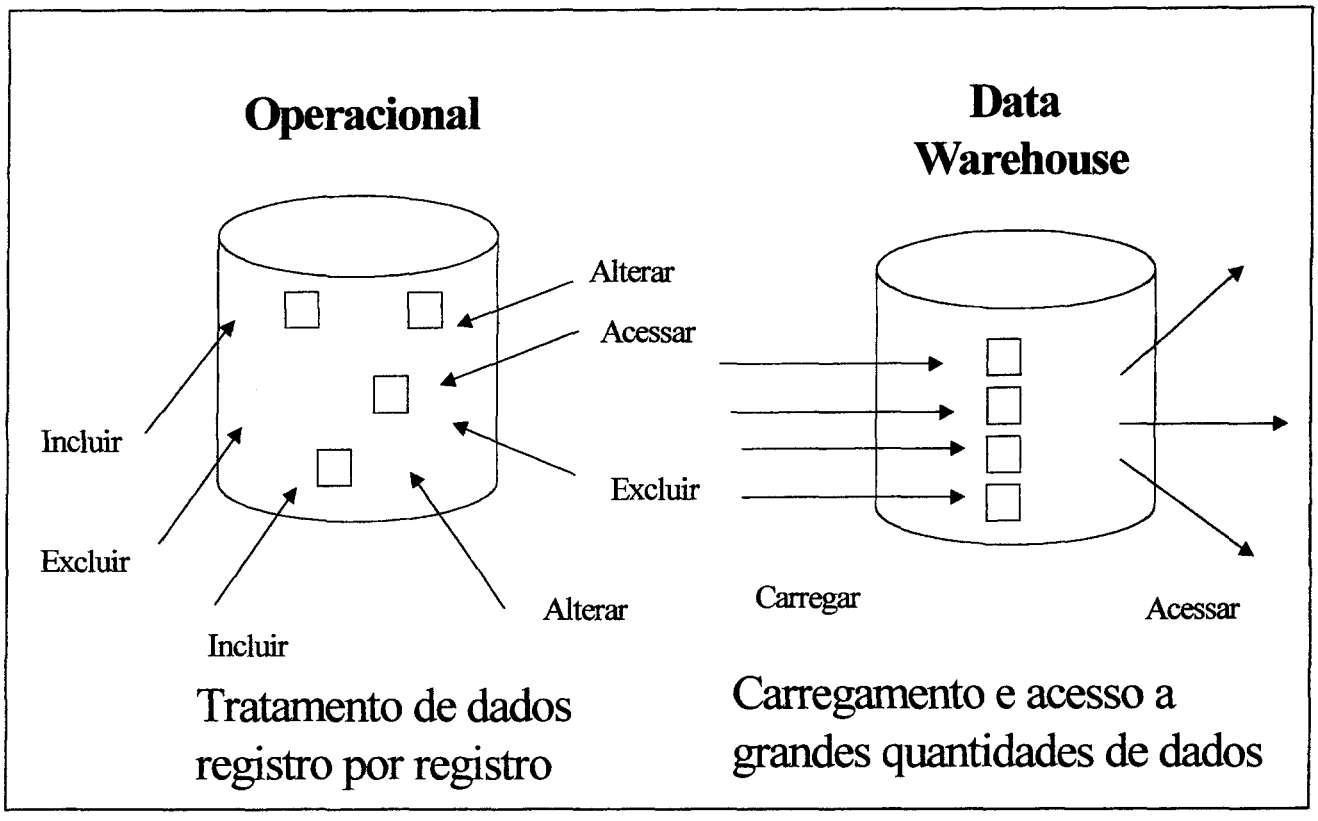
Variante no Tempo

Os dados operacionais são atualizados em tempo real, ou seja, a cada ocorrência registrada, os bancos de dados podem ser modificados. No entanto, isso não ocorre num Data Warehouse. Por possuir registros com horizontes de tempo muito maiores que os dados operacionais, um Data Warehouse não possui um acesso em tempo real para atualização. O acesso é feito, uma "foto" do momento é gerada e em seguida as consultas a esse banco de dados são feitas. Se novos dados forem introduzidos, é necessária a geração de uma nova "foto" de tempo para que novas consultas possam ser feitas, isto é, será executada uma nova carga de dados.

Não Volátil

Os dados operacionais são atualizados regularmente, registro por registro nos bancos de dados operacionais. Como podemos observar na Figura 2 a seguir, novos dados são inseridos, dados antigos são modificados e alguns substituídos. Num *Data Warehouse* apenas duas etapas ocorrem: uma de carga da base de dados e outra de acesso a essa informação criada.

Figura 2: A questão da não volatividade



Geralmente os dados estão espalhados em vários bancos de dados operacionais dentro das organizações, por exemplo, dados de Recursos Humanos, de produção, de vendas, e outros. Um *Data Warehouse* deve ser desenvolvido de maneira que os dados possam ser extraídos desses 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 processamentos analíticos e tomadas de decisões.

Análise de dados

O analista em geral possui uma hipótese sobre o conjunto de dados e algum tipo de ferramenta de análise é utilizada para a construção de um modelo que possa representar

a distribuição desses dados. Em geral, a idéia é entender como certos grupos de entidades se comportam. Os processos principais na Análise de Dados são:

- *Especificação do modelo*: onde um modelo específico é escrito de uma maneira formal;
- *Ajuste do modelo*: onde, quando necessário, alguns parâmetros específicos do modelo são determinados de acordo com o conjunto de dados;
- *Avaliação do modelo*: onde o modelo é avaliado com um conjunto de dados, através de um conjunto de teste, onde os valores de entrada e saída são conhecidos previamente; e
- *Refino do modelo*: onde o modelo inicial é iterativamente alterado até que algum parâmetro de erro seja alcançado.

Em ferramentas de análise baseadas em algoritmos, um modelo é especificado através da associação de variáveis de entrada (*input*) e variáveis de saída (alvos ou *target*). Em certos casos, o conjunto de dados é subdividido em um conjunto de dados para treinamento e um conjunto de dados para teste. O conjunto de dados de treinamento é utilizado para ajustar os parâmetros do modelo. O modelo é então avaliado através de sua aplicação no conjunto de teste.

Em ferramentas de análises baseadas em visualização, a hipótese é especificada através do conjunto de dados e da seleção de elementos nos dados. O resultado produzido é o modelo e seu poder explicativo pode ser observado através de sua apresentação. A visualização é um ingrediente chave para as 3 principais tarefas no processo de KDD: desenvolvimento do modelo, análise de dados e geração de resultados. A representação

apropriada dos dados e suas relações podem dar ao analista informações preliminares que não são possíveis de serem obtidas através de tabelas de resumos estatísticos. Em alguns casos, a visualização dos dados é a única solução para problemas de confirmação de hipóteses (BRACHMAN; ANAND, 1996).

Desenvolvimento do Modelo

Raramente um processo de descoberta de informações sobre uma determinada população inicia-se com a hipótese já definida. Uma das operações principais é descobrir subconjuntos da população que se comportem de forma semelhante no foco da análise. Em muitos casos, a população inteira pode ser muito diversa para compreensão, mas detalhes dos subconjuntos podem ser trabalhados.

A interação com o conjunto de dados leva à formulação das hipóteses. Nessa fase, os três principais sub-processos são: (1) segmentação dos dados, (2) seleção do modelo e (3) seleção de parâmetros. Para a segmentação dos dados, podem ser utilizadas ferramentas de *clustering* (grupamento). Para a seleção do modelo, uma grande variedade de modelos de análises podem ser utilizados, como regressão, árvore de decisão, redes neurais e regras de associação (BRACHMAN; ANAND, 1996). O analista deve escolher o melhor tipo de modelo antes de iniciar a utilização de uma ferramenta específica. As fases de análise e desenvolvimento do modelo são complementares e o analista pode voltar e alterar cada fase iterativamente. Esse ciclo é muito importante no processo de descoberta.

Geração de Resultados

Num cenário simplista, uma análise resulta em um relatório que pode incluir medidas estatísticas do modelo, dados sobre exceções, etc. Em geral, os resultados devem ser gerados de forma variada e simplificada. Uma descrição textual de uma tendência ou um gráfico que capture as relações no modelo são mais apropriados. Ações também são indicadas, ou seja, gerar e detalhar procedimentos que o usuário deva tomar dependendo de certas características.

O resultado de um processo de KDD deve ser visto como uma especificação de uma aplicação a ser construída que responda às questões chave sobre o informante. Isso usualmente vem na forma de um modelo que será utilizado como uma especificação principal para uma aplicação.

Mineração de Dados é parte do processo de Busca de Conhecimento, o qual possui uma metodologia própria para preparação e exploração dos dados, interpretação e assimilação de seus resultados. No entanto, tornou-se mais conhecida do que o próprio processo de KDD em função de ser a etapa onde são aplicadas as técnicas de busca de conhecimento.

Atualmente, existem poucos produtos de Mineração de Dados no mercado. Esses produtos certamente são muito úteis na hora de extrair informações de grandes bancos

de dados e conjuntamente com outras técnicas e ferramentas, formam o chamado *Business Intelligence*[8] ou Inteligência de Negócios.

A figura 8 abaixo, mostra as tecnologias utilizadas no contexto da inteligência de negócios.

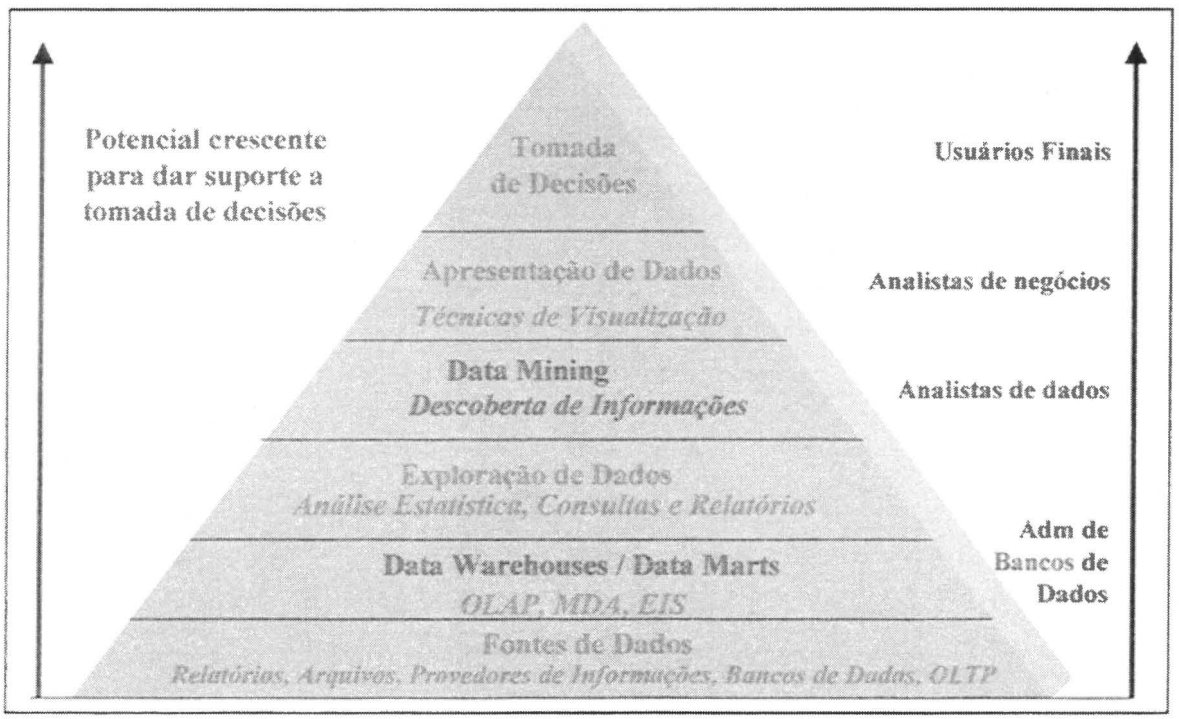


Figura 8: Mineração de Dados no contexto da Inteligência de negócios

Diversas tecnologias tem sido usadas conjuntamente em inteligência de negócios, entre elas se destacam as tecnologias de On-Line *Analitical Processing* (OLAP), de Análise e Exploração de Dados (AED), de *Data Warehousing* (DW) e Mineração de Dados (MD) ou *Data Mining*.

[8] Conjunto de conceitos e metodologias que, fazendo uso de acontecimentos (fatos) e sistemas baseados nos mesmos,

As aplicações de OLAP permitem que o usuário final possa extrair informações e conhecimento que são úteis para sua pesquisa visualizando e manipulando de forma multidimensional os dados do Data Warehouse. A abordagem de análise utilizada neste caso é a de verificação ou descoberta dedutiva, isto é, uma consulta é solicitada e um resultado é apresentado para verificação, em forma de tabelas cruzadas ou gráficos. Nesta abordagem, o usuário final formula uma hipótese sobre o seu negócio e então utiliza a ferramenta de OLAP para aprovar ou desaprovar esta hipótese.

Entretanto, ao utilizar a abordagem de verificação, existem algumas informações importantes que podem passar desapercebidas para o usuário, até pelo número de variáveis e relacionamentos de parâmetros, os quais são impossíveis de serem processados pelo cérebro humano. Para evitar esta perda de informações é que se torna necessário o processo de Mineração de Dados . Este processo trabalha no modo de descoberta indutiva, ou seja, os dados são analisados através de um conjunto de algoritmos e critérios especificados. Os resultados são apresentados como conclusões ou regras (conhecimento), ao invés de tabelas que o usuário deve ainda analisar para chegar a estas mesmas conclusões.

Geralmente, as fontes de dados em grandes organizações estão pulverizados entre vários bancos de dados e até em vários ambientes. A integração desses dados em um único meio de armazenagem, se faz necessário, e na maioria dos casos é a fase mais demorada de todo o processo.

---

apóia a tomada de decisões [BL97].

2.4 Mineração de Dados (*Data Mining*)

Mineração de Dados 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 e predição. Os bancos de dados são geralmente volumosos, chegando atualmente a terabytes e na forma que se encontram nenhum uso direto pode ser feito deles; as informações escondidas nos dados é que são realmente úteis.

Diversos pesquisadores estão empenhados em desenvolver novas técnicas de análise de dados, especialmente projetadas para tratar questões relativas a mineração de dados. No entanto, a mineração de dados ainda está baseada em princípios conceituais de Análise Exploratórios de Dados (*Exploratory Data Analysis* – EDA) e de modelagem de dados

Diversas definições de Mineração de Dados podem ser encontradas na literatura, destacamos as seguintes:

- Mineração de dados é um modo de procurar relações escondidas em um grande conjunto de dados. Neste caso, podemos definir qualquer estrutura necessária para investigação, como padrões de clustering, aproximações de funções e as técnicas de Data Mining podem ser a princípio semelhantes às análises de regressão (KING,1999);
- Mineração de dados é a busca de informações valiosas em grandes bancos de dados. É um esforço de cooperação entre homens e computadores. Os homens projetam bancos de dados, descrevem problemas e definem seus objetivos. Os computadores verificam dados e procuram padrões que casem com as metas estabelecidas pelos homens (WEISS; INDURKYA, 1997);

- Mineração de dados é a descoberta de padrões, associações, mudanças, anomalias e estruturas estatísticas e eventos em dados (GROTH, 2000);

- Mineração de dados consiste em uma seqüência iterativa dos seguintes passos: Limpeza, Integração, Seleção, Transformação e Mineração dos dados, gerando Padrões e assimilando conhecimentos (HAN; KAMBER, 2001); e
- Mineração de dados é um processo altamente cooperativo entre homens e máquinas, que visa a exploração de grandes bancos de dados, com o objetivo de extrair conhecimentos através do reconhecimento de padrões e relacionamentos entre variáveis, conhecimentos esses que possam ser obtidos por técnicas comprovadamente confiáveis e validados pela sua expressiva estatística (CORTES; PORCARO; LIFSCHTZ, 2002).

Um conceito muito difundido e equivocado sobre mineração de dados é o que define os sistemas de mineração de dados como sistemas que podem *automaticamente* minerar todos os conceitos que estão escondidos em um grande banco de dados sem intervenção ou direcionamento humano (HAN; KAMBER, 2001).

A aplicação típica de Mineração de dados começa com um grande conjunto de dados e poucas definições. A maioria dos algoritmos tratam os dados iniciais como verdadeiras "caixas – pretas", com nenhuma informação disponível sobre o que os dados descrevem, quais relações existem entre os dados e se contém erros. Ao examinar os dados, um algoritmo pode explorar milhares de prováveis regras, utilizando diversas técnicas para escolher entre elas.

Certa dificuldade na distinção entre mineração de dados e análise estatística deve-se basicamente as semelhanças entre elas e o 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 mineração forem diferenciadas, ou pelo menos entendidas como uma adaptação das técnicas estatísticas tradicionais, visando a análise deste enorme volume de dados.

Apesar de mineração de dados 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, elas representam dois procedimentos diferentes para análise de dados (DINIZ; NETO, 2000).

A análise de dados tradicional é baseada na suposição, em que uma hipótese é formada e validada através dos dados. Por outro lado, as técnicas de mineração de dados são baseadas na descoberta na medida que os padrões são extraídos do conjunto de dados.

Mineração de Dados ou Data Mining pode ser visto como aplicação direta 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.

O processo de aplicação da mineração de dados envolve vários estágios antes de se iniciar a busca do conhecimento e definir claramente a que resultados deseja-se chegar, isto é, que tipo de problema se deseja resolver. Uma vez definidos os problemas, será

preciso escolher que técnicas utilizar e o método de abordagem (ferramentas) a ser usado para aplicar essas técnicas e obter os conhecimentos desejados.

2.4.1 Definição do Problema

Para determinação dos tipos de problemas envolvidos no universo da Mineração de dados, utilizamos os trabalhos de autores como CARVALHO (2001) e BERRY & LINOFF (1997). Isso foi feito devido ao tipo de abordagem, voltado para profissionais de várias áreas.

Foram determinados sete tipos gerais de problemas, conforme apresentado abaixo:

- Classificação;
- Estimação;
- Previsão;
- Afinidade de Grupos;
- Segmentação;
- Descrição; e
- Reconstrução de funções.

Classificação

É o tipo de problema mais comum em mineração de dados, simplesmente porque é uma das tarefas cognitivas humanas mais realizadas no auxílio à compreensão do ambiente em que vivemos, O ser humano está sempre classificando o que percebe à sua volta: criando classes de relações humanas diferentes (colegas de trabalho, amigos, familiares, etc.) e dando a cada classe uma forma de tratamento diferenciado. A classificação

consiste em examinar as características de um objeto e associar essas características a classes pré determinadas. Pode ser do tipo simples ou múltipla.

A classificação simples consiste na identificação de uma característica binária. Ou seja a determinação da existência ou não de uma determinada característica. A classificação múltipla consiste em identificar a classe de um determinado objeto. Dessa forma, esse tipo de problema é caracterizado por uma definição detalhada das classes, possuindo um conjunto de dados para treinamento com exemplos pré classificados. O objetivo é a construção de um modelo que seja capaz de gerar classificações de novos objetos ou novos dados, sendo desenvolvido a partir de dados históricos classificando as respostas em positivas ou negativas. Em seguida, esse modelo pode ser aplicado em nova massa de dados para classificá-la, isto é, o modelo deve ser capaz de predizer corretamente situações já conhecidas antes de iniciar a predição de novas situações.

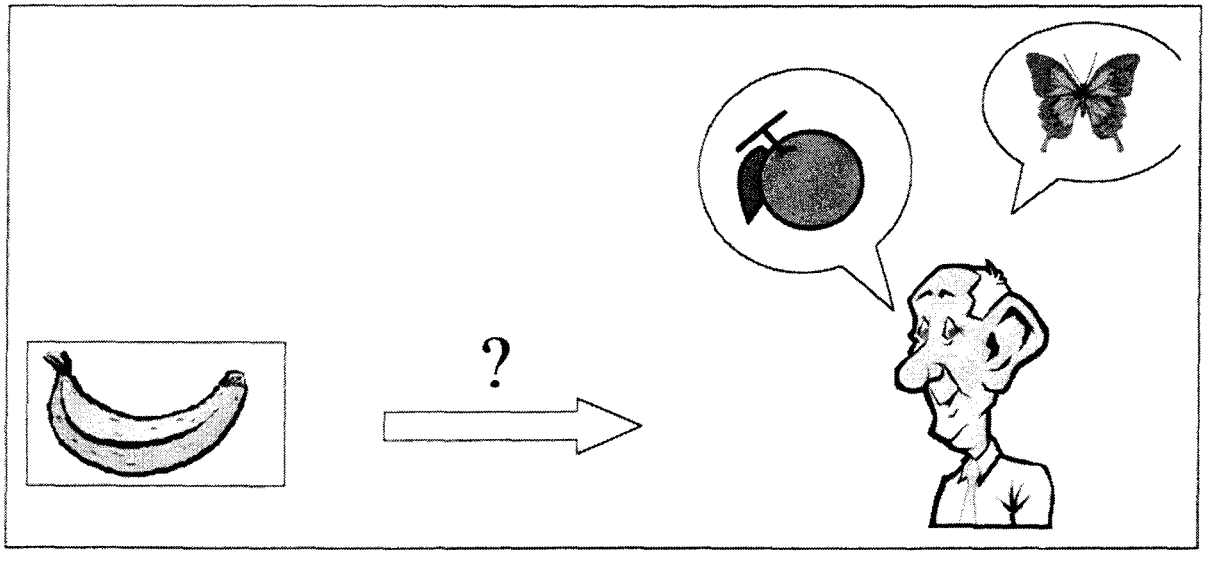


Figura 9: Classificação

Podemos citar exemplos desse tipo de problema:

- Definição de palavras-chave em artigos para publicação acadêmica;
- Análise de risco de crédito para empréstimo;
- Campanhas promocionais em mala direta;
- Detecção de fraudes; e
- Previsão de falências.

Estimação

Estimar algum índice é determinar seu valor mais provável diante de dados do passado ou de dados de outros índices semelhantes sobre os quais se tem conhecimento. Enquanto o problema de classificação lida com valores categóricos (sim ou não, classes, etc.), o problema de estimação consiste na determinação de valores contínuos (renda, altura, etc.).

Na prática, a estimação é utilizada para conduzir uma tarefa de classificação. Uma loja de roupas , por exemplo, interessada em anunciar uma promoção de ternos deseja construir um modelo que classifique seus clientes em dois conjuntos: usuários de terno e não usuários. Uma alternativa a essa abordagem é a construção de um modelo que associe a cada cliente um valor (entre 0 e 1) para a probabilidade do uso de terno. Dessa forma, estaríamos estimando a probabilidade do cliente usar terno.

Outro exemplo seria supor que se deseje determinar o gasto de famílias cariocas com lazer, em função da faixa etária e padrão sócio-cultural, tomando como base os dados das famílias paulistanas. Certamente que esta estimativa pode nos levar a erros, uma vez

que Rio de Janeiro e São Paulo são cidades com geografias diferentes e oferecem diferentes opções de lazer .

A arte de estimar é exatamente esta: determinar da melhor forma possível um valor baseando-se em outros valores de situações semelhantes, mas nunca exatamente iguais.

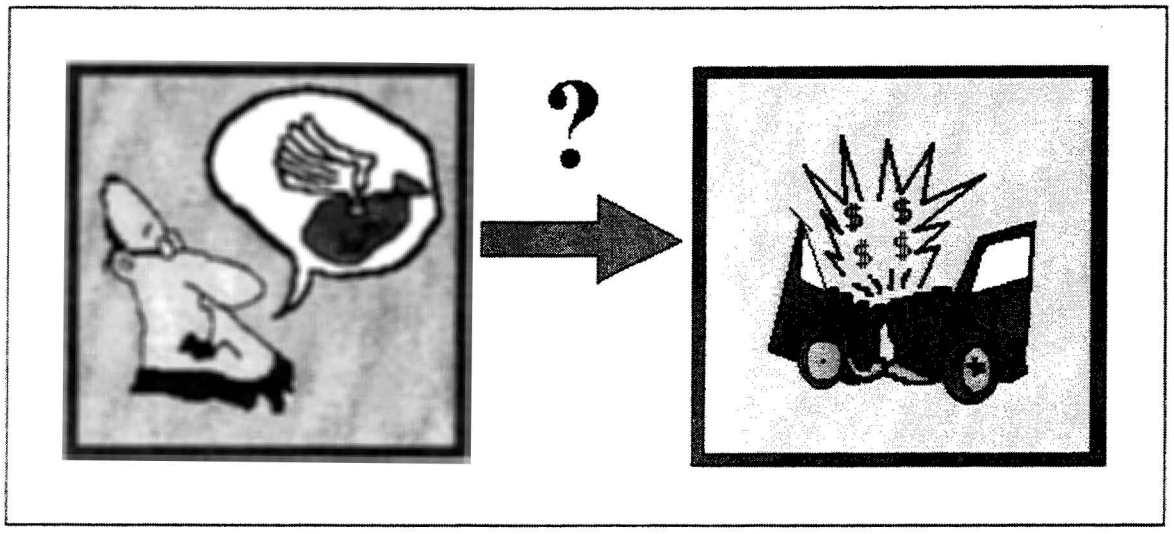


Figura 10: Estimação

Como alguns exemplos desse tipo de problema podemos citar:

- Estimação do número de crianças em uma família;
- Estimação da renda total de uma família; e
- Estimação do valor do cliente para a empresa.

Previsão

A técnica de previsão resume-se na avaliação do valor futuro de algum índice, baseando-se em dados do comportamento passado deste índice. Técnicas utilizadas nos problemas de classificação e estimação podem ser adaptadas para o uso de previsões. O único meio

de verificarmos se uma previsão foi bem feita é aguardar o acontecimento e verificar o quanto foi acertada ou não a previsão realizada.

Como exemplo, temos:

- Previsão da população de uma cidade daqui a 10 anos;
- Previsão de vendas;
- Previsão de níveis de gastos; e
- Previsão de fenômenos físicos e naturais.

Afinidade de Grupos (Agrupamento)

O objetivo desse tipo de problema é a identificação de relações diretas entre objetos, isto é, formar grupos baseados no princípio de que esses grupos devem ser os mais homogêneos em si e mais heterogêneos entre si. A diferença entre afinidade de grupos e classificação, é que aqui não existem classes pré-definidas.

Os registros são agrupados em função de suas similaridades básicas (BERRY; LINOFF, 1997) e (HAN; KAMBER, 2001). Por exemplo, podemos utilizar dados do Censo Demográfico para formar grupos de domicílios, utilizando atributos como situação conjugal, escolaridade, número de filhos. Observe que não existem classes pré-definidas e poderemos ter num mesmo grupo, domicílios de estados geograficamente opostos, porém semelhantes nestas variáveis. Outro exemplo desse tipo de problema é a identificação de padrões de compra em supermercado[9]. Através de identificação de padrões de compras, estratégias de planificação de produtos em prateleiras e mostruários podem ser otimizados.

Esse tipo de problema pode identificar produtos cuja venda está relacionada e indicar potenciais para aumento de venda. Dos números obtidos através da análise de afinidades, podem-se extrair regras de associação que regem o consumo de alguns itens.

Como exemplo, podemos ter:

- Vendas associadas entre diferentes produtos; e
- Serviços associados a diferentes clientes.

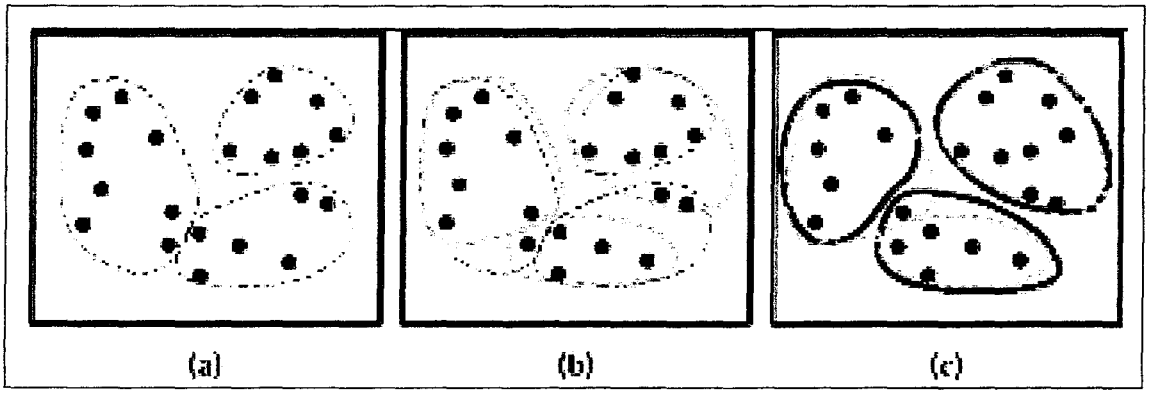


Figura 11: Três critérios diferentes de formação de agrupamentos (*clusters*)

Segmentação

Segmentação é a criação de uma divisão em um conjunto de dados no qual todos os membros de cada subconjunto são semelhantes de acordo com alguma medida. O que diferencia os segmentos (clusters) de uma classificação é o fato de não necessitar de classes pré-definidas, e a formação dos grupos ser conduzida pelo analista, que validará os significados e a representatividade dos agrupamentos, não sendo determinada pelo sistema (WEISS; INDURKYA, 1997).

---

[9] padrões de compra em supermercado (market basket analysys)

Por exemplo, agrupar sintomas pode gerar classes que não representam nenhuma doença explicitamente, uma vez que doenças diferentes podem ter os mesmos sintomas.

Como exemplos, temos:

- Determinação do número de região de vendas; e
- Determinação de grupos de consumidores potenciais.

Descrição

Algumas vezes, o objetivo de um projeto de Mineração de Dados é simplesmente descrever o que está acontecendo em um banco de dados para melhor entender as relações entre pessoas, produtos ou processos. Uma boa descrição do comportamento dos dados geralmente sugere uma explicação, ou no mínimo, onde iniciar a busca por explicação.

Exemplos desse tipo de problema poderiam vir de uma análise das características de uma pessoa que frauda cartões de crédito (do tipo "sexo masculino, idade entre 25 e 40 anos e possui nível superior"), e do perfil do eleitorado de um partido político: ("mulheres apóiam o partido X em um número maior que homens"). Essa descrição pode provocar grande interesse e necessidade de análises detalhadas por parte de jornalistas, economistas, cientistas políticos, além dos próprios partidos.

Reconstrução de funções

Consiste em reconstruir uma função, através de um número de valores conhecidos dessa mesma função, pelo menos em uma dada região de seu domínio.

Problemas desse tipo aparecem em:

- Computação de superfícies de demanda;
- Cálculo da elasticidade de preço;
- Processos de ajuste de máquinas; e
- Interpretação de experimentos físicos.

2.4.2 Métodos de Abordagem

Neste tópico serão apresentados alguns métodos de abordagem de Mineração de Dados. Por método de abordagem, definimos as técnicas gerais que são utilizadas no processo de Busca do Conhecimento. Esses métodos originaram-se de diferentes áreas de conhecimento.

Westphal & Blaxton (1998), subdivide tais métodos em 2 conjuntos: Métodos de visualização e Métodos analíticos não-visuais.

A tabela abaixo nos mostra os componentes dos 2 métodos.

Tabela 2.1
Métodos de Abordagem

Métodos de Visualização	**Métodos Analíticos Não-Visuais**
Clustering	Técnicas Estatísticas
Análise de Hierarquias	Árvores de Decisão
Redes Auto-Organizadoras	Regras de Associação
Sistemas de Geoprocessamento	Redes Neurais
Visualização	Algoritmos Genéticos

Fonte: WESTPHAL, C.; BLAXTON, 1998. 617 p.

Geralmente, a análise de dados é feita utilizando métodos analíticos não –visuais, como veremos em seguida. Entretanto, os métodos analíticos requerem que uma hipótese deva ser criada e que exista um conhecimento anterior do conjunto de dados. Os métodos de visualização, por outro lado, permitem a descoberta de padrões gerais no banco de dados, ao mesmo tempo que possibilita a definição de pequenos padrões locais que podem ser de extrema importância.

Normalmente os métodos de visualização não são utilizados separadamente dos métodos analíticos. Diferentes métodos são combinados durante um processo de busca, sendo os de visualização os primeiros, como forma de exploração inicial dos dados.

2.4.2.1 Métodos de Visualização

As técnicas de visualização são métodos úteis para a descoberta de padrões em um conjunto de dados. As técnicas de visualização têm se desenvolvido rapidamente.
Técnicas gráficas avançadas em realidade virtual possibilitam o tratamento de dados em espaços artificiais, enquanto que o desenvolvimento histórico de um conjunto de dados pode ser acompanhado através de uma animação ou filme.

Alguns benefícios da utilização dos métodos de visualização podem ser constatados, como:

- Observações podem ser feitas sem a existência de hipóteses ou concepções anteriores sobre os dados; e
- *Outlines*, freqüências de ocorrências, e interdependências entre os dados são rapidamente apontados.

No entanto uma das grandes desvantagens está no fato de que os métodos de visualização possuem contribuição limitada quando o número de dados é elevado.

A visualização é uma abordagem iterativa, onde diferentes cenários devem ser testados e observados.

Para que os dados possam ser analisados devem ser modelados de um modo posicional, geralmente em 2 ou 3 coordenadas. Dessa forma, traduzindo os dados através de atributos em um modelo posicional, podem ser processados utilizando métodos como *clustering*, hierarquias, superfícies e outros formatos.

Clustering

A técnica de *clustering* consiste em detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, em caso da existência, determinar estes grupos (DINIZ; NETO,2000). Os dados são colocados no gráfico baseados em valores de algum de seus atributos em uma das coordenadas em observação (x, y ou z). Tipicamente, quando esses valores representam um conjunto arbitrário de descrição (ex.: nomes, doenças, números de identificação, tipos de contas) aparecerão concentrações de dados com valores em comum. Esses métodos procuram de forma direta por uma partição aproximadamente ótima dos n elementos sem a necessidade 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 baseada no princípio de que os grupos formados devem ser os mais homogêneos em si e mais heterogêneos entre si, uma vez que pesquisar todas as possíveis partições tornaria o problema de difícil solução com k_{n-1} partições a serem pesquisadas (BUSSAB; MIAZAKI; ANDRADE,1990). A procura por uma partição ótima pode muitas vezes ser impraticável devido a enorme quantidade de partições possíveis.

A abordagem mais simples é o diagrama de dispersão (*scatter plot*), utilizado para visualizar informação de dois atributos em um espaço cartesiano, no entanto diversas metodologias podem ser utilizadas nesse método de abordagem (ARABIE; HUMBERT; SORTE, 1996).

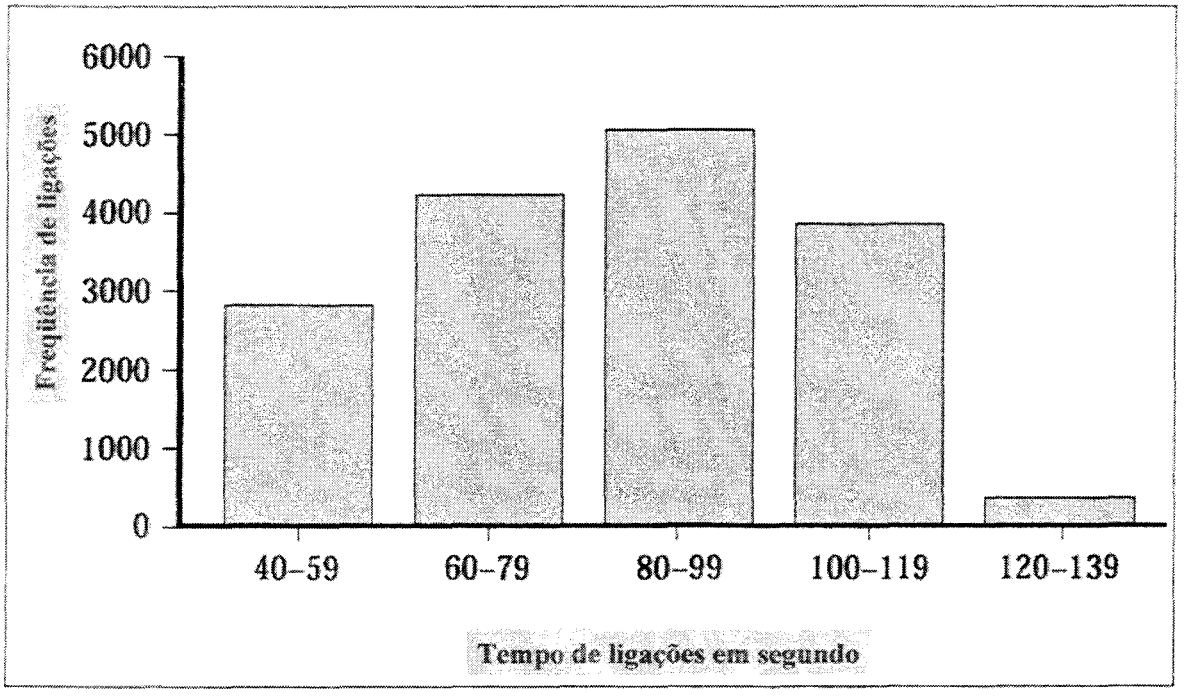


Figura 12: Exemplo de Clustering utilizando abordagem visual

Análise de Hierarquias

Quando a posição de um objeto é baseada na relação com outros objetos forma-se uma hierarquia na visualização. Essa abordagem é utilizada quando o conjunto de dados possui tipos e classes de objetos que estão ligados uns aos outros. A parte superior é chamada de nó raiz e sua seleção determina como uma hierarquia vai ser construída para os propósitos da visualização. Como cada nível está ligado ao nível anterior, não existe ligações entre objetos do mesmo nível ou relações assumindo ligações bidirecionais. Dependendo do tipo de aplicação, podemos ter mais de um nó raiz, como podemos observar na figura abaixo.

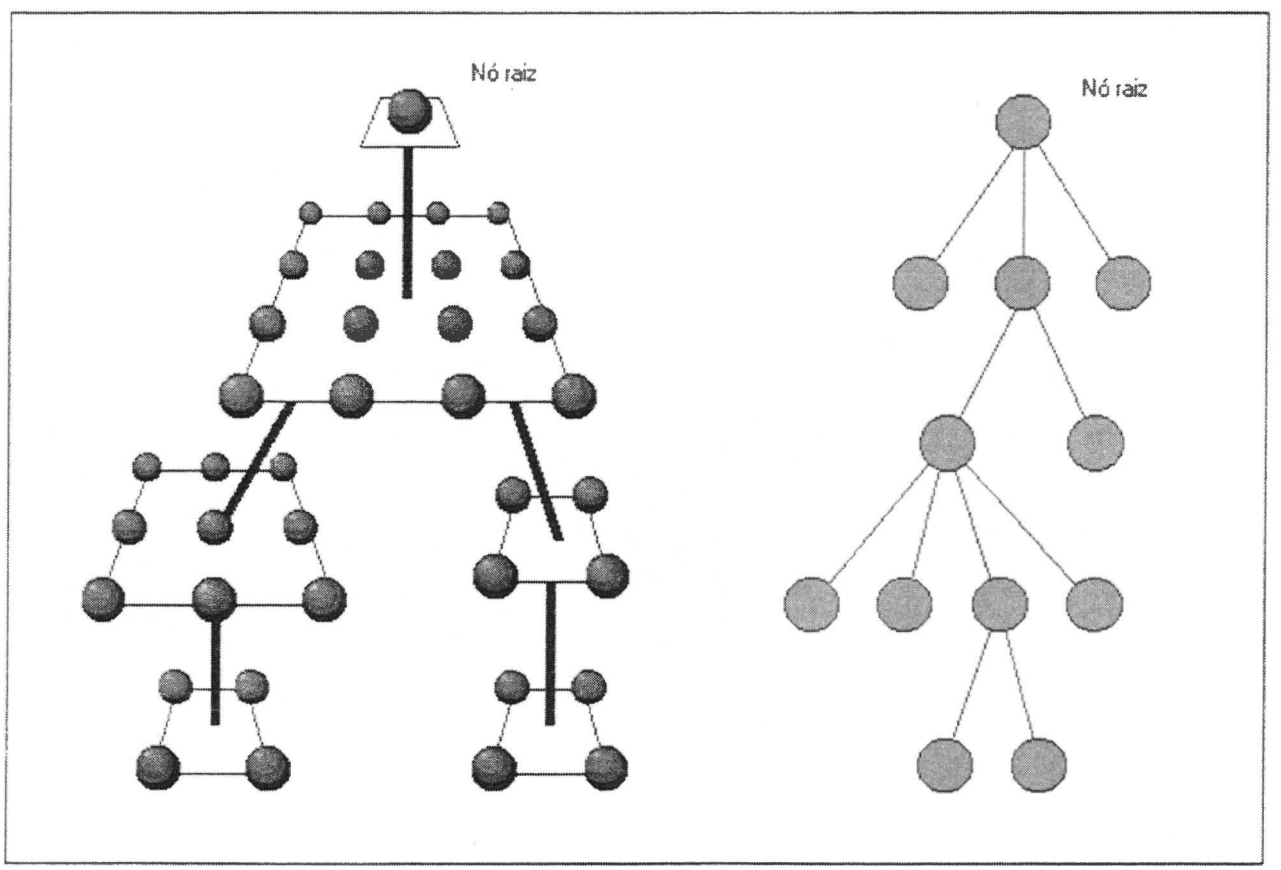


Figura 13: Hierarquia (Westphal; Blaxton, 1998)

As hierarquias não devem necessariamente ser formadas em relação de ligação. Algumas hierarquias são construídas observando os valores dados a certos atributos dos objetos que por sua vez são utilizados de maneira absolutas em camadas.

O método do centróide é o mais direto deles, pois substitui cada fusão de objetos por um único ponto representado pelas coordenadas de seu centro que é a média dos valores dos objetos que estão no grupo. A distância entre grupos é definida pela distância entre seus centros. Em cada etapa procura-se fundir grupos que tenham a menor distância entre si. Este método pode produzir resultados confusos, pois a distância entre os centros de um par pode ser menor que a distância entre outro par unido numa combinação anterior.

Redes auto-organizadoras

Sendo uma técnica recente, normalmente é confundida com a técnica de Redes Neurais, mas nas Redes auto-organizadoras, os dados estudados, tendem a se repetir. Sua ligação é estudada através das relações entre si.

Dessa forma, essas redes organizam os dados em números discretos de grupos ou segmentos. Seu funcionamento inicia com os dados reunidos e automaticamente a rede se ajusta até obter uma relação de equilíbrio no sistema. Dependendo do tipo de dados, essa abordagem pode produzir resultados que poderão ser melhor analisados, onde a rede se subdividirá em pequenos grupos.

Sistemas de Geoprocessamento

A visualização pode ser conduzida utilizando sistemas de GIS (Geographic Information Systems). São utilizados para auxiliar na análise de dados onde existe uma relação com sua localização geográfica.

Sistemas de GIS consistem na utilização de banco de dados com informações de coordenadas geográficas com mapas digitalizados. Dessa forma, podemos analisar segmentos a partir de uma mesma característica comum, visualizados dinamicamente em mapas digitais.

Visualização e Sumarização

Um dos principais objetivos da tecnologia de mineração de dados é oferecer seus resultados numa forma fácil de ser interpretado pelos usuários finais. Utilizar a sumarização de dados para facilitar o entendimento dos dados é uma estratégia muito usual que facilita e identifica inúmeras características nos dados em estudo.

Uma das principais abordagens para descrição de informações é a visualização, principalmente quando o conjunto de dados a ser explorado não está organizado em uma forma padrão. Os resultados da sumarização e da visualização são normalmente utilizados em conjunto com outras técnicas (WEISS; INDURKYA, 1997).

Por exemplo, podemos imaginar um gráfico de colunas impresso num mapa do Brasil, indicando em cada estado o número de chamadas telefônicas realizadas no ano de 2000.

Facilmente, podemos comparar esses resultados entre os estados. Se colocarmos os dados de dois anos, nossa análise será ainda mais rica.

A Figura 14, a seguir, é um exemplo de mineração de dados fornecendo seus resultados com técnicas de sumarização e visualização.

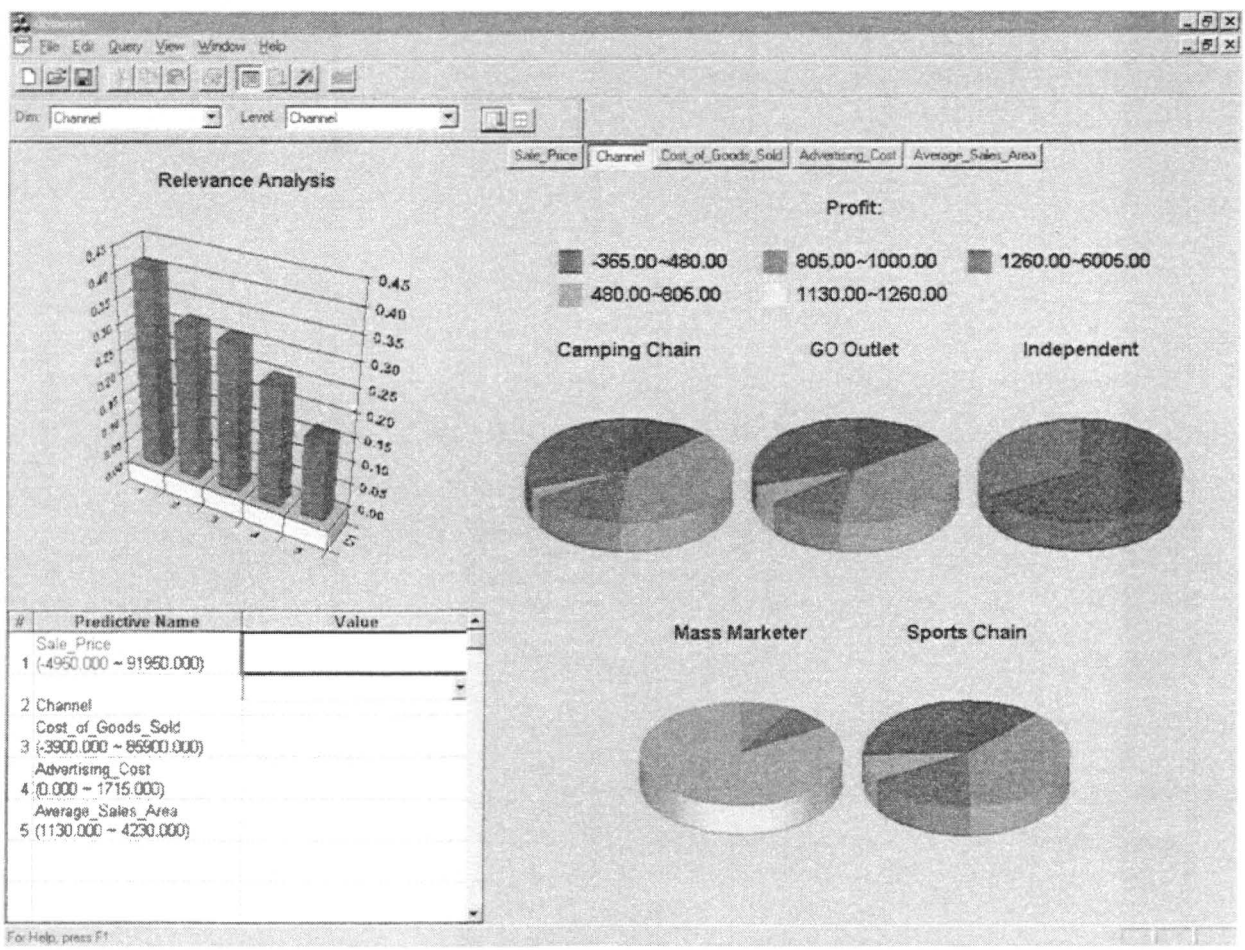


Figura 14: Mineração de dados com resultados da sumarização e visualização

2.4.2.2 Métodos Analíticos Não Visuais

Técnicas estatísticas

O uso da estatística descritiva e inferencial é um dos mais usados em projetos de Mineração de dados. A utilização de técnicas estatísticas requer o uso de dados quantitativos. Dados qualitativos devem ser codificados para valores numéricos. Os testes estatísticos podem ser usados para a comparação de valores de amostras de grupos diferentes. Em geral nos casos mais simples, a estatística descritiva é utilizada para obter uma visão geral das características dos dados analisados.

A estatística descritiva inclui medidas como a média, mediana, moda, desvio-padrão, intervalo de confiança e distribuição dos dados.

Dois grupos de testes estatísticos estão sendo amplamente utilizados em etapas iniciais nos projetos de Mineração de dados:

- Análise de Diferenças de Grupos : Consiste em utilizar testes estatísticos de hipótese para comparação entre dois grupos de dados. Por exemplo, suponha que tenhamos semanalmente a receita de uma loja durante o período de um ano. Teremos então a evolução das vendas ao longo dos meses semanalmente. Poderemos através desse tipo de análise inferir se as vendas no mês de dezembro são diferentes do mês de novembro. Testando a hipótese nula, onde não existe diferença e uma hipótese alternativa, onde existem diferenças. Por meio de testes

estatísticos, como o teste-*T*, análise de variância e teste *F*, podemos rejeitar ou não a hipótese nula e a significância do resultado (a probabilidade do efeito observado estar correto e não ser um efeito de aleatoriedade).

- Análise de Previsão com Regressão Linear: É utilizada quando previsões de valores numéricos devem ser feitas. Estará sendo determinada a curva que melhor se ajusta ao conjunto de dados. Esse tipo de análise possui dois resultados: a função da curva e uma medida de correlação. A medida de correlação descreve o grau de variação entre a curva e os dados reais.

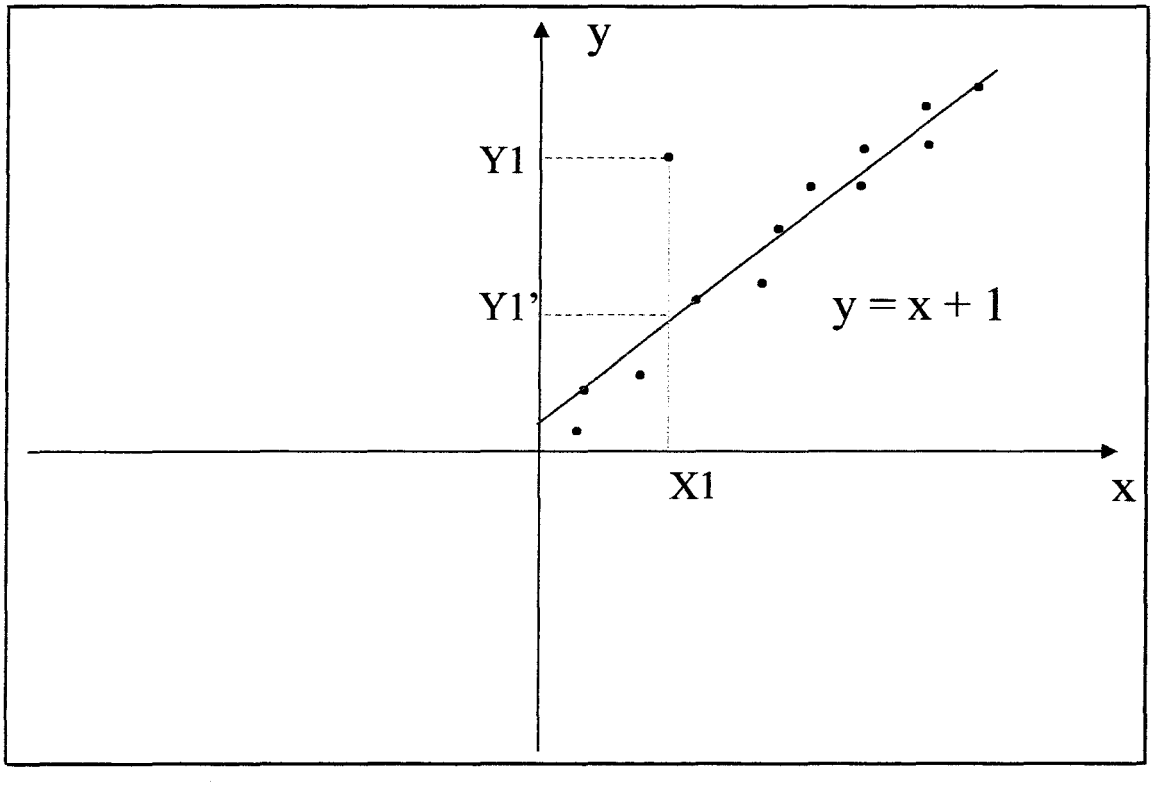


Figura 15 : Representação de uma Regressão Linear

A técnica de regressão pode ser usada com outras funções diferentes da reta, como por exemplo curvas logarítmicas, polinomiais e exponenciais.

Dessa forma, para utilização de testes estatísticos, algumas condições devem ser satisfeitas, como por exemplo o uso de variáveis quantitativas e a existência de uma hipótese a ser testada.

Amostras que possuam uma variância muito grande ou que possuam poucas observações podem ter pouco poder explicativo, e os resultados das análises podem ter pouca validade.

Árvore de decisão

Uma árvore de decisão é um fluxograma semelhante a uma estrutura de árvore. Seu nome é originado pela característica de sua estrutura, formada por ramos e nós. Em cada nó, uma decisão é tomada ou a ocorrência de uma probabilidade é avaliada. Se todos os itens pertencem a mesma classe então nenhuma decisão futura seria necessária para particionar os itens e, sendo assim, a solução estaria completa. Em caso dos itens do nó intermediário pertencerem a duas ou mais classes, então uma regra seria elaborada e este nó resultaria em uma divisão.

Este processo é repetido diversas vezes até que uma árvore completa seja obtida. É necessário todo cuidado pois podem existir componentes que foram criados por outliers presentes nos dados de treinamento. Por isso, vários métodos de decisão por árvore utilizam o recurso de poda, que tenta generalizar a árvore eliminando sub-árvores que parecem ser muito específicas.

Baseado no modo como a árvore é "podada", obtemos um algoritmo diferente de árvore de decisão, tais como: CART, CHAID, C4.5, etc...

- CART – Classification And Regression Tree – Nesta técnica uma função de avaliação usada para dividir os nós, é o índice *Gini*. Para um dado nó t, este índice é definido como:

$$\textbf{Gini (t)} = 1 - \sum p_i^2$$

onde p_i é a probabilidade da i-ésima classe no nó t.

A poda por 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[10]. Mas , sua razão de erro pode ser muito maior quando aplicamos ao conjunto de dados de teste. Sendo assim, o objetivo deste processo de poda é o de encontrar a sub-árvore que produz o menor erro, considerando o tamanho da árvore (BREIMAN et al, 1984).

- CHAID - (*Chi-squared automatic interaction detection*). O critério de separação é baseado em *p*-valores de *F*-distribuições (intervalos) ou distribuições chi-quadradas (nominais). Os *p*-valores são ajustados para acomodar múltiplos testes.

Para as entradas nominais um valor perdido constitui uma nova categoria. Já para os ordinais, ele está livre de qualquer restrição de ordem. Muitos software aceitam entrada de dados intervalares, e automaticamente agrupam os valores em faixas antes de criar a árvore. Inicialmente uma ramificação é alocada para cada valor de entrada. Ramificações são fundidas alternadamente e re-divididas como parece garantido pelos *p*-valores.

[10] dados considerados para a construção da árvore

O algoritmo para, quando nenhuma operação de fundição ou re-divisão cria um *p*-valor adequado.

Uma alternativa comum, chamada de método exaustivo, continua fundindo para uma divisão binária, e depois adota a divisão com o *p*-valor mais favorável dentre todas as divisões consideradas pelo algoritmo. Depois que uma divisão é adotada, seu *p*-valor é ajustado, e a entrada com o melhor *p*-valor ajustado é selecionada como a variável de divisão. Se o *p*-valor ajustado é menor que um limite especificado, o nó é dividido.

A construção da árvore termina quando todos os *p*-valores ajustados, das variáveis de divisão nos nós não divisíveis, estão acima do limite especificado pelo usuário.

- C4.5 - Dado um nó t, o critério de divisão usado e:

RazãoGanho(t) = ganho (t) / Informação_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[11] para o teste (DINIZ; NETO, 2000).

[11] Entropia (t) = $\sum_i - p_i \log p_i$ onde *pi* é a probabilidade da i-ésima classe dentro do nó t.

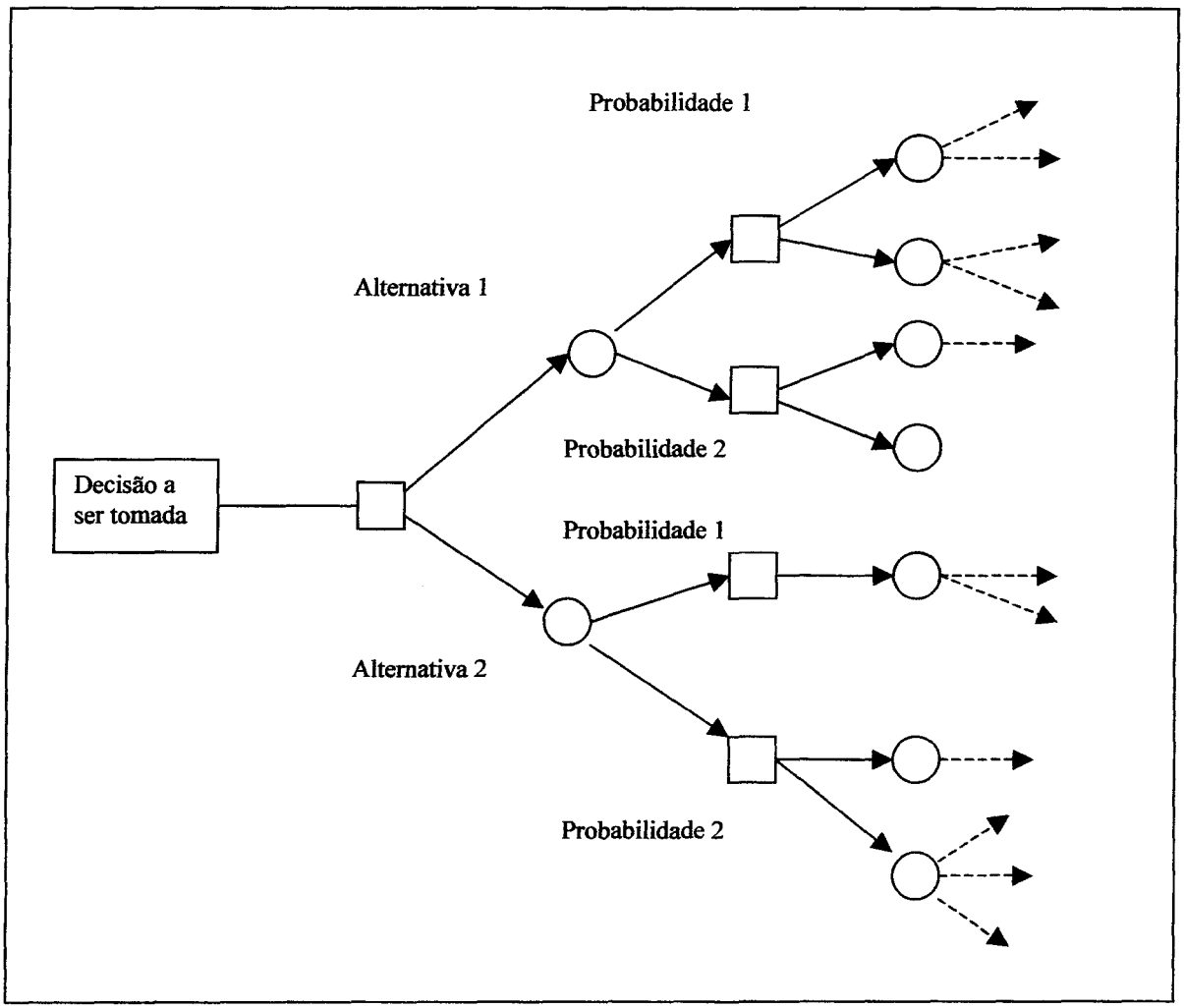


Figura 16: Estrutura de uma árvore de decisão

Tomemos como exemplo um loja de automóveis que gostaria de traçar o perfil de seus clientes. Os dados analisados contém características de vendas passadas, como sexo do comprador, idade e tipo de carro comprado (motor, número de portas, cor, estilo, país de origem). Decisão : Compra de um carro

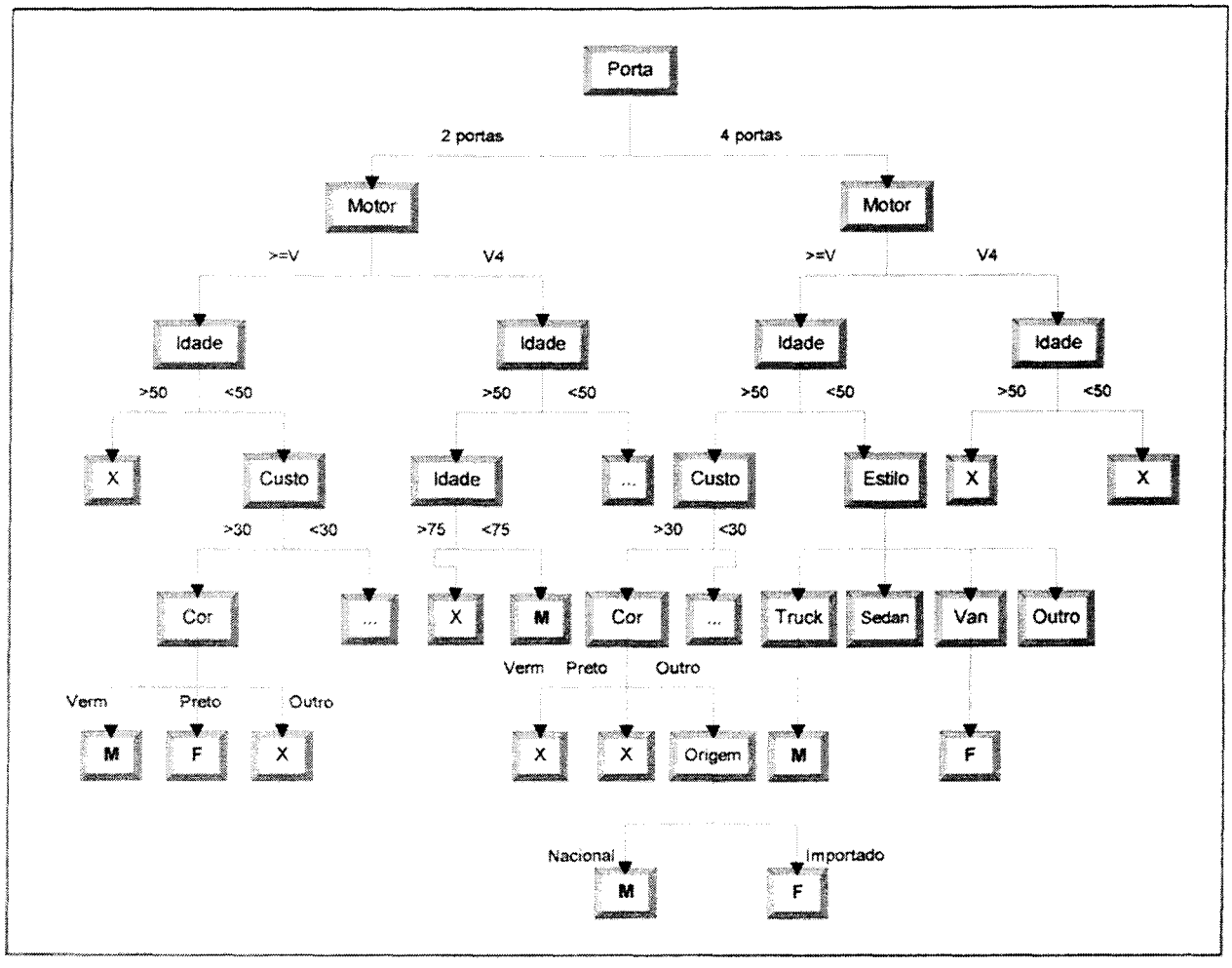


Figura 17: Exemplo Árvore de Decisão: Perfil de compra (Westphal & Blaxton)

O algoritmo utilizado para subdividir os dados procura as variáveis que consigam uma máxima separação do arquivo. No exemplo, a primeira variável escolhida foi o número de portas que o carro deve ter, pois é a variável que divide igualmente o conjunto de dados. Qualquer das variáveis poderia ter sido a escolhida, mas o algoritmo escolhe as variáveis que segregam ao máximo o número de registros, trazendo significância para a árvore.

Analisando o resultado da construção da árvore de decisão, podemos observar que se olharmos na figura o lado esquerdo teremos:

1. Se o veículo possuir 2 portas....e
2. Se o veículo possuir 6 cilindros ou mais..... e
3. Se o comprador tiver menos de 50 anos .....e
4. Se o veículo custar mais de R$ 30.000 .....e
5. Se a cor do veículo for Vermelha, então ⇒ O comprador possui alta probabilidade de ser do sexo **Masculino**

Por outro lado se olharmos o lado direito teremos:

1. Se o veículo possuir 4 portas ....e
2. Se o veículo possuir 6 cilindros ou mais.....e
3. Se o comprador tiver menos de 50 anos. ...e
4. O veículo for uma Van ⇒ O comprador possui alta probabilidade de ser do sexo **Feminino**

Regras de Associação

Regras de associação é o processo que extrai informação de padrões que se repetem. Um exemplo comum é aquele referente à cesta do supermercado. Neste caso, a cesta do supermercado corresponde àquilo que o consumidor compra em um supermercado durante uma visita (ELMASRI;NAVATHE,2000). Na área de marketing é conhecido como análises de transações de compras (*market basket analysis*).

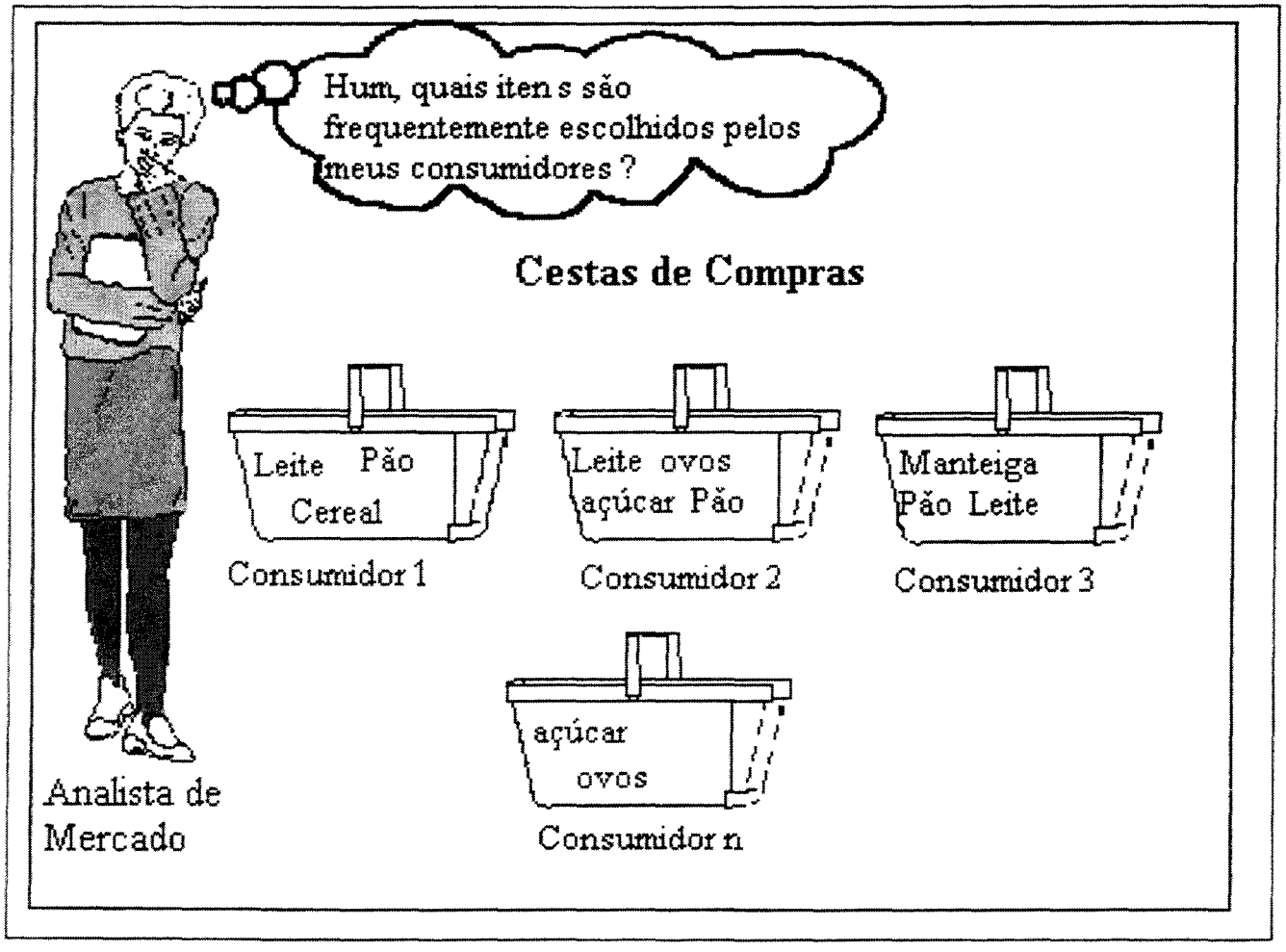


Figura 18: Um exemplo de análises de transações de compras extraído de (HAN; KAMBER,2001)

As Regras de Associação são obtidas através de matriz de inter-relação, onde a probabilidade do acontecimento conjunto de cada evento é calculada (DINIZ; NETO,2000).

Um estudo completo sobre regras de associação pode ser encontrado em (MENDES; PLASTINO;OCHI,2001). No exemplo abaixo, estamos considerando a análise de 12 produtos. No entanto, a análise poderia ser estendida para n produtos.

A matriz nos mostra que são muitas vezes comprados em conjunto (células possuindo mais de um **x**), produtos que algumas vezes são comprados em conjunto (células que possuem apenas um **x**) e aqueles pares que não possuem relações em padrões de compra (células em branco).

	Leite	Morango	Pão	Carne	Champanhe	Lubrif	Café	Ração	Pasta	Ovos	Cereal	Xarope
Leite			xxxx	x		x	x		x	xx	xx	
Morangos				x	xxx							
Pão									x	xx	xx	
Carne												
Champagne								xx				
Lubrificante												
Café										x		xxx
Ração												
Pasta												
Ovos												x
Cereal												
Xarope												

Figura 19: Exemplo de Matriz de Inter-relação para Regras de Associação (Westphal & Blaxton)

Pela matriz podemos notar que existe uma forte relação entre a compra de pão e de leite. O mesmo acontece com relação a cereais e leite. São relações inicialmente intuitivas, e que a princípio, não trazem maiores esclarecimentos. No entanto, notamos relações não tão diretas, como a relação entre morangos e *champagne* e entre ração e *champagne*. O grande benefício desta técnica está na descoberta desses padrões não intuitivos e sua posterior análise.

Por outro lado, o uso de uma matriz de inter-relação possui suas limitações, pois embora a relação entre morango e *champagne* possa ser compreendida, a relação entre *champagne* e ração parece ser de pouco uso. Fazendo uma análise desta relação, suponhamos que as pessoas que possuam um animal de estimação comprem ração semanalmente. No entanto, a pessoa não consome ração em conjunto com *champagne*. A regra de associação apresentada, relaciona os dois produtos, já que são comprados em conjunto regularmente. Tomar como base essas regras para padrões gerais de consumo, pode não resultar em uma boa estratégia de marketing, pois não é intuitivo que todas as pessoas que consomem *champagne* estejam interessadas em comprar ração animal. Logo, cabe ao analista determinar quais são as "*boas*" regras de associação e as regras "*sem uso*".

Redes Neurais

As redes neurais artificiais consistem em um método de solucionar problemas de inteligência artificial, construindo um sistema que tenha circuitos que simulem o cérebro humano, inclusive seu comportamento, ou seja, aprendendo, errando e fazendo descobertas. São mais que isso, são técnicas computacionais que apresentam um

modelo inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência. Uma grande rede neural artificial pode ter centenas ou milhares de unidades de processamento, enquanto que o cérebro de um mamífero pode ter muitos bilhões de neurônios. Assim como o sistema nervoso é composto por bilhões de células nervosas, a rede neural artificial também seria formada por unidades que nada mais são que pequenos módulos que simulam o funcionamento de um neurônio. Estes módulos devem funcionar de acordo com os elementos em que foram inspirados, recebendo e retransmitindo informações (HAYASHI,1991).

Redes Neurais são uma classe de modelagem de prognóstico que trabalha por ajuste repetido de parâmetros. Estruturalmente, uma rede neural consiste em um número de elementos interconectados (chamados neurônios) organizados em camadas que aprendem pela modificação da conexão entre elas

A função básica de cada neurônio é: (a) avaliar valores de entrada, (b) calcular o total para valores de entrada combinados, (c) comparar o total com um valor limiar, (d) determinar o que será a saída. Enquanto a operação de cada neurônio é razoavelmente simples, procedimentos complexos podem ser criados pela conexão de um conjunto de neurônios. Tipicamente, as entradas dos neurônios são ligadas à uma camada intermediária (ou várias camadas intermediárias) que é então conectada com a camada de saída, como mostra a figura baixo:

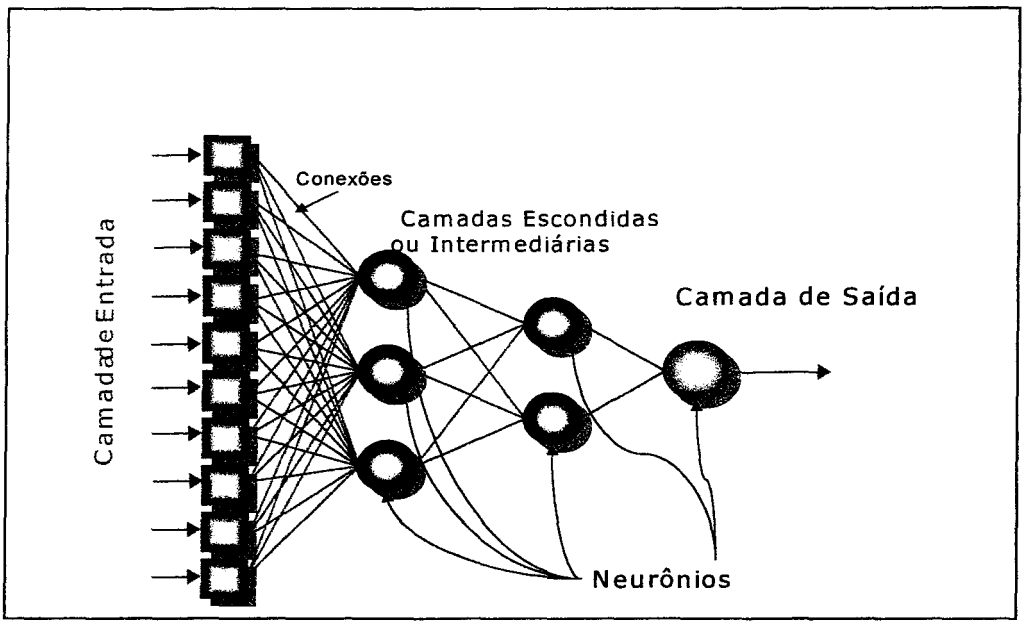


Figura 20: Mostra uma representação de um processamento de uma rede neural

A maioria dos modelos de redes neurais possui alguma regra de treinamento, onde os pesos de suas conexões são ajustados de acordo com os padrões apresentados. Em outras palavras, elas aprendem através de exemplos.

Para construir um modelo neural, nós primeiramente "adestramos" a rede em um *dataset* de treinamento e então usamos a rede já treinada para fazer predições.

Cada neurônio geralmente tem um conjunto de pesos que determina como o neurônio avalia a combinação dos sinais de entrada. A entrada para um neurônio pode ser positiva ou negativa. O aprendizado se faz pela modificação dos pesos usados pelo neurônio em acordo com a classificação de erros que foi feita pela rede como um todo. As entradas são geralmente pesadas e normalizadas para produzir um procedimento suave.

Durante a fase de treinamento, a rede estabelece os pesos que determinam o comportamento da camada intermediária. Um estudo completo sobre regras de redes neurais pode ser encontrado em (HAYASHI,1991). Falaremos um pouco mais no Estudo de Caso

Algoritmos Genéticos

Enquanto os métodos apresentados anteriormente estão relacionados com problemas de classificação e previsão, a utilização de algoritmos genéticos está associada a otimização. Através de analogias desenvolvidas com a teoria da evolução, os algoritmos genéticos iniciam com uma população de itens e procuram alterar e eventualmente otimizar sua composição para solucionar determinado problema. Uma definição formal é que um algoritmo genético é um tipo de algoritmo de computação que modela o processo biológico genético, incluindo as trocas cruzadas (crossover) e operações de mutação (ELMASRI; NAVATHE,2000).

Indivíduos dentro da população (por exemplo possíveis soluções para um problema) evoluem através de diferentes gerações. O material genético de informação apresentada por cada indivíduo é passado para as gerações subseqüentes em diferentes modos no processo de otimização. Dentro desse método, existem 3 mecanismos básicos pelos quais a informação é escolhida, alterada e passada no processo de otimização: seleção, trocas cruzadas e mutação.

O processo de seleção implementado nos algoritmos genéticos é análogo ao processo de seleção natural que ocorre na evolução. A seleção é baseada no princípio de

sobrevivência daquele que melhor se adequa ao ambiente. Esses indivíduos são aqueles que mais freqüentemente passam o material genético para a próxima geração. Os valores de adequação são calculados para todos os indivíduos, ou genomas, da população e aqueles que possuem os maiores valores podem reproduzir. Genomas que possuem valores baixos de adequação possuem um número menor de cópias que sobrevivem na próxima geração. O método de escolha de genomas é feito de modo aleatório em uma população onde o número de indivíduos representados é proporcional aos valores de adequação. Quanto maior esse valor, maior o número de indivíduos na população e maior a chance que a informação contida em seu genoma sobreviva na próxima geração. Outras formas de seleção também podem ser incorporadas, através de ordenações.

Nos algoritmos genéticos mais simples, toda a população é substituída a cada ciclo e o tamanho da população permanece constante. Entretanto, em modelos mais sofisticados, o tamanho da população pode crescer ou diminuir. No caso de modelos onde a população cresça, indivíduos com valores baixos de adequação sobrevivem mais tempo, podendo contribuir com uma maior chance na solução ótima.

Além disso, deve existir uma grande uniformidade no banco de dados, já que os itens a serem analisados devem ser codificados em vetores de mesma dimensão. Por esta razão, é improvável sua utilização em situações onde os dados formam combinados de diversas fontes e tipos de informação.

Sua maior utilização está em áreas como maximização do lucro através de combinações de mix de produtos. Além disso, tem sido utilizado em aplicações que envolvem

organização de cronograma e análise de séries temporais. Sua combinação com redes neurais é amplamente utilizada, gerando diminuições em tempos de processamento.

2.5 Conclusões

Neste capítulo apresentamos os principais conceitos para entendimento desta dissertação. Inicialmente destacamos o tratamento de dados ausentes e os métodos mais usados para a imputação de dados.

Logo a seguir foram introduzidos conceitos de *Busca de Conhecimento em Banco de Dados*, também conhecido como *Knowledge Discovery in Database - KDD*, com destaque para as etapas de preparação dos dados para Mineração de Dados propriamente dita.

No próximo capítulo trataremos da abordagem dos métodos para o uso de Mineração de Dados.

3. Abordagens da Mineração de Dados

As abordagens da mineração de dados ou metodologias de aplicação descrevem como o usuário irá conduzir o processo da mineração na obtenção de seus resultados.

Essencialmente existem as abordagens *top-down* (direta) e *botton-up* (indireta), e uma terceira que pode ser a combinação dessas abordagens chamada de híbrida. Na abordagem *top-down*, também chamada de *teste de hipótese*, o usuário parte do princípio que existe uma hipótese, uma idéia pré-concebida que precisa ser confirmada ou não.

Já na abordagem *botton-up*, também chamada de *busca de conhecimento*, o usuário inicia o processo de exploração dos dados na tentativa de descobrir alguma coisa que ainda não é de seu conhecimento (BERRY; LINOFF, 1997).

3.1 Busca de conhecimento direta

Na busca de conhecimento *direta* ou *supervisionada* sua meta é orientada. Existe um valor para ser prognosticado, uma classe a ser atribuída aos registros ou um determinado relacionamento para ser explorado. Existe apenas uma vaga idéia do que se está procurando. Os passos para aplicação da busca de conhecimento direta são:

- Identificar as fontes dos dados selecionados para mineração;
- Preparar os dados para análise;
- Construir e treinar o modelo computacional; e
- Avaliar o modelo computacional.

Maiores detalhes sobre esses passos podem ser encontrados em (BERRY; LINOFF, 1997).

3.2 Busca de conhecimento indireta

Na busca de conhecimento *indireta* ou *não-supervisionada* não existe uma meta bem definida. As ferramentas são mais livres na sua aplicação sobre os dados e espera-se que seja descoberto alguma estrutura significante nos dados. Os passos para aplicação da busca de conhecimento direta são:

- Identificar as fontes dos dados;
- Preparar os dados para análise;
- Construir e treinar o modelo computacional;
- Avaliar o modelo computacional;
- Aplicar o modelo computacional no novo conjunto de dados;
- Identificar potenciais objetivos para busca de conhecimento direta; e
- Gerar novas hipóteses para teste.

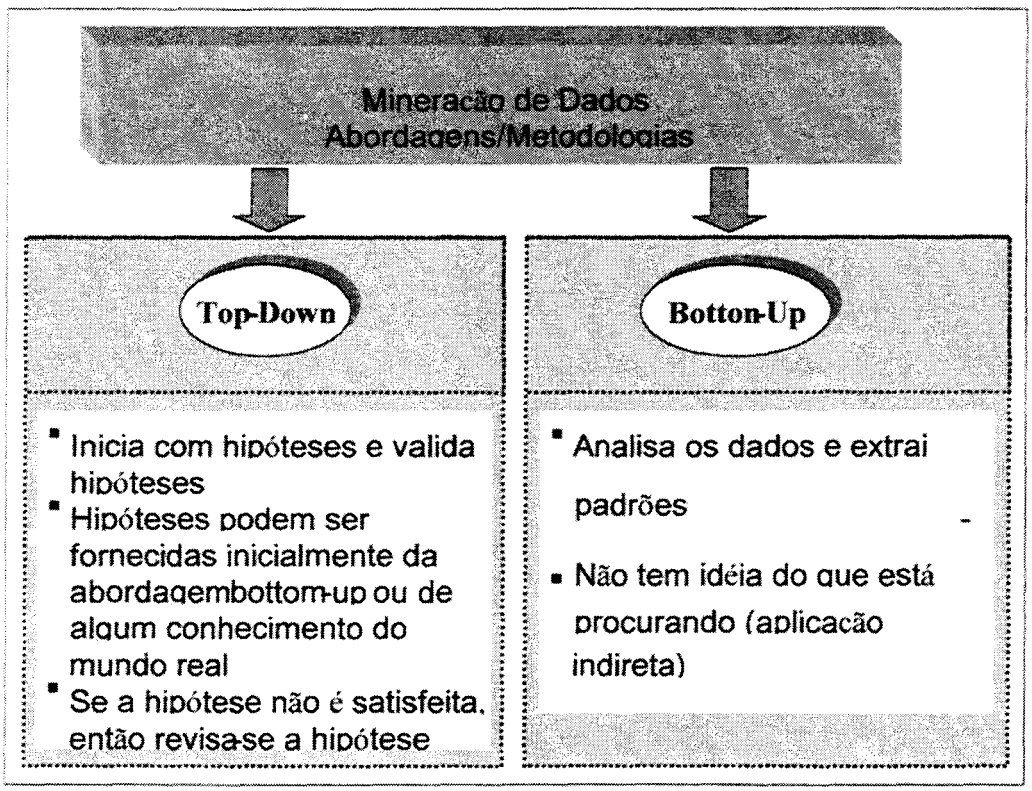


Figura 21: Abordagens para aplicação da mineração de dados

Maiores detalhes sobre esses passos podem ser encontrados em (BERRY; LINOFF, 1997) e (BAILEY; THOMPSON, 1990).

A Figura 21 acima, extraída de (BAILEY; THOMPSON, 1990), resume a forma de aplicação do processo de mineração de dados.

Em um processo de mineração de dados, deve-se ter bem claro qual o resultado a que se deseja chegar. A escolha da técnica, na maioria dos casos exige a participação de pessoas que efetivamente entendam do negócio em estudo, mesmo que não sejam especialistas na utilização e manuseio computacional dos dados.

Diversas técnicas podem ser utilizadas para se chegar aos resultados pretendidos, entretanto, cada técnica possui suas características, suas peculiaridades e precisa de pessoas que saibam interpretar seus resultados. A Figura 22 a seguir, sintetiza um ciclo de vida para operação das técnicas de mineração de dados.

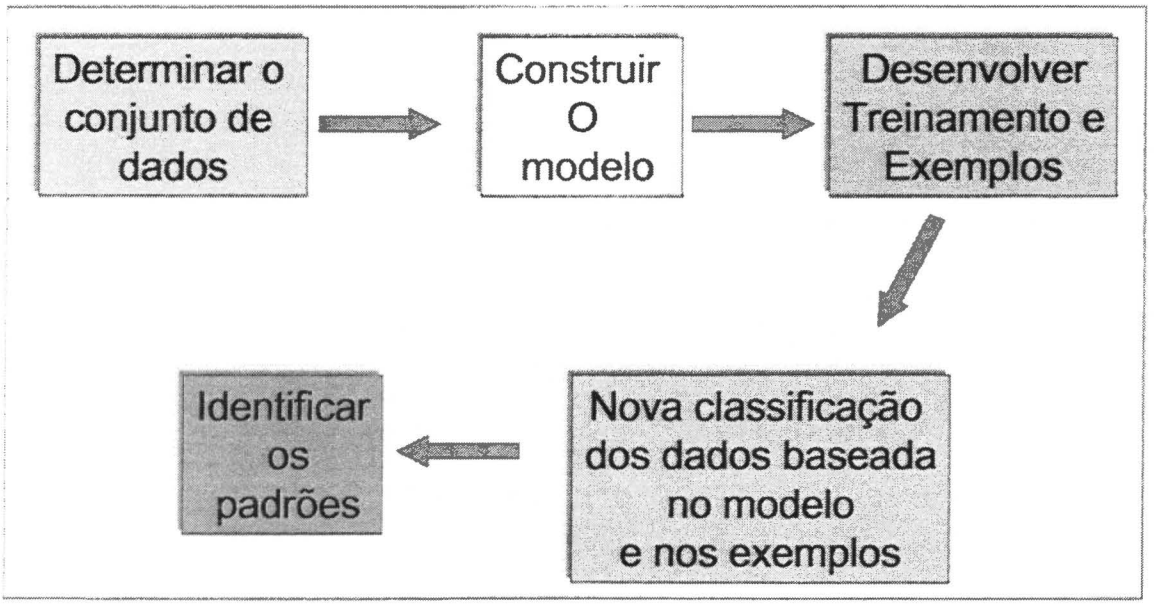


Figura 22: Ciclo de vida de operação das técnicas de mineração de dados

Uma vez identificadas as melhores técnicas a serem aplicadas, deve-se escolher uma abordagem, uma metodologia de aplicação para condução de todos os processos.

3.3 O uso de um valor padrão para a imputação de dados.

Segundo Pao (1989, p.309), o conhecimento de padrões é importante devido às ocorrências na vida humana tomarem forma de padrões. A formação da linguagem, o modo de falar, o desenho das figuras, o entendimento das imagens, tudo envolve padrões. Reconhecimento de Padrões é um tarefa complexa, onde o homem busca, sempre, avaliar as situações em termos dos padrões das circunstâncias que as constituem, para melhor entendê-lo e adaptar-se.

O reconhecimento de padrões é, freqüentemente, altamente especializado. As contribuições têm vindo de muitas áreas de pesquisas distintas. Entre essas áreas destacam-se os sistemas e processamento de sinais, inteligência artificial, teoria de estimação/otimização, conjuntos difusos, modelagem estrutural, linguagem formal, etc... [Kas96]. Isto é uma clara indicação do sucesso da extensão, profundidade de interesse no tópico e do vigor das pesquisas associadas.

Existem, hoje, muitas técnicas de reconhecimento de padrões desenvolvidas, que se baseiam em técnicas matemáticas, estatísticas e/ou técnicas de Inteligência Artificial (redes Neurais, conjuntos difusos, entre outros). Cada uma destas tenta simular o reconhecimento de padrões de forma distinta.

Independentes da técnica usada no reconhecimento de padrões, Bezdek & Pal (1992, p. 532) conceitualizaram o problema em três estágios ou espaços: o espaço do padrão, o

espaço das características e o espaço da classificação (Figura 23.) Segundo ANDREWS, (1972, P.242), o mundo físico é representado por um contínuo de parâmetros que é essencialmente de dimensionalidade infinita. Porém, representa-se o problema do mundo real por *R* características. Esta é a dimensionalidade do espaço dos padrões. Como *R* é, geralmente, muito grande, é desejável reduzí-lo (redução de dimensionalidade) de forma que os dados resultantes ainda mantenham o poder discriminatório dos padrões que estão inerentes aos dados. O espaço das características é postulado de dimensão *N* ($N<R$) no qual as regras de classificação podem ser executadas em tempo razoável. O espaço de classificação, então, simplesmente é o espaço de decisão no qual *K* classes podem ser selecionadas e, portanto, de dimensão *K*.

Conceitualmente, o problema de reconhecimento de padrões pode, então, ser descrito como uma transferência do espaço de padrões *P* (dimensão *R*), para o espaço de características *F* (dimensão *N*) e finalmente para o espaço de classificação *C* (dimensão *K*):

P – F – C (2.1)

Neste processo encontram-se duas transformações:

1) Redução de Dimensionalidade
2) Classificação dos dados em *K* classes

Tanto para a primeira, como para a segunda transformação, existem muitas ferramentas desenvolvidas que são utilizadas para resolver problemas de reconhecimento de padrões reais.

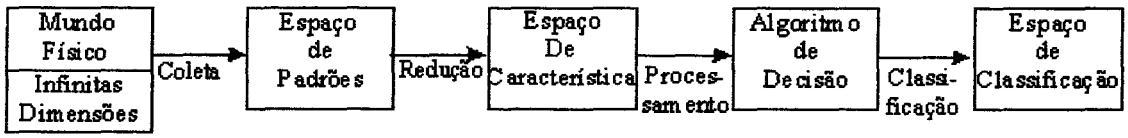


Figura 23: O problema de reconhecimento de padrões
Fonte: Adaptado de ANDREWS, (1972)

Dois elementos essenciais dentro de reconhecimento de padrões devem ser definidos:

- **Padrão**: É um conjunto de características que define um objeto ou um grupo de objetos. É essencialmente um arranjo ou uma ordenação em que alguma organização de estrutura pode ser dita existir. Um padrão pode ser referenciado como uma quantidade ou descrição estrutural de um objeto ou algum outro item de interesse. Um padrão pode ser tão básico quanto um conjunto de medidas ou observações, geralmente sendo representado na forma de vetor ou matriz (PANDYA; MACY, 1996).

- **Características**: Podem ser entendidas como qualquer medição útil extraída no processo de identificação do padrão. Intensidade de sinais são exemplos de características. As características podem ser variáveis contínuas ou discretas.

Sendo o objetivo principal do reconhecimento de padrões a distinção entre os diferentes tipos de padrões, a capacidade de discriminação é colocada em questão. Diferentes padrões são compostos de características diferentes, com valores numéricos diferentes ou relação entre as próprias características diferentes. A maioria dos classificadores (Estatísticos, Sintáticos ou Neurais) são baseados no conceito de similaridade. Por exemplo, se um padrão *x* é muito similar a outro padrão que é conhecido pertencer a uma classe *C1*, intuitivamente a tendência é classificar x como pertencente à classe *C1*.

O problema em reconhecimento de padrões está na sua definição ou composição, já que definir um conjunto de característica que o representa pode não ser uma tarefa trivial. A chave é escolher e extrair um conjunto finito de características que o represente totalmente e que possa ser passível de ser manuseado (BEZDEK; PAL, 1992).

II. Abordagem para Classificação

O Reconhecimento de Padrões, assim como a Inteligência Artificial, teve seu início nos anos 50. Inicialmente, utilizou-se técnicas probabilísticas, mais precisamente, da Estimação e Teoria da Decisão para sua fundamentação. Assim, usou a linguagem da probabilidade em sua origem, usando a abordagem Bayesiana (SCHALKOFF, 1992). Uma outra corrente apareceu depois, utilizando a abordagem estrutural, sendo usado o modelo sintático. Mais recentemente, a modelagem neural e difusa apareceram como abordagens promissoras para esta área. Assim pode-se dividir os métodos de classificação em quatro grupos de reconhecimento de padrões: estatístico, sintático, neural e difuso (PAO, 1989; FU, 1994; BISHOP, 1985).

- *Abordagens estatísticas*

Para a abordagem de tomada de decisão, um conjunto de medidas de características é extraído dos padrões. O reconhecimento de cada padrão, ou seja, a designação de uma classe padrão, é normalmente feita particionando o espaço de características. A informação estrutural que descreve cada padrão é importante e o processo inclui, não somente a capacidade de associar o padrão a uma classe particular e classificá-lo, mas

também a capacidade de descrever aspectos do padrão que o tornam inelegível para designá-lo para outra classe.

Robert Schalkoff (1992, p. 364) diz que o reconhecimento de padrões estatísticos, como sugere o próprio nome, assume uma base estatística para os algoritmos de classificação. Um conjunto de medidas, que denotam as características, é extraído dos dados de entrada e usado para associar cada vetor de características a uma dentre *c* classes. Presume-se que as características são geradas por um estado da natureza e, portanto, existe um modelo subordinado a um estado com um conjunto de probabilidade e/ou funções de densidade de probabilidade correspondente, passível de ser usado para representá-lo.

Segundo Bishop (1985, p 492), a forma mais geral e natural de formular soluções para o reconhecimento de padrões é o reconhecimento de padrões estatístico, através do qual é reconhecida a natureza estatística tanto da informação que se quer representar quanto dos resultados que devem ser expressos.

- *Abordagens Sintática*

Muitas vezes, as informações significativas em um padrão não consistem apenas na presença ou ausência de valores numéricos de um conjunto de características. Ao invés disto, a inter-relação ou inter-conexão das características produzem informações estruturais importantes, que facilitam a descrição ou classificação. Esta é a base do reconhecimento de padrões sintático. Portanto é necessário, ao se usar a abordagem de

reconhecimento de padrões sintático, quantificar e extrair as informações estruturais importantes para poder avaliar a similaridade estrutural entre os padrões.

Para a abordagem sintática, ao enfatizar a descrição estrutural dos padrões, tenta-se fazer uma analogia entre as estruturas do padrão e a sintaxe de uma linguagem. Tipicamente, o reconhecimento de padrões sintático formula uma descrição hierárquica de padrões complexos, construída a partir de sub-padrões mais simples, sendo que no nível mais baixo se encontram os elementos mais simples, extraídos dos dados de entrada que são chamados de primitivas (FU; 1982 p.596). Por exemplo, as frases e as sentenças são construídas juntando-se palavras, e palavras são construídas juntando-se letras.

- *Abordagem Neural*

Com a capacidade intrínseca de aprender e o fato de serem aproximadores universais, as redes neurais surgiram como uma ferramenta poderosa na área de reconhecimento de padrões. Sua capacidade de fazer suposições mais precisas a respeito da distribuição dos dados de entrada do que métodos estatísticos tradicionais e a capacidade de formar fronteiras de decisão altamente não-lineares no espaço de características, levou ao seu crescente uso. Além disto, as redes neurais são usadas para a implementação de alguns modelos estatísticos, como a Análise de Componentes Principais (ACP)[12]

[12] ACP – Análise de Componentes Principais, também conhecida como transformação de Karthunen-Love, que busca uma representação de dimensão em vetores ortogonais, onde os vetores de entrada possam ser projetados sem perda de generalidade.

Uma rede neural, também conhecida como rede conexionista, é definida em (CICHOCKI; UNBEHAUEN, 1993), como um sistema de processamento de sinal ou informação, composto por um grande número de elementos simples de processamento, chamados neurônios artificiais, ou simplesmente nós, que são interconectados por elos diretos chamados de conexões, que colaboram para realizar um processamento paralelo distribuído (PPD) para realizar uma tarefa computacional desejada.

A tarefa computacional citada no parágrafo anterior pode ser, por exemplo, a classificação de padrões, ou seja: dado um vetor de características, deve-se associá-lo a uma classe em um determinado problema, por exemplo, reconhecimento de padrões. Na maioria das aplicações de reconhecimento de padrões usando-se redes neurais, são estabelecidas conexões entre os valores de todas as características que definem os padrões com camadas intermediárias de neurônios e a todas as classes objetivos que são consideradas como a camada de saída. O treinamento da rede é realizado corrigindo os pesos nas conexões para estabelecer-se as relações entre as características e classes que promovam a melhor discriminação possível entre os padrões de classes diferentes. Assim, ao ser apresentado à rede um novo padrão, esta indicará a classe que o representa melhor na camada de saída.

- *Abordagem Difusa*

A utilidade da teoria dos conjuntos difusos no reconhecimento de padrões já foi reconhecida em meados da década de 60, e hoje a literatura a respeito de reconhecimento de padrões difuso é bastante extensa, segundo Klir & Yuan (1995). Existem duas formas clássicas de reconhecimento de padrões difuso, são elas: os

métodos de reconhecimento de padrões através de listas de pertinência e os métodos difusos sintáticos.

No método clássico da lista de pertinência, cada classe é caracterizada por um conjunto de padrões que é armazenado no sistema de reconhecimento de padrões. Um padrão desconhecido a ser classificado é comparado com os padrões armazenados um a um. O padrão é classificado como um membro de uma classe se ele combinar com um dos padrões pertencentes a esta classe.

No método sintático difuso, um padrão é representado por uma cadeia de sub-padrões concatenados chamados de primitivas. Estas primitivas são vistas como o alfabeto de uma linguagem formal. Um padrão, então, é uma sentença gerada por alguma gramática. Todos os padrões, cujas sentenças são geradas pela mesma gramática, pertencem a uma mesma classe. Um padrão desconhecido é classificado como pertencente a uma classe particular se ele puder ser gerado pela gramática correspondente a ele (KLIR; YUAN, 1995).

Existem diversas formas de representação e um espaço que se julga necessário para discriminá-las. Matematicamente sempre é possível encontrar um conjunto de dimensão alta o suficiente onde os padrões (dados) são totalmente discriminados (separáveis linearmente), porém, este tipo de conjunto não interessa, principalmente devido às limitações computacionais. Procura-se o inverso, reduzir o conjunto de dados ao máximo possível e ainda manter o poder classificatório.

3.3 Conclusão

Conforme descrito nesta seção para utilização de um processo de mineração de dados, deve-se ter bem claro quais resultados se deseja chegar. A escolha da metodologia, na maioria dos casos, exige a participação de pessoas que efetivamente entendam do negócio em estudo, mesmo que não sejam especialistas na utilização e manuseio computacional dos dados. Uma vez definidas qual método será usado parte-se para identificar a melhor técnica, a mais aderente para obtenção dos resultados. Diversas técnicas podem ser utilizadas para se chegar aos resultados pretendidos, entretanto, cada técnica possui suas características, suas peculiaridades e precisa de pessoas que saibam interpretar seus resultados.

Estão incluídas no Anexo, uma aplicação das técnicas de Mineração de Dados dividida por áreas de concentração com alguns exemplos encontrados ao longo desse trabalho. Acompanha também uma tabela com alguns software que podem ser utilizados para a aplicação das técnicas de *Data Mining.*

O próximo capítulo trata da aplicação da abordagem *top-down ou direta*, que foi a escolhida para a condução do processo de imputação dos dados de fecundidade. Onde faremos uma comparação entre os métodos de Regressão Logística e Rede Neural.

4. Estudo de Caso

O presente capítulo tem o objetivo mostrar uma alternativa para o tratamento da não resposta, seguindo as etapas de aplicação da abordagem *top-down ou direta*, que foi a escolhida para a condução do processo de imputação dos dados de fecundidade. São elas: Identificar as fontes de dados, preparar os dados para análise, construir e treinar o modelo, avaliar o modelo.

Nossa proposta ao seguir este roteiro, é apresentar os resultados da regressão logística, da rede neural e fazer uma comparação nos resultados entre os modelos no percentual de erro no número de mulheres que foram classificadas como não tendo filhos.

4.1 Identificar as fontes dos dados

O IBGE tem uma vasta cultura na utilização de técnicas de imputação automática, seja no Censo Demográfico ou em outras pesquisas, demográficas ou não. No caso específico do Censo Demográfico de 1991 utilizou-se a imputação automática de variáveis para corrigir os casos de não-resposta ou inconsistências entre as variáveis, como por exemplo a variável renda. Os métodos usados, em geral, são baseados em imputações determinísticas, probabilísticas ou o chamado "*hot deck*".

No Censo Demográfico de 1991, a imputação automática foi usada em sua maior parte para as variáveis ditas categorizadas. Depois de vários estudos, decidiu –se considerar o código 99 para substituir a falta de informação em variáveis do bloco de fecundidade, e utilizar a distribuição conjunta com as variáveis grupos de idade qüinqüenais e estado conjugal para a imputação das variáveis de fecundidade, tendo em vista a forte correlação existente entre elas (OLIVEIRA et al, 1996).

Não podemos deixar de considerar, alguns problemas relacionados à aplicação dos métodos usados nas variáveis numéricas imputadas:

O processo de correção usado no Censo de 1991 foi o de padronizar as variáveis de fecundidade. Isto significa que se uma das variáveis estiver sem informação (ou inconsistente), todas as variáveis são transformadas para "ignorado" (OLIVEIRA et al, 1996), ocasionando assim:

- aumento da população de ignorados;
- perda de informações das variáveis eliminadas;

Resolveu-se substituir as variáveis numéricas por um valor determinístico de "ignorado" para um valor numérico de "99", acarretando assim:

- necessidade de uso de um filtro para processar corretamente o número de filhos (<99);

A necessidade de imputação das variáveis numéricas de fecundidade se baseia no fato de que a maior parte das não-respostas é ocasionada pela inferência (ou constrangimento) do recenseador em fazer a pergunta (principalmente para as mulheres mais jovens, do grupo de 10 a 15), assumindo que a mulher não deve ter filhos, acarretando assim que:

- a não-resposta é confundida com parturição zero (sem filhos);

As variáveis de fecundidade, são aquelas pertencentes ao questionário da Amostra (CD1.02), numeradas de 3.35 a 3.44[13].

Mais informações sobre a metodologia aplicada para a imputação das variáveis do Censo Demográfico de 1991 procurar no Manual de Apuração dos Dados Investigados pelo Questionário da Amostra – CD 1.02.

---

[13] Entende-se como variáveis *numéricas* (ou quantitativas) aquelas que dizem respeito à quantidade de filhos (3.37 a 3.42).

Optamos por trabalhar com o Estado do Rio de janeiro, pois estudos realizados na área de demografia, apontam que as tendências apresentadas no Rio de Janeiro, com o passar dos anos se apresentam também em outros estados da federação. E parte importante do que se deve estudar quando da implementação do modelo, seria a área para a modelagem (mínimo, máxima, ideal); assumindo-se que existam especificidades regionais.

Durante a fase de preparação da base de dados para desenvolver a aplicação desejada, aplicaram-se os filtros que estão no programa em anexo. Primeiramente, foram excluídos da base do Censo demográfico 1991 do Rio de janeiro todos os homens, o que corresponde a 48,23% da população total de 1316778 habitantes. Das 681718 mulheres, foram separadas as que apresentassem as variáveis Filhos Tidos que moram no domicílio, Filhas Tidas que moram no domicílio, Filhos Tidos que moram em outro domicílio, Filhas Tidas que moram em outro domicílio, Filhos Tidos Nascidos Vivos que já morreram e Filhas Tidas Nascidas Vivas que já morreram, todos iguais a 99, correspondendo a 21489 mulheres[14]. O restante das 660.229 mulheres apresentavam todos os dados do bloco de fecundidade completos, e são esses os dados que usaremos para a análise de interesse.

4.2 Preparando os dados para análise

Existem várias formas de analisar um conjunto de dados, seja com estatísticas descritivas, seja com ferramentas mais refinadas. Uma das mais utilizadas é a modelagem, que no nosso caso, nos permitirá relacionar através de um modelo, a

---

[14] Estas mulheres já passaram pela crítica específica do Censo para este quesito, e como já mencionado devem ter apresentado pelo menos uma dessas variáveis com dados omissos.

variável resposta (ter filhos ou não) com os fatores que mediam a ocorrência da existência de filhos.

A técnica utilizada mede a relação ou influência que algumas variáveis exercem sobre uma outra, denominada de interesse. Uma opção para tal avaliação é a classe de Modelos Lineares Generalizados (MLG).

A escolha das variáveis foi feita baseada nas variáveis usadas para imputação dos dados de fecundidade do Censo Demográfico de 1991 e também no próprio objetivo do trabalho que é buscar, além da literatura existente, novas variáveis para tornar o modelo o melhor possível, isto é, o que possa ter o melhor poder explicativo com o menor número de variáveis. Para isso foram selecionadas algumas variáveis do bloco de fecundidade, educação e da força de trabalho expostas na tabela abaixo, descritas na Documentação dos Microdados da Amostra.

Além destas foram criadas as variáveis TerFilhos, indicadora da existência ou não de filhos, e Ocupa, que é uma variável derivada das variáveis V0358 e V0345. Cabe ressaltar que a modelagem foi feita diretamente do arquivo de amostra do Rio de Janeiro.

Tabela 4.1

Variáveis selecionadas

Variáveis[15]	Descrição	valores possíveis
V0202	Localização	
V0309	raça ou cor	1 a 5 ou 9
V0310	religião	00,11 a 89 e 99
V0323	sabe ler/escrever	1, 2 ou b
V0330	vive/viveu cônjuge	1, 2 ou b
V0345	Trabalhou nos últimos 12 meses	1,2 ou 3
V0358	situação de ocupação	0 a 9 ou b
V1061	situação do domicilio	1 a 8
V3562	faixa de renda	
V3072	idade em anos (até a 3ª potência)	
V3241	anos de estudo (até a 3ª potência)	
V3311	idade contraiu 1ª união	
V3342	situação conjugal atual	1 a 5 ou b
TerFilhos	ter filhos ou não	1 ou 0

Fonte: IBGE/Censo Demográfico de 1991

Algumas variáveis acima, foram agrupadas para facilitar a análise dos dados, segundo as categorias usadas para o Censo91, a saber: religião ou culto (RELIGIAO), situação do domicílio (SITDOM), situação conjugal atual (SITUACASA), localização (LOCAL). Em anexo estão as categorias utilizadas para cada uma das variáveis.

A análise exploratória desta base de dados, entretanto revelou que a variável V3311, que informa a idade que contraiu a 1ª união, por exemplo, apresentava um valor alto de dados *missing* (18%), entre os que deveriam ter a informação. Para diminuir este percentual e podermos aproveitar um maior número de informações válidas, optou-se por imputar a idade atual menos 1 (um) como a data da primeira união, desde que a variável V0330 (Vive ou viveu com cônjuge) estivesse com resposta sim.

[15] Os nomes das variáveis seguem o padrão definido no dicionário de dados do Censo Demográfico de 1991 disponibilizado com o CD de microdados.

Das 660.229 mulheres que teriam informações no bloco de fecundidade, foram separadas as mulheres que tivessem mas de 10 anos. Ficamos com um total de 536749 mulheres com mais de 10 anos e os dados completos.

4.3 Construir e treinar o modelo

A seleção de modelos é uma parte importante de toda a pesquisa, pois envolve a procura de um modelo, o mais simples possível, que descreva bem os dados observados.

Para alcançar o formato final dos dados foi realizado um estudo de cada variável utilizando-se além do questionário do Censo com o dicionário de dados que fazem parte dos microdados, alguns estudos auxiliares como os produzidos nas tabelas abaixo para verificar a distribuição desses dados.

Optou-se por separar as mulheres de 10 anos e mais pelo estado civil , separando por grupo de idade e classificando pela variável TerFilhos (Tabela 42).

Tabela 4.2

Distribuiçao das mulheres por grupo de idade e situação conjugal, segundo a condição de ter filhos ou não

Mulheres que Não teriam Filhos				
Grupo de idade	Situação Conjugal atual			
	Casada	Separada/ viúva	Solteira	Total
10|——15	135	33	56333	56501
15|——18	1167	115	29826	31108
18|——20	1702	146	16063	17911
20|——25	6166	524	26175	32865
25|——40	11116	1795	24791	37702
40|——60	5072	1828	8331	15231
60|——	2266	2891	4368	9525
Total	27624	7332	165891	200843

Mulheres que Teriam Filhos				
Grupo de idade	Situação Conjugal atual			
	Casada	Separada/ viúva	Solteira	Total
10|——15	90	20	57	167
15|——18	1309	182	505	1996
18|——20	3183	469	858	4510
20|——25	18459	2567	2947	23973
25|——40	105895	17258	6865	130018
40|——60	82945	30200	3629	116774
60|——	22743	33824	1901	58468
Total	234624	84520	16762	335906

Fonte: Microdados IBGE/Censo Demográfico de 1991. Estimativas da autora

As variáveis v0309 (Raça), v0323(Saber ler/escrever) e V0202(Localização) também foram estudadas, e verificou-se que a variável v0202 precisaria ter uma nova classificações (LOCAL), pois os dados apresentavam faixas com freqüências muito baixas, sem expressividade no conjunto e acarretando problemas no modelo.

Tabela 4.3

Distribuiçao das mulheres por raça, segundo a condição de ter filhos ou não

Raça	Existência de filhos		
	Não ter	Ter	Total
Branca	111198	190923	302121
Preta	20663	35356	56019
Amarela	251	432	683
Parda	67658	107633	175291
Indígena	111	321	432
Ignorada	962	1241	2203
Total	200843	335906	536749

Fonte: Microdados IBGE/Censo Demográfico de 1991. Estimativas da autora

Optou-se por conservar a classificação original para as variáveis v0309 e v0323, pois sua distribuição verificada nas tabelas 4.3 e 4.4 respectivamente não apresentaram evidências de problema.

Tabela 4.4

Distribuiçao das mulheres que sabem ler e escrever por grupo de idade, segundo a condição de ter filhos ou não

Mulheres que Não teriam Filhos			
Grupo de idade	Alfabetização		
	Saber ler	Não saber	Total
10|-----15	53714	2787	56501
15|-----18	30189	919	31108
18|-----20	17466	445	17911
20|-----25	31994	871	32865
25|-----40	36207	1495	37702
40|-----60	13622	1609	15231
60|------	7726	1799	9525
Total	190918	9925	200843

Mulheres que Teriam Filhos			
Grupo de idade	Alfabetização		
	Saber ler	Não saber	Total
10|-----15	144	23	167
15|-----18	1830	166	1996
18|-----20	4198	312	4510
20|-----25	22393	1580	23973
25|-----40	121447	8571	130018
40|-----60	99100	17674	116774
60|------	42484	15984	58468
Total	291596	44310	335906

Fonte: Microdados IBGE/Censo Demográfico de 1991. Estimativas da autora

Este trabalho procura fazer uma comparação entre o resultado de redes neurais com método de *backpropagation* (retroprogramação) e o resultado de classificação da regressão logística usando o método *none* (método com todas as variáveis).

A seguir definiremos os modelos utilizados.

Usando Regressão Logística

Nesta dissertação utiliza-se a metodologia de modelos lineares generalizados[16] nos quais as variáveis respostas são binárias. A variável resposta é a ocorrência ou não de se ter filhos. Define-se uma variável aleatória binária da forma:

Y = 1 se a resposta é sucesso
Y = 0 se a resposta é fracasso

com $P(Y=1) = p$ e $P(Y=0)= 1-p$

Sejam $Y_1,....Y_n$, n variáveis aleatórias independentes com $P(Y_i=1) = p_i$ (i = 1,... n). Então, a distribuição de probabilidade conjunta é dada por:

$$P(Y_1,....Y_n) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} = \exp\left[\sum_{i=1}^{n} y_i \log\left(\frac{p_i}{1-p_i}\right) + \sum_{i=1}^{n}\log(1-p_i)\right]$$

O modelo linear generalizado (MLG) é definido por 3 elementos: uma distribuição de probabilidade, obrigatoriamente membro da família exponencial de distribuições, para a variável resposta; um conjunto de variáveis independentes descrevendo a estrutura linear do modelo; e uma função de ligação entre a média da variável resposta e a estrutura linear (por exemplo, logarítmica para os modelos log-lineares). O modelo de regressão logístico que é um caso especial dos MLG, pode ser utilizado quando a variável resposta é binária.

---

[16] Para uma definição mais detalhada consultar (Dobson,1983)

Quando a variável resposta é binária, tomando os valores 1 e 0, com probabilidades p_i e (1 - p_i) respectivamente, Y é uma variável *Bernoulli* com parâmetro $E(Y_i) = p_i$. A relação entre o parâmetro de distribuição e as variáveis explicativas do modelo logístico na sua forma usual é dado por:

$$g(p_i) = \beta_0 + \beta_1 X_i$$

Considerações teóricas e práticas sugerem que quando a variável resposta é binária, a forma da função de ligação deve ser uma função f definida no intervalo [0,1], espaço de valores possíveis para o parâmetro p_i, com imagem em (+∞, -∞), espaço de valores possíveis para o resultado de $\beta_0 + \beta_1 X_i$.

Existem várias funções com estas características, por exemplo a logito:

$$\textbf{Logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki}$$

Esta função é chamada de *transformação* Logito da probabilidade *p*. A razão *p* / (1 - *p*) é chamada de Chance (Odds). A função resposta transformada é denominada como função resposta *logit*. Outra função de resposta curvilínea com a mesma forma da função logística é obtida transformando *p* por meio da inversa da distribuição normal acumulada.

$probit(p_i) = \phi^{-1}(p_i)$ onde:

ϕ = função de distribuição normal acumulada da Normal (0,1) e ϕ^{-1} é a função inversa de ϕ. Esta transformação é chamada de transformação probito. Outra função de resposta curvilínea é a transformação complemento log log da probabilidade π dada por:

$$\ln [- \ln (1 - p_i)]$$

Diferentemente das transformações logito e probito esta transformação não é simétrica em torno de $\pi = 0,5$. Existem também outras funções como o arctag, etc., mas estaremos trabalhando com as 3 primeiras já implementadas no SAS.

Nosso interesse é verificar se um subconjunto das variáveis explicativas podem ser retiradas do modelo de regressão logística, isto é, vamos testar se os coeficientes de regressão β_k são iguais a zero. Para isso ajustaremos modelos de regressão logística aos dados, usando a estatística de teste *Deviance*, baseada na razão de verossimilhança, isto é, que compara o logaritmo da verossimilhança deste modelo com o logaritmo da verossimilhança do modelo completo, e a estatística de *Wald*, com a finalidade de obtermos o modelo mais adequado para avaliar os fatores que estão associados à decisão da escolha de ter filhos ou não.

No caso específico da dissertação, foram testadas as três funções de ligação. As variáveis usadas como explicativas para a aplicação do modelo de regressão logística foram: Localização, raça, religião, saber ler/escrever, vive/viveu com cônjuge, ocupação, situação do domicílio, faixa de renda per capta, idade em anos (também idade ao quadrado e ao cubo), idade que contraiu a 1ª união, situação conjugal atual, anos de estudo (também ao quadrado e ao cubo). A variável dependente é Ter filhos.

Os resultados das três funções conforme a tabela abaixo, foram bastante similares, por isto optou-se pelo modelo que apresentasse a menor desviância observada (DOBSON, 1983), no caso, o modelo logito.

Tabela 4.5

Análise da Desviância

Modelo de Regressão Logística	Desviância
LOGIT	**150176.82**
PROBIT	150207.54
CLOGLOG	156471.23

Fonte: IBGE/Censo Demográfico 1991. Estimativas da autora.

Escolhida a função de ligação, o métodos de otimização para o estimador de máxima verossimilhança usado neste caso, foi o de *Quasi-Newton* seguindo os critérios apresentados a seguir:

- **Parâmetros do modelo acima de 40 e número máximo de iterações 50** – *Trust-Region, Newton-Raphson- Ridging*, e *Newton-Raphson -Line Search* são usados quando o Hessiano é fácil e rápido de calcular. Às vezes Newton-Raphson-Ridging pode ser mais rápido que *Trust-Region*, mas *Trust-Region* é numericamente mais estável. Se o Hessiano não for fácil de calcular, então o método de *Newton-Raphson-Line* pode ser um método muito competitivo.
- **Parâmetros do modelo até 400 e número máximo de iterações 200** – Os métodos de *Quasi-Newton* e *Double-Dogleg* são usados para problemas de otimização onde a função objetiva e o gradiente são muito mais rápidos de calcular que o Hessiano. O *Quasi-Newton* e o *Double-Dogleg* requerem mais iterações que

o método de Trust-Region ou os métodos de Newton-Raphson mas cada iteração é muito mais rápida.

- **Parâmetros do modelo acima de 400 e número máximo de iterações 400** - O método de *Conjugate Gradient* é usado também onde a função objetiva e o gradiente são muito mais rápidos de calcular do que o Hessiano, e precisam de muita memória para armazenar a matriz de Hessiano aproximada.

Uma regressão logística foi executada com todas as variáveis simples, para determinar sua significância para o modelo.

Pela estatística da deviance , a variável SITDOM (situação do domicílio) foi rejeitada. Rodou-se outra regressão com todas as interações mais significativas e verificou-se quais variáveis e interações poderiam ser usadas para tentar melhorar o modelos, potencializando as variáveis que melhor se apresentam no modelo original.

A variável SITDOM retornou ao modelo, pois a interação desta com a variável v3562 (renda per capta) se mostrou significativa na nova regressão com o uso das interações. O resultado obtido neste estudo está na tabela abaixo:

Tabela 4.6

Resultado da seleção de variáveis para o modelo com interações

Variável	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	283707067			
V0309	283629145	5	77.92	<.0001
V0323	279070008	1	4559.14	<.0001
V0330	134217168	1	144853	<.0001
V3342	133897789	3	319.38	<.0001
RELIGIAO	133850271	8	47.52	<.0001
SITDOM	133843286	1	6.99	0.0082
Ocupa	126390691	6	7452.59	<.0001
LOCAL	126275037	3	115.65	<.0001
V3072	121824469	1	445057	<.0001
V3241	120778912	1	1045.56	<.0001
V3311	120629193	1	149.72	<.0001
V3562	120499775	1	129.42	<.0001
V307_UYW	116100743	1	4399.03	<.0001
V307_KZW	112571332	1	3529.41	<.0001
V324_NBA	112301905	1	269.43	<.0001
V324_UY4	112105729	1	196.18	<.0001
V3562*LOCAL	112087359	3	18.37	0.0004
V3241*Ocupa	111875625	6	211.73	<.0001
V3072*Ocupa	111641284	6	234.34	<.0001
V3562*Ocupa	111625703	6	15.58	0.0162
V3562*SITDOM	111603910	1	21.79	<.0001
V3072*V0309	111479316	5	124.59	<.0001
V3072*V0323	111419645	1	59.67	<.0001
V3562*V0330	111178022	1	241.62	<.0001
V3241*V3342	110986719	4	191.30	<.0001
V3072*V3311	110981942	1	4.78	0.0288
V3072*V3342	110799241	4	182.70	<.0001
V3072*V3241	110728988	1	70.25	<.0001
V3311*V3342	105937025	3	4791.96	<.0001

Fonte: IBGE/Censo Demográfico de 1991. Estimativas da autora

Para a aplicação do modelo, dividiu-se o arquivo de dados completos em três arquivos:

1 - arquivo de treinamento (40%), que é usado para ajustar o modelo preliminar

(tentativa de achar os melhores pesos para esse conjunto de variáveis);

2 - arquivo de validação (30%), que é usado para monitorar e afinar os pesos do modelo durante a estimação e avaliação do modelo; e

3 - arquivo de teste (30%), que é uma segurança adicional que pode ser usado para obter uma estimativa imparcial do erro de generalização do modelo.

Os resultados da análise do modelo de regressão listados na Tabela 4.7 foram obtidos usando o Enterprise Miner do SAS.

No modelo de regressão estudado, foram incluídas somente termos que expressam uma relação entre as variáveis explicativas e a variável dependente. No entanto, em algumas situações, a relação entre X_1 e Y varia em função de diferentes valores de X_2.

No nosso caso, 13 termos de interação, envolvendo o produto das variáveis explicativas, foram incluídos.

Usando a função Logito e os resultados da regressão logística com todas as variáveis significativas, o modelo de interação em questão pode ser formulado como:

Logit(prob.TerFilhos = -6.8266+0.1844(local_1)+0.3315(local_2)+0.5526(local_3) - 1.0706(ocupa_0)-0.4278(ocupa_1)+ 1.0378(ocupa_2) + 0.4548(ocupa_3) + 2.4379(ocupa_4) - 3.6844(ocupa_5) + 0.1414(religião_0) + 0.00349(religião_1) + 0.0514(religião_2) + 0.0425(religião_3) – 0.0652(religião_4) + 0.1009(religião_5) - 0.0641(religião_6) - 0.1653(religião_7) - 0.1317(sitdom_1) - 0.0877(v0309_1) + 0.4583(v0309_2) - 0.8412(v0309_3) + 0.0785(v0309_4) + 0.4267(v0309_5) + 0.2842(v0323_1) + 0.1674(v0330_1) + 0.5833(v3072) - 0.0117(v3072**2) +

0.000066(v3072**3) - 0.0189(v3241) + 0.000502(v3241**3) - 0.0166(v3241**2) - 0.0107(v3311) + 0.5060(v3342_1) + 2.1869(v3342_2) - 0.1972(v3342_3) - 0.00025(v3072*ocupa_0) + 0.00823(v3072*ocupa_1) - 0.00268(v3072*ocupa_2) - 0.00367(v3072*ocupa_3) - 0.0379(v3072*ocupa_4) + 0.0482(v3072*ocupa_5) + 0.0513(v3241*ocupa_0) - 0.00154(v3241*ocupa_1) - 0.0268(v3241*ocupa_2) - 0.00786(v3241*ocupa_3) - 0.0826(v3241*ocupa_4) + 0.0506(v3241*ocupa_5) + 0.000099(v3072*v0309_1) - 0.0105(v3072*v0309_2) + 0.0161(v3072*v0309_3) - 0.00052(v3072*v0309_4) + 0.00532(v3072*v0309_5) - 0.00849(v3072*v0323_1) + 0.00171(v3072*v3241) + 0.000132(v3072*v3311) + 0.1040(v3072*v3342_1) + 0.0449(v3072*v3342_2) + 0.0344(v3072*v3342_3) + 0.0434(v3072*v3342_4) + 0.0470(v3241*v3342_1) + 0.0123(v3241*v3342_2) - 0.00418(v3241*v3342_3) - 0.0195(v3241*v3342_4) - 0.1527(v3311*v3342_1) - 0.1204(v3311*v3342_2) + 0.00292(v3311*v3342_3) - 0.0173(v3562) - 0.0170(v3562*local_1) -0.0127(v3562*local_2) - 0.0262(v3562*local_3) + 0.0333(v3562*ocupa_0) - 0.0119(v3562*ocupa_1) - 0.0574(v3562*ocupa_2) - 0.0252(v3562*ocupa_3) + 0.0126(v3562*ocupa_4) + 0.0896(v3562*ocupa_5) + 0.0162(v3562*sitdom) + 0.0265(v3562*v0330)

O critério para seleção do melhor modelo foi o de Akaike (AIC), este critério penaliza o modelo pelo acréscimo de parâmetros. Sendo assim, o modelo com valor de AIC menor é o escolhido.

Usando Rede Neural

Entender o funcionamento do cérebro humano sempre foi um desafio à comunidade científica. Sua agilidade e eficiência constituem o objetivo a ser alcançado por sistemas da Inteligência Artificial (IA), que baseia seus estudos na simulação computacional de aspectos da inteligência humana.

Os sistemas conexionistas, emergiram como um ramo da IA, cujo propósito primordial era entender o funcionamento do cérebro humano e assim desenvolver uma estrutura artificial nos mesmos moldes, com capacidade de processamento, habilidade e credibilidade semelhantes à da estrutura cerebral biológica.

Definição de Redes Neurais Artificiais

Redes Neurais Artificiais ou RNAs são sistemas paralelos distribuídos compostos por unidades de processamento simples (nodos), que calculam determinadas funções matemáticas (normalmente não-lineares). Tais unidades são dispostas em uma ou mais camadas e interligadas por um grande número de conexões, geralmente unidirecionais, sendo inspiradas no funcionamento do cérebro,. Na maioria dos modelos estas conexões estão associadas a pesos, os quais armazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida por cada neurônio da rede (BRAGA; LUDEMIR; CARVALHO, 2000). Portanto, uma RNA pode ser definida como uma estrutura computacional que tem como objetivo permitir a implementação de modelos matemáticos que representem a forma como o cérebro humano processa as informações que adquire.

Neurônios Biológicos

Os neurônios são divididos em três seções: o *corpo da célula*, os *dendritos* e o *axônio*, cada um com funções específicas, porém complementares. Os dendritos têm por função

receber as informações, ou *impulsos nervosos*, oriundos de outros neurônios e conduzi-las até o corpo celular, onde a informação é processada, e novos impulsos são gerados.

Estes impulsos são transmitidos a outros neurônios, passando através do axônio até os dendritos dos neurônios seguintes. O ponto de contato entre a terminação axônica de um neurônio e o dendrito de outro é chamado de *sinapse*.

É pelas sinapses que os nodos se unem funcionalmente, formando redes neurais.

A figura 4.1 a seguir ilustra, os componentes do neurônio.

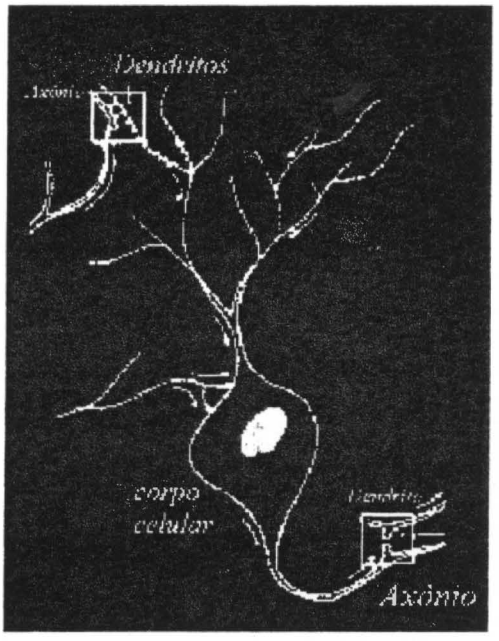


Figura 4.1: Componentes do neurônio biológico
(MACHADO, 1980).

As sinapses funcionam como válvulas, e são capazes de controlar a transmissão de impulsos, isto é, o fluxo de informações, entre os nodos na rede neural. O efeito das sinapses é variável, e é esta variação que da ao neurônio capacidade de adaptação.

Os sinais oriundos dos neurônios pré-sinápticos são passados para o corpo do neurônio pós-sinaptico, onde são comparados com os outros sinais recebidos. Se o percentual em um intervalo curto de tempo é suficientemente alto, a célula "dispara", produzindo um impulso que é transmitido para as células seguintes (nodos pós-sinápticos).

Neurônio artificial: modelo MCP

O modelo de neurônio proposto por McCulloch e Pitts é uma simplificação do que se sabia então a respeito do neurônio biológico. Sua descrição matemática resultou em um modelo com *n* terminais de entrada $x_1,x_2,....x_n$ (que representam os dendritos) e apenas um terminal de saída *y* (representando o axônio). Para emular o comportamento das sinapses, os terminais de entrada do neurônio têm pesos acoplados $w_1,w_2,...,w_n$ cujos valores podem ser positivos ou negativos, dependendo de as sinapse correspondentes serem inibitórias ou excitatórias.

O efeito de uma sinapse particular i no neurônio pós-sináptico é dado por x_iw_i. Os pesos determinam "em que grau" o neurônio deve considerar sinais de disparos que ocorrem naquela conexão.

Uma descrição do modelo está ilustrada na Figura 4.2

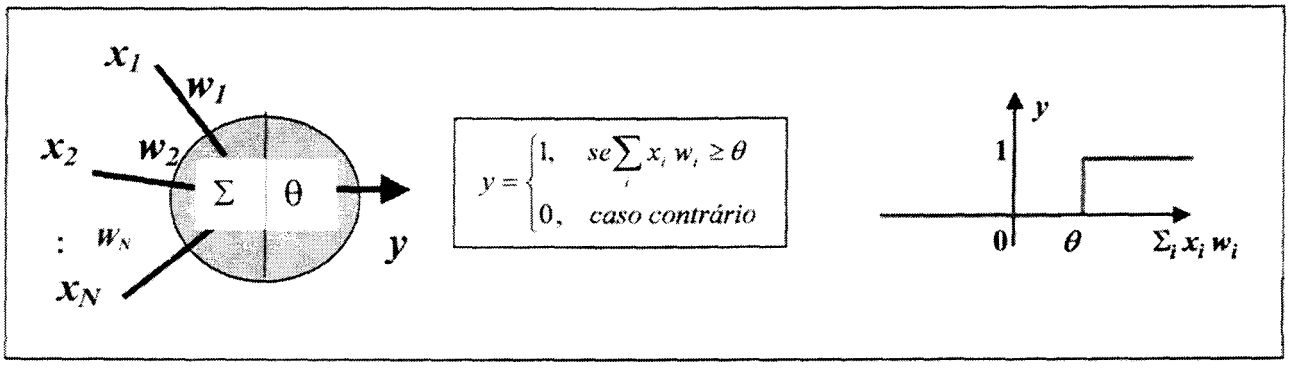


Figura 4.2: Neurônio de McCulloch e Pitts.- MCP

Um neurônio biológico dispara quando a soma dos impulsos que ele recebe ultrapassa o seu limiar de excitação (θ) *threshold*. O corpo do neurônio, por sua vez, é emulado por um mecanismo simples que faz a soma dos valores $x_i w_i$ recebidos pelo neurônio (soma ponderada de x_i *com os pesos* w_i) e decide se o neurônio deve ou não disparar (saída igual a 1 ou a 0) comparando a soma obtida ao limiar ou *threshold* do neurônio. Já no modelo MCP, a ativação do neurônio é obtida através da aplicação de uma "função de ativação", que ativa ou não a saída , dependendo do valor da soma ponderada das suas entradas. Na descrição original do modelo de MCP, a função de ativação é dada pela função de limiar descrita na Equação a seguir:

$$\sum_{i=1}^{n} x_i w_i \geq \theta$$

onde n é o número de entradas do neurônio, w_i é o peso associado à entrada x_i e θ é o limiar (*threshold*) do neurônio.

O parâmetro externo θ, definido como "*threshold*" ou limiar, normalmente é utilizado como um valor abaixo do qual a saída do neurônio é nula. Ele pode ser definido como um peso referente a uma entrada de valor constante igual a 1, denominada *bias*[17].

Pode-se encontrar também, a denominação de "função de transferência" para o que até agora foi chamado de função de ativação. Entretanto, há autores que definem a existência de duas funções no neurônio artificial, sendo a soma ponderada das entradas (v_j) denominada de função de ativação e φ (v_j) sendo a função de transferência (NELSON & ILLINGWORTH, 1990).

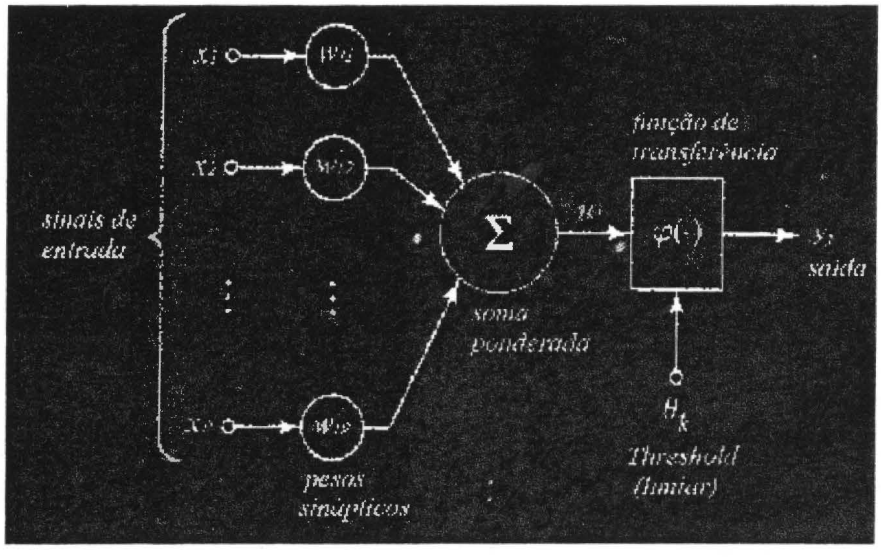


Figura 4.3: Modelo de neurônio artificial. No caso do modelo de Mcculloch e Pitts, φ (.) é uma função de saída binária, (TODESCO, 1995).

A partir do modelo de neurônio de Mcculloch e Pitts, um modelo generalizado, utilizando outras opções de função de ativação, tornou-se padrão no desenvolvimento de RNAs. Este modelo generalizado permite que a resposta do neurônio varie de modo contínuo. A

[17] Indica a posição da curva ao longo do eixo horizontal

função de ativação limita a amplitude da saída do neurônio, geralmente no intervalo normalizado de [0,1], e em casos específicos [-1,1]. A seguir são dados alguns exemplos de funções de ativação comumente utilizadas por neurônios artificiais:

Funções de ativação

A partir do modelo de McCulloch e Pitts, foram derivados vários outros modelos que permitem a produção de uma saída qualquer, não necessariamente zero ou um, e com diferentes funções de ativação.

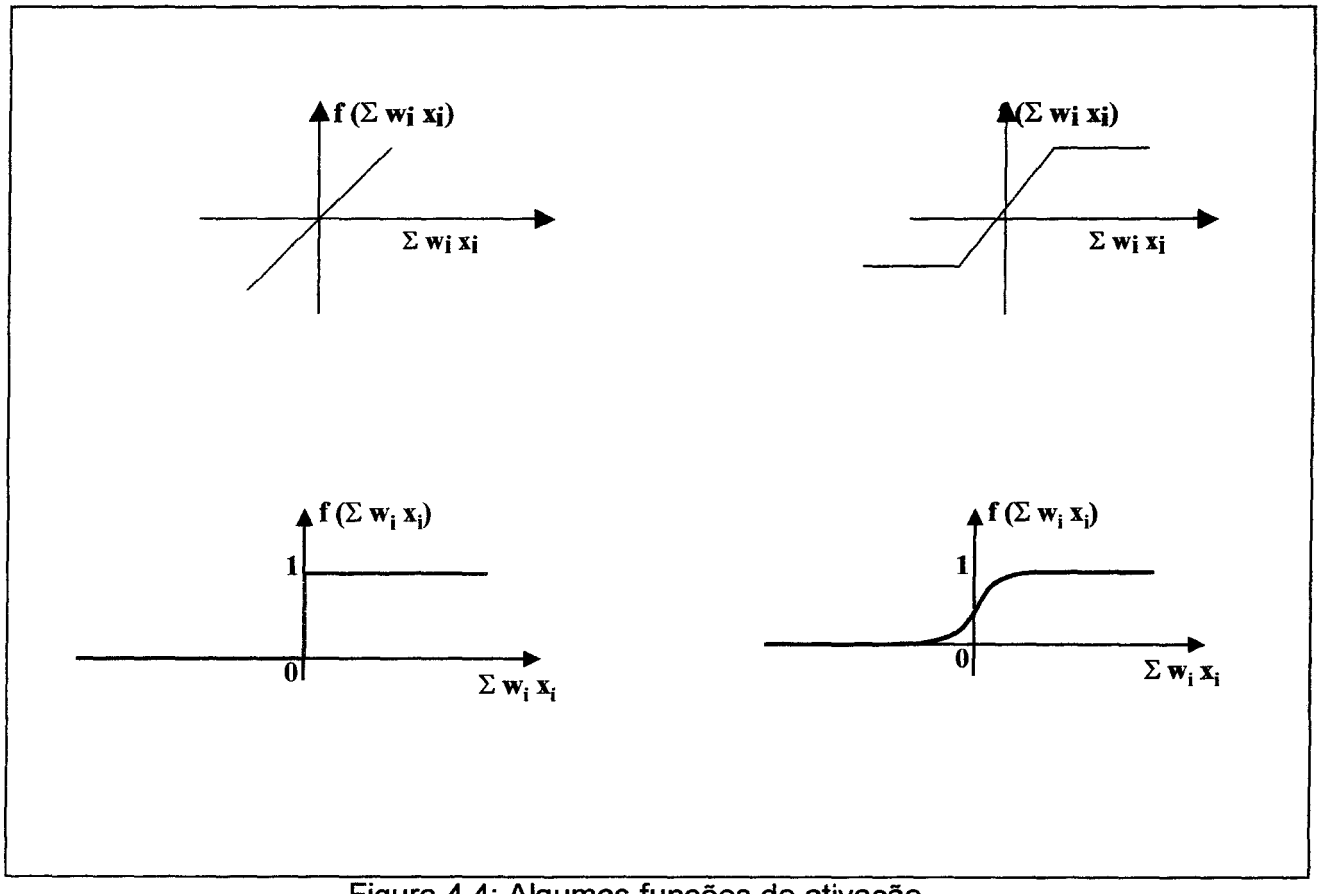


Figura 4.4: Algumas funções de ativação.

A figura 4.4 mostra quatro funções de ativação diferentes: a função linear, a função rampa, a função degrau (*step*) e a função sigmoidal.

A Figura4.4 (a) é definida pela equação $y = \alpha x$, onde α é um número real que define a saída linear para os valores de entrada, y é a saída e x é a entrada.

A função linear pode ser restringida para produzir valores constantes fera da faixa [$-\gamma$, $+\gamma$], e neste caso a função passa a ser a função rampa (Figura4.4b) e a equação :

$$Y = \begin{cases} +\gamma & \text{iff } X \geq +\gamma \\ X & \text{iff } |X| < +\gamma \\ -\gamma & \text{iff } X \leq -\gamma \end{cases}$$

A função passo, Figura 4.4c, é similar a função sinal, (+ para números positivos e – para números negativos), no sentido de que a função produz a saída $+\gamma$ para os valores de X maiores que zero , caso contrário a função produz o valor $-\gamma$. A função degrau é definida pela equação:

$$Y = \begin{cases} +\gamma & \text{iff } X > 0 \\ -\gamma & \text{iff } X \leq 0 \end{cases}$$

A função sigmoidal (na verdade, uma família de funções com estas características), conhecida também como S-shape, ilustrada na Figura 4.4d, é uma função semilinear, limitada e monotônica.

Uma das funções sigmoidais mais importantes é a função logística definida pela equação:

$$Y = \frac{1}{1 + e^{-x/T}}$$

onde o parâmetro *T* determina a suavidade da curva. Exemplo destas funções são encontradas na descrição da função de ligação do MLG.

Aprendizado

Assim como uma rede neural biológica desenvolvida deve possuir a capacidade de armazenar novos conhecimentos com o intuito de torná-los úteis à tomadas de decisões, uma RNA deve ser treinada com o mesmo objetivo: aprender.

Segundo HAYKIN (1994), os pesos sinápticos designados às conexões entre os neurônios artificiais, são os responsáveis pelo armazenamento do conhecimento nas RNAs. São eles os parâmetros que devem ser ajustados através do processo de "treinamento", para que a rede habilite-se a responder o mais corretamente possível a quaisquer outros estímulos que lhe forem apresentados em uma fase posterior ao treinamento, denominada de "teste".

Esta capacidade que a RNA tem de classificar os dados de teste, é chamada de capacidade de "generalização". Fica subentendido, portanto, que a generalização da rede deve ser significativa, após seus pesos serem ajustados na fase de treinamento.

Três formas de aprendizado podem ser definidas:

- Supervisionada,
- Não Supervisionada e

- Híbrida.

Para que os valores dos pesos possam ser ajustados, eles devem primeiro ser "inicializados". Entretanto, a forma de designar os pesos iniciais de uma RNA, é muito discutida. Alguns autores defendem a inicialização através de valores randômicos dentro do intervalo de valores das entradas de treinamento (HAYKIN, 1994), outros defendem a tomada das primeiras entradas do conjunto de treinamento como os pesos iniciais da rede (BISHOP, 1995). A inicialização também pode ser feita tomando-se os primeiros elementos de cada classe, disponíveis no conjunto de treinamento, como os primeiros pesos da rede. Entretanto, a escolha apropriada da forma de inicialização dos pesos depende, entre muitos outros fatores, do objetivo para o qual a rede está sendo projetada e da forma de aprendizado utilizada.

Aprendizado Supervisionado

O aprendizado supervisionado das RNAs caracteriza-se pela disponibilidade de um conjunto de treinamento composto por dados de entrada previamente classificados. Ou seja, estímulos de entrada e respostas desejadas à estes estímulos.

O processo de ajuste dos pesos dá-se da seguinte forma: os estímulos de entrada, disponíveis no conjunto de treinamento, são apresentados à rede, que calcula uma resposta, utilizando como parâmetros os valores de pesos atuais, ou seja, sem ajuste. A partir daí faz-se uma comparação entre a resposta oferecida pela rede atual e a resposta desejada àqueles estímulos. Com base na similaridade entre as duas respostas, os pesos são ajustados. Esse procedimento ocorre até que os pesos possibilitem à rede classificar o mais corretamente possível as entradas de treinamento, e conseqüentemente as entradas apresentadas durante a fase de teste.

Geralmente estas regras de ajuste supervisionado, baseiam-se na determinação do erro, calculado como a diferença entre a resposta que a rede oferece, utilizando os pesos atuais, e a resposta que deveria oferecer ou resposta desejada, disponível no conjunto de treinamento.

A soma dos erros quadráticos de todas as saídas é normalmente utilizada como medida de desempenho da rede e também como função de custo a ser minimizada pelo algoritmo de treinamento.

Um fator importante no processo de aprendizado, é definir se a rede será treinada "por padrões" ou "por grupos" (*batch*) (HAYKIN, 1994). No primeiro modo os pesos são ajustados a medida em que os padrões vão sendo apresentados à rede. Assim, em cada apresentação do conjunto de treinamento, tantas alterações de pesos ocorrerão quantos forem os elementos deste conjunto.

No modo "por grupos" ou "*batch*", os pesos, normalmente, são ajustados após todos os elementos do conjunto de treinamento serem apresentados à rede, ou seja, apenas um cálculo ocorre ao final de cada *época* (intervalo no qual todo o conjunto de treinamento é apresentado à RNA). No aprendizado "por grupos", há a possibilidade da apresentação de grupos de padrões à rede, desta forma, em uma época ocorrerão tantos ajustes de pesos quantos forem os grupos apresentados.

Aprendizado Não Supervisionado

O aprendizado não supervisionado permite à rede aprender sem que haja um conjunto de "respostas desejadas" como referência para as saídas da rede. Neste caso, os pesos são ajustados a medida que a rede vai sendo alimentada de padrões de entrada selecionados como representativos de cada classe. Neste caso, os pesos das conexões, são ajustados de acordo com sua similaridade aos padrões de entrada apresentados à rede. Normalmente, a distância Euclidiana entre ambos é calculada, e o neurônio representando o peso mais próximo do vetor de entrada é considerado como o vencedor.

Aprendizado Híbrido

A forma de aprendizado híbrida ou por reforço, é uma variante do aprendizado supervisionado, no qual não se dispõe de respostas corretas, mas pode-se saber se as respostas que a rede produziu são corretas ou não. Pode-se dizer, neste treinamento que existe um "crítico" que verifica se a resposta gerada pela rede é satisfatória, em caso afirmativo as conexões são reforçadas, em caso contrário estas conexões devem ter menor peso (ROITMAN, 2001).

Métodos dos mínimos quadrados também são bastante utilizados no ajuste de pesos de redes neurais. Nestes casos, a soma dos erros quadrados, E, deve ser minimizada (BISHOP,1995).

A regra Delta ou Least Mean Squares, ou ainda regra de Widrow e Hoff (HUSH & HORN, 1993) é também um método de minimização do erro quadrado muito difundido.

A Figura 4.5, ilustra o processo de aprendizado aplicando a regra Delta, e o resumo de seu algoritmo é descrito em seguida:

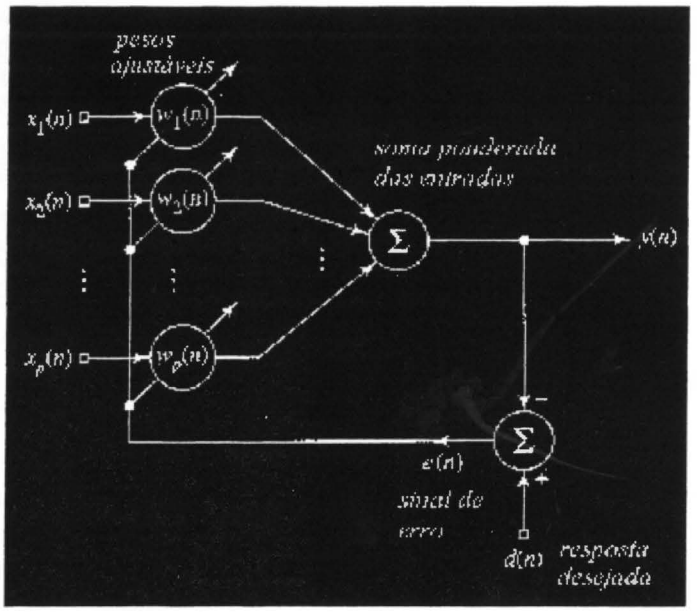


Figura 4.5 Processo de aprendizado ocorrido pela aplicação da Regra Delta. O erro é utilizado no cálculo do ajuste de pesos (HAYKIN, 1994).

Segundo HAYKIN (1994), a regra Delta só pode ser aplicada sobre um modelo de neurônio linear simples. Portanto, redes não lineares, têm que ser treinadas por processos mais elaborados, tais como a regra Delta generalizada aplicada pelo algoritmo de aprendizado Backpropagation (SKAPURA, 1996).

Arquiteturas de RNAS utilizadas em problemas de classificação de dados

Normalmente, as arquiteturas de redes se classificam em função do número de camadas de neurônios e de sua estruturação, além da forma como os sinais calculados são propagados (*feedforward* ou recorrentes) (PACHECO, 1996).

As redes que permitem uma retroalimentação são chamadas de redes recorrentes. Neste caso, o sinal de saída da camada de neurônios pode ser utilizado como uma nova entrada para os mesmos, até que um equilíbrio seja atingido.

Muitas arquiteturas de RNAs foram desenvolvidas para solucionar problemas de classificação de dados. Segundo SCHALKOFF (1992), o reconhecimento de padrões, pode ser visto como o particionamento dos dados, representados por pontos em um espaço multidimensional, em regiões indicativas das classes.

Entre as redes com apenas uma camada de neurônios, o *Perceptron,* desenvolvido por Rosenblatt em 1958, pode ser citado como um exemplo amplamente difundido (HAYKIN, 1994). É definida como uma rede linear e, portanto, classifica apenas padrões que sejam linearmente separáveis.

O *Perceptron* consiste basicamente de um neurônio recebendo várias entradas através de conexões com pesos ajustáveis. Normalmente, tais pesos são ajustados através de um processo de aprendizado supervisionado.

Além de possuir apenas uma camada de neurônios, o *Perceptron* tem a característica de ser *feed-forward*, ou seja, seus sinais calculados apenas passam adiante na rede e não retornam nunca.

Arquiteturas constituídas de várias camadas, também são bastante utilizadas em processos de classificação. O *Perceptron* Multicamadas (SKAPURA, 1996), é uma rede *feed-forward* que se classifica nesta categoria.

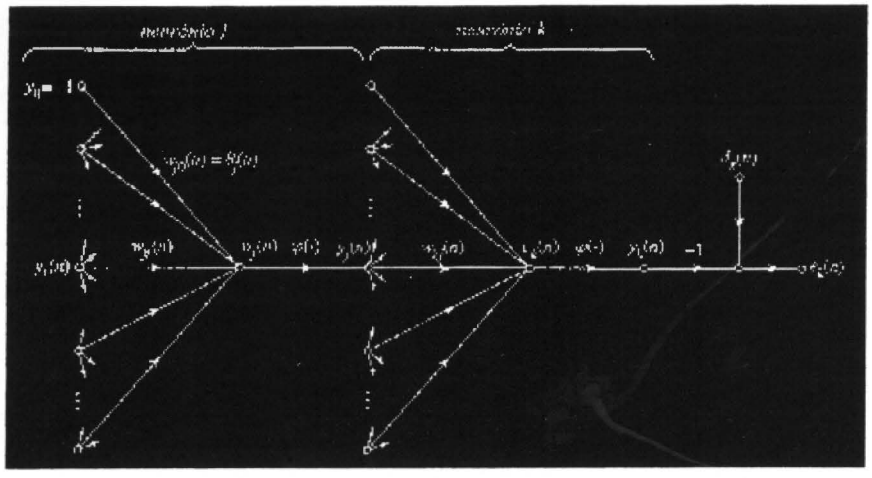


Figura 4.6: Camada de saída (*k*) conectada à segunda camada escondida (*j*) em um *Perceptron* Multicamadas (HAYKIN, 1994).

Normalmente, o *Perceptron* de Rosenblatt, possui uma arquitetura constituída por duas camadas de neurônios escondidos (com saídas não observáveis) e uma camada de neurônios de saída, e é normalmente treinada com o algoritmo *Back-propagation*.

O *Perceptron* de múltiplas camadas, é utilizado para a classificação de padrões não linearmente separáveis. Ou seja, padrões que caem em determinadas posições de um hiperespaço nas quais não podem ser separados por uma linha reta.

Esta capacidade de classificar dados não lineares a posiciona como uma rede adequada a solucionar problemas mais próximos da realidade.

Diversos modelos de redes neurais são encontrados na literatura. Neste estudo foram utilizadas redes *feed–forward*, onde várias camadas são organizadas horizontalmente. Cada neurônio se conecta e envia informação para todos os neurônios da camada seguinte. O modelo de *back-propagation* ou de retro-propagação é um modelo *feed-*

forward, sendo o modelo mais conhecido e referenciado na literatura (SHARD,1994). Este estudo usa um modelo de retro-propagação baseado nos Modelos de Processamento Distribuído Pararelo, proposto por (RUMELHART; MAcLELLAND, 1986), onde os valores de entrada são transmitidos de uma camada para a outra e transformados através de pesos de conexões entre os neurônios.

Seja W_{ji} o peso da conexão entre os neurônios **I** e neurônio **J**. A rede acumula seu conhecimento através dos pesos de conexão.

Podemos tomar como exemplo um neurônio **J** na camada **Jc**:

1 – O neurônio **J** recebe as saídas $O_i(t)$ dos neurônios precedentes e o valor N_j é computado: $N_j = \sum W_{ji}O_i$;

2 - N_j é introduzido numa função de ativação $f(N_j)$ que produz um valor de ativação A_j. No modelo de RUMELHART o valor de ativação é igual a N_j ; e

3 - A intensidade do sinal enviado de um neurônio para o outro é uma função da intensidade do valor de ativação. Uma função de transferência toma em conta o valor de ativação e produz o sinal de saída O_j.

$$O_j(t+1) = \frac{1}{1+e^{-(\sum WiOi+\theta j)}} \quad \text{(eq. 1)}$$

Método de Aprendizado por Retro-Propagação

Os pesos de conexão apropriados entre os neurônios são determinados, usando-se um método de aprendizado. Esses pesos são geralmente valores arbitrários no início do processo de aprendizado. São corrigidos ao passo que o aprendizado evolui usando-se

exemplos (ou fatos) representativos do problema em estudo. Estes fatos são relidos sucessivas vezes até que todos sejam aprendidos pela rede. A retro-propagação é um método de aprendizado supervisionado pois os resultados observados na saída dos neurônios são usados para ajustar os pesos de conexão. O vetor obtido na saída dos neurônios é comparado com o vetor de saída desejado. Se houver uma diferença entre os dois vetores, os pesos de conexão são corrigidos ocorrendo assim o aprendizado.

O modelo de RUMELHART é baseado em uma regra de aprendizado chamada Regra Delta Generalizada (RUMELHART; MAcLELLAND, 1986):

$$\Delta W_{ji}(n+1) = \mu \delta j O_i + \alpha \Delta W_{ji}(n)$$

$\Delta W_{ji}(n+1)$ é o ajuste introduzido em n+1 no peso de conexão entre os neurônios i e j.

μ é uma constante chamada taxa de aprendizado que controla a intensidade de correção feita nos pesos de conexão a cada iteração do processo. Quanto maior a taxa de aprendizado tanto maior as mudanças que serão introduzidas nos pesos em cada iteração. α é uma constante chamada fator suavizante que faz o processo de aprendizado considerar o valor do peso no momento *n*.

Na criação das redes, existem uma série de parâmetros que poderão influenciar sua capacidade preditiva. No entanto a falta de metodologia na concepção de redes neurais (BRAGA; LUDEMIR; CARVALHO, 2000), impossibilita sua escolha de maneira racional (número de neurônios na camada intermediária, número de iterações, taxa de aprendizado, entre outros).

Para uma melhor compreensão do processo de aprendizado, pode-se supor, por exemplo, que cada combinação de pesos e limiares corresponda a um ponto na superfície de solução. Considerando que a altura de um ponto é diretamente proporcional ao erro associado a este ponto, a solução está nos pontos mais baixos da superfície.

O algoritmo *back-propagation* procura minimizar o erro obtido pela rede ajustando pesos e limiares para que eles correspondam às coordenadas dos pontos baixos da superfície de erro. Para isto, ele utiliza um método de **gradiente descendente**.

O gradiente de uma função está na direção e sentido em que a função tem taxa de variação máxima. Isto garante que a rede caminha na superfície na direção que vai reduzir mais o erro obtido. Para superfícies simples, este método certamente encontra a solução com erro mínimo. Para superfícies mais complexas, esta garantia não mais existe, podendo levar o algoritmo a convergir para mínimos locais.

O algoritmo *back-propagation* fornece uma aproximação da trajetória no espaço de pesos calculado pelo método do gradiente descendente. Estes pontos ou áreas podem incluir platôs, mínimos locais ou arestas.

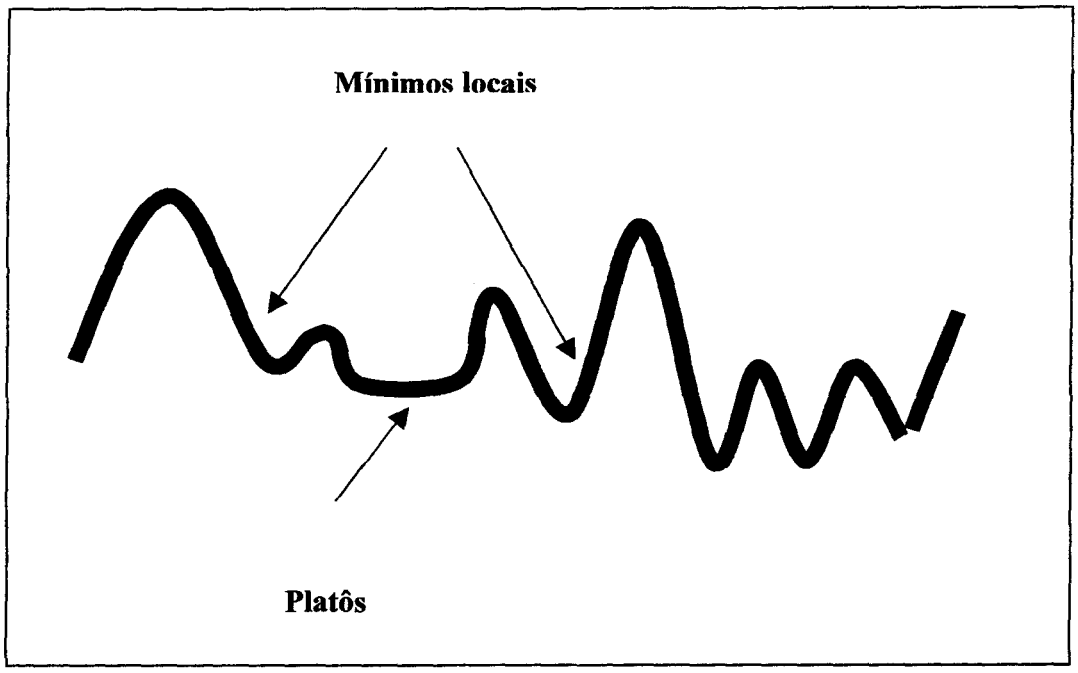


Figura 4.7 : Ilustra mínimos locais e platôs.

A seleção dos parâmetros do algoritmo de *back-propagation* é um processo muito pouco compreendido, e pequenas diferenças nestes parâmetros devem levar a grandes diferenças tanto no tempo de treinamento quanto na generalização obtida. Não é raro encontrar na literatura, para um mesmo problema, utilizando o mesmo método de treinamento, tempos de treinamento que diferem em uma ou mais ordens de magnitude (BRAGA; LUDEMIR; CARVALHO, 2000).

Uma dúvida que surge naturalmente diz respeito a quando parar de treinar a rede. Existem vários métodos para determinação do momento em que o treinamento deve ser encerrado. Estes métodos são chamados de *critérios de parada*. Os critérios de parada mais utilizados são:

1. encerrar o treinamento após *N* ciclos;

2. encerrar o treinamento após o erro quadrático médio ficar abaixo de uma constante α;
3. encerrar o treinamento quando a percentagem de classificações corretas estiver acima de uma constante α (mais indicado para saídas binárias); e
4. combinação dos métodos acima.

Como dissemos anteriormente, o algoritmo *back-propagation* apresenta uma série de dificuldades ou deficiências que podem desestimular uma maior disseminação de seu uso. O principal problema diz respeito à lentidão do algoritmo para superfícies mais complexas. Uma forma de minimizar este problema é considerar efeitos de segunda ordem para o gradiente descendente. Não é raro o algoritmo convergir para mínimos locais.

Mínimos locais são pontos na superfície de erro que apresentam uma solução estável, embora não seja a saída correta. Algumas técnicas são utilizadas tanto para acelerar o algoritmo quanto para reduzir a incidência de mínimos locais:

- utilizar taxa de aprendizado decrescente;
- adicionar nós intermediários;
- utilizar um termo *momentum* (η);
- adicionar ruído aos dados.

Entre as várias técnicas utilizadas para acelerar o processo de treinamento e evitar mínimos locais, a adição de um termo *momentum* (RUMELHART; MAcLELLAND, 1986) é uma das mais freqüentes. A inclusão do termo *momentum* na fórmula de ajuste dos pesos aumenta a velocidade de aprendizado (aceleração), reduzindo o perigo de instabilidade.

Outro problema enfrentado pelas Redes Neurais de um modo geral é o problema do "esquecimento catastrófico" , quando, ao aprender novas informações, a rede esquece informações previamente aprendidas. Uma rede apresenta capacidade de generalização quando classifica corretamente padrões não-utilizados no treinamento ou com ruído.

A generalização ocorre através da detecção de características relevantes do padrão de entrada. Assim, padrões desconhecidos são atribuídos a classes cujos padrões apresentam características semelhantes. A generalização também permite a tolerância a falhas. Mais informações sobre variações do algoritmo de *back-propagation* como: *Quickprop* (FAIRHURST, 1988 P.193), Levenberg-Marquardt (HAGAN; MENHAJ, 1994), *momentum* de segunda ordem (PEARLMUTTER, 1992 P. 887-894), Newton (BATTITI, 1991) e *Rprop* (RIEDMILLER, 1994), poderão ser encontradas nos livros indicados.

Por tudo que foi dito anteriormente, a escolha adequada dos valores dos parâmetros tende a ficar baseada na experiência e sensibilidade do analista que conduz a construção da rede. Neste estudo assumimos os seguintes valores de parâmetros:

Taxa de aprendizado (μ) = 1

Momentum (η) = 0,9

Foram criadas redes com 6 neurônios na camada intermediária, pois foram considerados como *defaut* pela ferramenta de concepção utilizada (SAS), quando os dados apresentam ruídos moderados.

Como o funcionamento é baseado na propagação da informação de erro a partir da camada de saída até a camada de entrada, não é comum usar muitas camadas escondidas, pois cada vez que o erro é propagado para a camada anterior, torna-se uma estimativa menos precisa. Neste sentido o número de camadas que possa melhorar o modelo utilizado, ainda é algo a se estudar.

O número de iterações, que é o número de vezes em que a amostra de base será lida durante o processo de aprendizado, foi limitado a 200 iterações nas experiências realizadas, uma vez que o processo de aprendizado , não evoluiu com um número maior de iterações. O processo de aprendizado foi encerrado automaticamente quando a taxa de precisão requerida foi alcançada..

Os resultados da análise do modelo de rede neural foram obtidos usando o Enterprise Miner , e estão listados abaixo:

Tabela 4.7

Resultados Rede Neural

	Fit Statistic	Training	Validation	Test
1	[TARGET=TERFILHO]	.	.	.
2	Average Profit	0.627121565	0.6243937277	0.6254968203
3	Misclassification Rate	0.1178155566	0.1180686229	0.1178892587
4	Average Error	0.2940351472	0.2950969935	0.2946000164
5	Average Squared Error	0.087304902	0.0875726349	0.0874255475
6	Sum of Squared Errors	37488.724933	28202.767073	28155.222715
7	Root Average Squared Error	0.2954740294	0.2959267391	0.2956781146
8	Root Final Prediction Error	0.2957976178	.	.
9	Root Mean Squared Error	0.2956358679	0.2959267391	0.2956781146
10	Error Function	126258.69222	95035.986758	94875.34609
11	Mean Squared Error	0.0874005664	0.0875726349	0.0874255475
12	Maximum Absolute Error	0.9972644789	0.9958488593	0.9972005491
13	Final Prediction Error	0.0874962307	.	.
14	Divisor for ASE	429400	322050	322048
15	Model Degrees of Freedom	235	.	.
16	Degrees of Freedom for Error	214465	.	.

17	Total Degrees of Freedom	214700	.	.
18	Sum of Frequencies	214700	161025	161024
19	Sum Case Weights * Frequencies	429400	322050	322048
20	Akaike's Information Criterion	126728.69222	.	.
21	Schwarz's Baysian Criterion	129143.78652	.	.

Fonte: IBGE/Censo Demográfico 1991 SAS-Enterprise Mine. Estimativas da autora.

Uma medida bastante empregada para medir a exatidão dos modelos é RMSE (raiz do erro médio quadrático), que penaliza o erro de maior magnitude.

Avaliar o modelo

O experimento foi efetuado através da construção de resultados dos dados do Rio de Janeiro, comparando os resultados da matriz de confundimento e Akaike dos modelos de Regressão Logística (none), onde todas as variáveis entram no modelo, e o modelo de rede Neural, conforme tabelas abaixo:

Tabela 4.8:

Resultados da Classificação

Grupo	Porcentagem de classificação correta	
	Regressão Logística	Rede Neural
Ter Filhos	95,12	95,53
Não Ter Filhos	83,01	76,00
Akaike	106874.89	126728.69

Fonte: IBGE/Censo Demográfico 1991. Estimativas da autora.

AIC – *Akaike's Information Criterion* :

$$\text{AIC} = (n)(\ln(SSE/n) + 2p$$

Onde *n* é o número de casos, *SSE* é a soma dos erros quadráticos e *p* é o número de parâmetros do modelo.

Tabela 4.8

Medidas de Qualidade

Medidas de Qualidade		
	Regressão Logística	**Rede Neural**
Classificação exata	**90,57**	**88,20**
Taxa de erro	**9,43**	**11,80**

Fonte: IBGE/Censo Demográfico 1991. Estimativas da autora.

Como resultado das tabelas 4.7 e 4.8, obtivemos porcentagem de classificação superior através da técnica Logito, principalmente quando utilizado o modelo de regressão logística. Devemos levar em conta também, que algumas das variáveis na regressão logística não puderam ser utilizadas por indisponibilidade de dados (*missing*). Este problema não ocorre com a Rede Neural que podem levar em consideração informações incompletas.

4.4 Aplicar o novo modelo

Após a escolha do modelo que com os parâmetros usados, apresentou o melhor resultado, aplicamos essa informação no arquivo 2, que possuía somente dados onde as variáveis do bloco de fecundidade haviam sido informados como " 99" . Como conclusão temos que desse número de 21489 mulheres onde a informação de fecundidade foram ignoradas, aproximadamente 16% (3378) delas teriam informações de filhos, e 84% (18111) desse total de mulheres as informações sobre fecundidade poderiam ser efetivamente ignoradas. A tabela 4.9 representa a distribuição das mulheres neste arquivo.

Tabela 4.9

Mulheres com ou sem filhos por situação conjugal e grupo de idade

Mulheres que Não teriam Filhos				
Grupo de idade	Situação Conjugal atual			
	Casada	Separada/ viúva	Solteira	Total
10|-----15	5	8	7807	7820
15|-----18	12	5	3298	3315
18|-----20	8	2	1588	1598
20|-----25	17	0	2464	2481
25|-----40	7	0	1958	1965
40|-----60	10	0	495	505
60|------	8	0	419	427
Total	**67**	**15**	**18029**	**18111**

Mulheres que Teriam Filhos				
Grupo de idade	Situação Conjugal atual			
	Casada	Separada/ viúva	Solteira	Total
10|-----15	2	0	0	2
15|-----18	25	5	0	30
18|-----20	52	8	0	60
20|-----25	250	26	0	276
25|-----40	1025	144	105	1274
40|-----60	766	245	128	1139
60|------	223	364	10	597
Total	**2343**	**792**	**243**	**3378**

Fonte: IBGE/Censo Demográfico 1991. Estimativas da autora.

Tabela 4.10

Total de mulheres por localização da residência - geral

	Localização da residência				
Grupo de idade	Isolada/ condomínio	Conjunto popular	Favela	Improvisado ou coletivo	Total
10|-----15	52427	4448	7371	244	64490
15|-----18	29715	2526	4104	104	36449
18|-----20	19652	1606	2739	82	24079
20|-----25	48765	3941	6650	239	59595
25|-----40	141246	11845	17188	680	170959
40|-----60	112477	9486	11057	629	133649
60|------	59945	3932	4117	1023	69017
Total	464227	37784	53226	3001	558238

Fonte: IBGE/Censo Demográfico 1991. Estimativas da autora.

Tabela 4.11

Total de mulheres por localização da residência – dados excluídos

	Localização da residência				
Grupo de idade	Isolada/ condomínio	Conjunto popular	Favela	Improvisado ou coletivo	Total
10|-----15	6309	525	955	33	7822
	12,03	11,80	12,96	13,52	12,13
15|-----18	2722	231	381	11	3345
	9,16	9,14	9,28	10,58	9,18
18|-----20	1375	103	177	3	1658
	7,00	6,41	6,46	3,66	6,89
20|-----25	2280	189	272	16	2757
	4,68	4,80	4,09	6,69	4,63
25|-----40	2671	183	350	35	3239
	1,89	1,54	2,04	5,15	1,89
40|-----60	1370	90	159	25	1644
	1,22	0,95	1,44	3,97	1,23
60|-----	824	45	70	85	1024
	1,37	1,14	1,70	8,31	1,48
Total	17551	1366	2364	208	21489

Fonte: IBGE/Censo Demográfico 1991. Estimativas da autora.

Outro dado a ser verificado, é que comparando as tabelas 4.10 e 4.11, verificamos que a incidência da falta de informação são maiores na faixa de 10 a 15 anos, seguido das de 15 a 18 anos, confirmando a afirmação anterior de que a maior parte das não–respostas é ocasionada pelo constrangimento do recenseador em fazer a pergunta se a mulher teve ou não filhos. E que apesar do percentual diferente entre as classes de localização da residência, os valores são bem próximos, sendo maiores nos domicílios improvisados ou coletivos, seguidos pelos domicílios localizados em favelas.

Algumas considerações não poderão ser esquecidas, para que esse trabalho possa ser implementado nas pesquisas do IBGE, precisaríamos verificar se este modelo poderia ser utilizado para todos os estados, e isto implicaria em testes com os dados de todas as UF´s, e necessitaria também ser feito uma avaliação não só se o modelo vale para todos

os estados ou se dependendo da situação do domicílio, urbano ou rural, precisaríamos descer a nível de município neste estudo, isto é, em que nível de desagregação, deveria ser feito a modelagem.

Como os programas de imputação das variáveis do bloco de fecundidade são executadas depois dos processos normais de imputação das outras variáveis, e é um bloco independente, nossa proposta só seria implementada a partir desta fase.

Como sugestão para aplicação do modelo para as variáveis de fecundidade, seguiríamos as seguintes etapas:

1. Separar as mulheres de 10 anos e mais com dados inconsistentes no bloco de fecundidade para aplicação do modelo.
2. Das mulheres que o modelo imputasse zero filhos as variáveis numéricas seriam zeradas.
3. Uso de uma técnica "hot deck" para a imputação do número total de filhos para os casos não selecionados na etapa anterior.
4. Uma vez que o total de filhos tenha sido imputado, seria necessário distribuir este valor para o número de filhos homens e mulheres.

5. Conclusão

Este capítulo apresenta as conclusões do trabalho. Inicialmente é apresentada uma síntese do estudo, ressaltando os métodos utilizados. Em seguida são relacionadas as principais contribuições da dissertação e finalmente são apresentados os trabalhos futuros que prevêem estudos para aplicação e extensão dos estudos desse trabalho.

5.1 Síntese do trabalho

Este trabalho apresentou uma comparação entre dois modelos de dados para predizer, a existência ou não de filhos nas mulheres com variáveis do bloco de fecundidade rejeitadas por respostas inconsistentes nos dados do Censo Demográfico de 1991, Regressão Logística e uma Rede Neural, a qual o modelo de Regressão Logística com estatística *Wald* apresentou um melhor resultado.

Nesta mesma linha de trabalho, as Nações Unidas recomenda o uso do método de El-Badry, onde se calculam as proporções de imputação de zero filhos para cada faixa de idade das mulheres, baseando-se na distribuição do número de mulheres sem filhos, e se aplica essa proporção às mulheres com resposta inválidas. Pelo estudo de caso, podemos concluir que o modelo proposto por esta dissertação, poderia ser usado em substituição ao modelo indicado pelas Nações Unidas, com a vantagem de individualização dos registros, uma vez que a imputação é feita para cada mulher independente do grupo de idade ao qual ela pertence.

Apesar da pequena diferença entre os resultados da Regressão Logística e a Rede Neural, a qualidade de classificação da Rede poderia eventualmente melhorar a partir da exploração de outros métodos

5.2 Contribuições do trabalho

As principais contribuições deste trabalho são:

- Estudar e apresentar uma metodologia para aplicação das técnicas de Mineração de Dados em problemas de imputação de dados demográficos;
- Aplicar a técnica de Busca de Conhecimento direta nos casos apresentados;
- Apresentar um resultado comparativo entre os modelos estudados e constatar que o modelo de Regressão Logística se mostrou mais aplicável para o domínio da aplicação estudado no estudo de caso apresentado ;
- Mostrar que o modelo de Regressão Logística pode ser aplicado como método de estudo nos Censos Demográficos.

5.3 Trabalhos futuros

Como trabalhos decorrentes desta dissertação, podemos recomendar as seguintes pesquisas:

1. Um estudo da viabilidade da aplicação do modelo de previsão em Mineração de Dados da Árvore de Decisão para imputar os dados de Filhos Tidos Nascidos Vivos por sexo. Estudos iniciais sobre esta pesquisa já foram iniciados pela autora desta dissertação;
2. Aplicar os métodos utilizados nesta dissertação em dados do Censo Demográfico de 2000, para verificar se esta comparação se sustenta e avaliar a possibilidade de sua aplicação em regime de produção em outras pesquisas; e
3. Estudar e avaliar a utilização das técnicas de limpeza e preparação dos dados da Mineração de Dados na preparação dos dados das pesquisas

demográficas, visando diminuir o impacto das não-respostas nos processos de imputação posteriores.

6. Referências Bibliográficas

ADRIAANS P., ZANTINGE D. *Data mining.* Harlow, England : Addison-Wesley, 1996. 158 p.

ALBIERI, S. *A Ausência de resposta em pesquisas*: Uma aplicação de método de imputação. Rio de janeiro : IMPA, 1992. 138 p. (Informes de matemática. Série D, n. 48/92). Originalmente apresentada como dissertação (Mestrado).

ALTMANN, AM.G.; FERREIRA C.E.C. *Evolução do Censo Demográfico e Registro Civil como fontes de dados para a análise da fecundidade e mortalidade no Brasil* : (dados coletados e estudos realizados). São Paulo : Fundação Sistema Estadual de Análise de Dados, Grupo Especial de Análise Demográfica, 1979. 81 f. Trabalho apresentado na III Reunion del Grupo de Trabajo Sobre Informacion Sociodemografica, Lima, Peru, 21 – 25 de mayo de 1979.

ANDREWS, H.C. *Introduction to mathematical techniques in pattern recognition.* New York : Wiley-Interscience, 1972. 242 p.

ARABIE, P.; HUMBERT, L.J.; SOETE, G. (Ed.). *Clustering and classification.* Singapore : Word Sci., 1996. 490 p.

BAILEY, D.; THOMPSON, D. Developing neural network applications. *AI Expert*, 5:34-41, 1990

BATTITI, R. *First and second-order methods for learning:* between steepest descent and Newton's method. Trento : University of Trento, 1991. (Technical report).

BERRY, M.J.A.; LINOFF, G. *Data mining techniques:* for marketing, sales and customer support. New York: J. Wiley & Sons, 1997. 454 p.

BEZDEK, J. C.; PAL, S. K. (Ed.) *Fuzzy models pattern recognition:* methods that search for structures in data. New York : IEEE, 1992. 532 p.

BISHOP, C.M. *Neural networks for pattern recognittion.* Oxford : University Press Inc., USA, NY, 1995.

BRACHMAN, R.J.; ANAND, T. The Process of knowledge discovery in databases. *In:* FAYYAD, U. M. (Ed.). *Advances in knowledge discovery and data mining.* Menlo Park, CA : AAAI ; MIT Press, 1996, p. 37-57.

BREIMAN, L. et al. *Classification and regression trees.* Monterey : Wadsworth & Books, 1984. 358 p.

BRAGA, A. P.; LUDEMIR, T.B.; CARVALHO, C.P.L.F. *Redes neurais artificiais :* teorias e aplicações. Rio de Janeiro : LTC, 2000. 262 p.

BUSSAB,W.O.; MIAZAKI, E.S.; ANDRADE, D.F. *Introdução à Análise de Agrupamentos.* IX Simpósio Brasileiro de Probabilidade e Estatística. IME-USP, São Paulo 1990.

CABENA, P. et al. *Discovering data mining from concept to implementation.* Upper Saddle River : Prentice- Hall ; San José : IBM, 1998. 195 p.

CARVALHO, L.A.V. Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração. 2. ed. São Paulo: Érica, 2001. 256 p.

CICHOCKI, A. ; UNBEHAUEN, R. *Neural networks for optimization and signal processing.* Stuttgart : B. G. Teubner ; Chichester : J. Wiley, 1993. 526 p.

CORTES, S. da C.; PORCARO, R.M.; LIFSCHTZ, S. *Mineração de Dados – Funcionalidades, Técnicas e Abordagens.* Monografias em Ciência da

Computação, nº 10/2002 ISSN 0103-9741 PUC-Rio – Departamento de Informática Ed: Carlos J.P. Lucena.

CROFT, T. *Demographic and Health Surveys Word Conference*, Washington, Proceedings vol.2. 1991 Bureau of the Census. Research Conference, 1995, Arlington. Proceedings.

DIAS, A .J. R.; ALBIERI, S. Uso de imputação em pesquisas domiciliares. In: ENCONTRO NACIONAL DE ESTUDOS POPULACIONAIS, 8., 1992, Brasília, DF. *Anais...* São Paulo: ABEP, 1992. v.1, p.11-26.

DINIZ, C.A.R.; NETO, F.L. *Data Mining: Uma introdução* 14º Sinape – Caxambu : ABE – Associação Brasileira de Estatística, 2000.

DOBSON, A .J. *Introduction to statistical modelling.* London ; New York : Chapman and Hall, 1983. 125 p.

ELMASRI, R.; NAVATHE, S. *Fundamentals of database systems. 3rd ed.* Reading : Addison-Wesley, 2000. 955 p.

FAIRHURST, M. C. *Computer vision for robotic systems : an introduction.* New York : Prentice-Hall, 1988. 193 p.

FELLEGI, I. P.; HOLT, D. A systematic approach to automatic data editing and imputation. *Journal of the American Statistical Association*, Washington, DC, v. 71, n. 353, p. 17-35, 1976.

FU, K. S. *Syntactic pattern recognition and applications.* Englewood Cliffs: Prentice – Hall, 1982. 596 p.

FU, LiMin. *Neural networks in computer intelligence.* New York: McGraw-Hill, 1994. 460 p.

GROSSMAN, R. *Data mining:* challenges and opportunities for data mining during the next decade. Disponível em: < http://www.magify.com>. Acesso em : dez. 1997.

GROTH, R. *Data mining:* building competitive advantage. Upper Saddle River, NJ : Prentice–Hall, 2000. 299 p.

HAIR, J. F. et al. *Multivariate data analysis*. 5th. ed. Upper Saddle River, NJ : Prentice-Hall. 1998. 730 p.

HAGAN, M.; MENHAJ, M. Training feed forward networks whit the marquardt algorithm. *IEEE transactions on Neural Networks*, New York, v. 5, n. 6, p. 989-993, Nov. 1994.

HAN, J.; KAMBER, M. *Data mining :* concepts and techniques. San Francisco *:* Morgan Kaufmann Publishers. 2001. 550 p.

HAYASHI, Y. A neural expert system with automated extraction of fuzzy if-then rules and ts application in medical diagnosis. In R.P. Lippmann, J. E. Moody, and D.S. Touretzky, editors, *Advances in Neural Information Processing Systems 2*, pages 578-584. Morgan Kaufmann, 1991.

HAYKIN, S. *Neural Networks – A Comprehensive Foundation.* Macmillan College Publishing Company, USA, NY, 1994.

HUSH, Don R. & HORNE, Bill G. *Progress in Supervised Neural Networks - What's New Since Lippmann?* IEEE Signal Processing Magazine, janeiro de 1993, p. 8-39.

INMON, W.H.; HACKATHORN, R. D. *Como usar o data warehouse.* Tradução: Olavo Faria. Rio de Janeiro: Infobook, 1997. 277 p.

INMON, W H. *Como construir o data warehouse* . [Tradução: Ana Maria Netto Guz]. Rio de janeiro: Campus, 1997. 388 p.

KALTON, G.; KASPRZYK, D. *Inputing for missing survey responses.* In Proceedings of the Survey Research Methods Section, American Statistical Association, 1982.

KASABOV, N. K. *Foundations of neural networks, fuzzy systems, and knowledge engineering.* Cambridge, Mass. : MTI Press, 1996. 550 p.

KING, D. *Numerical machine learning.* Disponível em: < http://www.cc.gatech.edu/~kingd/datamine/datamine.html>. Acesso em: 22 set. 1999.

KLIR, G.J.; YUAN, B. *Fuzzy sets and fuzzy logic : theory and applications.* Upper Saddle, NJ : Prentice-Hall, 1995. 574 p.

KOHN, A.F. *Reconhecimento de padrões :* uma abordagem estatística. São Paulo : USP. 1998. 205 p.

LITTLE, R. J. A.; RUBIN, D. B. *Statistical analysis with missing data.* New York :J. Wiley & Sons, 1987. 278 p.

MACHADO, A. *Neuroanatomia Funcional.* Livraria Atheneu, 1980

MATTISON, R. *Data warehousing and data mining for telecommunications.* Boston: Artech House, 1997. 273 p.

MENDES, I. M. B.; PLASTINO, A.; OCHI, L. S. *Regras de associação*: suas diferentes formas e seus algoritmos de mineração. Trabalho apresentado no minicurso SBBD 2001.

NELSON, D.M. & ILLINGWORTH,W.T. *A Pratical Guide to Neural Nets.* Addison-Wesley Publishing Company, Inc. Usa, 1991.

OLIVEIRA, L.C.S. et al. *Apuração dos dados investigados pelo questionário da amostra – CD 1.02 do censo demográfico de 1991.* Rio de Janeiro : IBGE, Diretoria de Pesquisas, 1996. 77 p. (Texto para discussão, n. 86).

PAO, Yoh-Han. *Adaptive pattern recognition and neural networks.* Reading. MA : Addison-Wesley, 1989. 309 p.

PANDYA, A. S. ; MACY, R. B. *Pattern recognition with neural networks in C++.* New York : CRC Press, 1996. 410 p.

PEARLMUTTER, B. Gradient descent: second order momentum and saturation error. In: MOODY, J. E.; HANSON, S.; LIPPMANN, R. (Ed.). *Advances in neural information processing systems 2. San Mateo, CA : Morgan Kaufmann. 1992. p.* 887-894.

RIEDMILLER, M. R. *Description and implementation details.* Karlsruhe : University of Karlsruhe, 1994. (Technical report)

ROITMAN, V. L. *Um modelo computacional de redes neirais para predição do índice de desemprego aberto.* 201. 1 v. Tese (Doutorado) – Universidade Federal do Rio de Janeiro, COPPE – RJ.

RUBIN, D.B. *Multiple Imputation for nonresponse in surveys.* New York : J. Wiley, 1987. 258 p.

RUMELHART, D. E.; MAcLELLAND, J. L. *Parallel distributed processing : exploration in the microtexture of cognition.* Londres : MIT Press, 1986. V. 1 : Foundations.

SCHALKOFF, R. J. Pattern r*ecognition: statistical, structural and neural approaches.* New York : J. Wiley , 1992. 364 p.

SHARDA, R. Neural networks for the MS/OR analyst: an application bibliography . *Interfaces*, Providence, RI, v. 24, n. 2, p.116-130, 1994.

SILVA, A. N.; CORTEZ, B. F. *Uma proposta para a imputação das variáveis numéricas de fecundidade do censo demográfico 2000.* Rio de Janeiro : IBGE, Diretoria de Pesquisas, 2000. 21 p.

SILVA, P.L.N. *Crítica e imputação de dados quantitativos utilizando SAS.* 1989. 1 v. Monografia (Mestrado) – Instituto de Matemática Pura e Aplicada, Rio de Janeiro.

SKAPURA, David M. *Building Neural Networks.* ACM Press Books, USA, NY - New York, 1996.

SULAIMAN, A; MOREIRA, J. Prospecção de Conhecimento em banco de dados. *Developers Magazine* , Rio de Janeiro, fev. 1997.

TÉCNICAS indirectas de estimacion demografica. Nueva York : Naciones Unidas, 1986. 318 p. (Estudios de poblacion, n.81). ST/ESA/Ser.A/81. Acima do Título: Manual X.

TODESCO, J. L. *Reconhecimento de Padrões Usando Rede Neural Artificial com uma Função de Base Radial: Uma aplicação na Classificação de Cromossomos Humanos.* Tese de Doutorado, Universidade Federal de Santa Catarina – UFSC, 1995.

WEISS, S. M.; INDURKHYA, N. *Predictive data mining: a pratical guide.* San Francisco : Morgan Kaufmann, 1997. 228 p.

WESTPHAL, C.; BLAXTON, T. *Data mining solutions:* methods ans tools for solving real: word problems. New York: J. Wiley & Sons, 1998. 617 p.

Falta colocar o questionário do Censo

ANEXOS

Tabela 1: Classificação de Variáveis

Variável	Nome	Código	Descrição
V0202	Local	1	Casa/apartamento isolado ou condomínio
		2	Casa/apartamento cjto residencial
		3	Casa/apartamento em favela
		4	Improvisados ou coletivos
V0310	Religião	0	Sem religião
		1	Católica
		2	Evangélica
		3	Espírita
		4	Judaica
		5	Oriental
		6	Outras
		7	Mal definidas
		9	Sem declaração
V1061	Sitdom	1	Urbano
		2	Rural
		99	Menos de 10 anos
V3072	Gidade	1	De 0 a 4 anos
		2	< 10 anos
		3	< 15 anos
		4	< 18 anos

		5	< 20 anos
		6	< 25 anos
		7	< 40 anos
		8	< 60 anos
		9	>= 60 anos
V0345*V0358	Ocupa	0	Sem ocupação
		1	Trabalha
		2	Procurando
		3	Pensionista/aposentado
		4	Vive de renda
		5	Estudante
		6	Afazeres domésticos
V3342	Situacasa	1	Casada
		2	Separada ou viúva
		3	Solteira
		4	Menos de 10 anos

Aplicação das Técnicas de Data Mining

Westphal & Blaxton (1998) e Mattison (1997) destacam cinco áreas ou setores onde as técnicas de Mineração de Dados tem sido amplamente desenvolvidas e aplicadas: Químico, Varejo, Financeiro, Médico e a de Telecomunicações. Além desses autores, temos referências dedicadas apenas a aplicação de ferramentas de Mineração de Dados, como apresentado por Groth (2000) e por Cabena et al (1998) e Weiss & Indurkhya (1997), confirmando a concentração de aplicações nessas 5 áreas.

Dessa forma, temos 5 áreas de aplicação ou setores onde problemas possíveis de serem tratados como Mineração de dados tem sido identificados e solucionados. O objetivo deste capítulo não é o de ser extensivo e sim o de prover, para as áreas descritas, um mapeamento das aplicações atuais, possíveis abordagens de solução e incluir a área de Imputação de dados como mais uma área de aplicação.

A seguir, detalharemos cada um dos grupos.

Químico

No setor químico e farmacêutico, as técnicas de Mineração de Dados são utilizadas na administração de informações, na descoberta de novos e para análise e interpretação de dados para sua avaliação, seja em termos da sua utilidade como de sua viabilidade econômica. Uma peculiaridade do processo de descoberta de um novo produto é a elevada gama de possíveis soluções e o pouco espaço para otimização. Essa procura é cara e o fator tempo é fundamental.

Além disso, as técnicas de Mineração de Dados tem sido aplicada no estudo o genoma humano. Através de seqüências e características do DNA humano, os cientistas buscam identificar genes que são específicos de doenças, para geração
de tratamentos e remédios. Através de grandes bancos de dados, técnicas de visualização em 3D são utilizadas para identificação de padrões genéticos.

Método de Abordagem: Redes Neurais
Tipo de Problema: Classificação

Exemplo 1: Classificação de seqüências de DNA

Esta aplicação consiste em identificar a seqüências de DNA que são codificadas (*exons*) das seqüências que não são codificadas (*introns*).

Com base nas informações das seqüências conhecidas, *Redes Neurais* são desenvolvidas com o objetivo de classificar novas seqüências. Através dos parâmetros de entrada é possível obter um resultado classificando a seqüência como um *exon* ou um *intron*.

Método de Abordagem: Redes Neurais
Tipo de Problema: Previsão

Exemplo 2: Reações químicas

Prever reações químicas a partir de informações como temperatura, pressão, composição da mistura, etc após um certo período de tempo.

Com base em informações com experimentos anteriores, podemos modelar uma *Rede Neural* utilizada para previsão de resultados, através de alterações dos parâmetros de entrada.

Método de Abordagem: Regras de Associação
Tipo de Problema: Previsão

Exemplo 3: Intoxicação

Produzir estudos para antecipar o efeito de substâncias químicas nos seres humanos e nos animais. Alguns desses efeitos podem ser tóxicos, e por analogia podem ser previstos, isto é, uma estrutura química pode ser relacionada com outras conhecidamente tóxicas devido a um mecanismo comum entre elas. Usando o método de *Regras de Associação*, substâncias com estruturas químicas semelhantes são detectadas, resultando uma previsão de dano tóxico ou não.

Varejo

O setor varejista é um dos setores onde as técnicas de Mineração de Dados são mais utilizadas. Neste setor a venda de cada item é sensível a diferentes fatores de mercado como propaganda e preço, e fatores como tendências de moda, renda, competição entre outras. Existe uma correlação interna entre a venda de diferentes produtos, que podem não estar aparente.

Os problemas de Mineração de Dados no varejo podem ser agrupados em duas categorias principais: previsão de vendas e otimização de estratégias de marketing.

O problema de previsão de vendas está relacionado com o estudo de uma série temporal de dados passados. Esse problema já foi amplamente abordado pelos estatísticos, principalmente com a utilização de métodos lineares. Quando os fenômenos são complexos e não lineares os métodos comumente utilizados não são suficientes.

O problema de otimização de estratégias de marketing, por outro lado, é um problema conceitual mais complexo, baseado na habilidade de fazer previsões em diferentes condições e na geração de modelos de negócios. Um exemplo é o caso de mala direta. As campanhas de mala direta são caras, sendo importante a limitação da gama de todos os consumidores em subconjuntos de consumidores potenciais. Isso tem sido otimizado com o uso de árvores de decisão e de redes neurais.

Um outro exemplo é a determinação de perfis de consumidores. O objetivo da segmentação de mercado é a exploração de diferenças na curva de demanda de diversos segmentos de mercado.

A análise de sensibilidade e elasticidade também é utilizada para a otimização de estratégia de mercado. A análise de sensibilidade é a avaliação do impacto da mudança de uma variável em uma outra variável, como por exemplo a flexibilidade de preço da demanda. Se uma grande quantidade de dados históricos está disponível, o cálculo de sensibilidades é um problema de generalização de função de um número de valores conhecidos.

Método de Abordagem: Árvore de Decisão
Tipo de Problema: Classificação

Exemplo 1: Perfil do Consumidor

Aumentar a taxa de resposta de uma campanha promocional, mantendo ou até diminuindo o número de material enviado. A abordagem por *Árvore de Decisão* tem sido amplamente utilizada através de um detalhamento histórico de respostas a campanhas contendo características dos clientes. Podemos assim classificar deferentes perfis de consumidores e aumentar a taxa de resposta. (ex; Idade, renda, número de filhos, carro, localização, etc...)

Método de Abordagem: Métodos Analíticos de Visualização
Tipo de Problema: Classificação

Exemplo 2: Detecção de Fraudes

Detecção de fraudes que podem ocorrer durante o transporte ou estocagem de produtos no varejo. Através do histórico de vendas, características geográficas e clima, é possível

traçar padrões de envio de produtos para cada loja. Se o envio de algum item aumenta consideravelmente sem o aumento de um produto diretamente correlacionado, isso pode indicar a existência de fraudes ou erros operacionais. O Monitoramento desses comportamentos podem ser analisados através da abordagem de Visualização de Métodos Analíticos. Definindo grupos de classificações com os limites superior e inferior considerados aceitáveis, utilizando gráficos do tipo *scatter plots,* podemos verificar variáveis que estão fora dos limites, que depois de verificadas podem confirmar a existência ou não de fraudes nos processos.

Método de Abordagem: Redes Neurais
Tipo de Problema: Classificação

Exemplo 3: Análise de Informações de Mercado

A analise de mercado é uma das aplicações mais desafiantes no varejo. A opinião humana é substancialmente diferente de outras aplicações, onde as relações seguem regras definidas (ocultas ou não). 0 objetivo desse tipo de aplicação é o de classificar os clientes de acordo com seu comportamento ou atitude. Com as respostas computadas através de questionários com informações quantitativas e qualitativas, os clientes são classificados em diferentes grupos. Com estes dados armazenados ao longo do tempo, modelos em *Redes Neurais* são desenvolvidos e utilizados para a classificação de novos clientes, ou verificação de mudanças de classes dos clientes existentes. Podemos usar neste caso também uma abordagem usando *Árvore de Decisão.*

Método de Abordagem: Análise de Hierarquias

Tipo de Problema: Classificação

Exemplo 4: Identificação de fluxo de Informações

Analisar o fluxo de informações entre diferentes empresas de uma cadeia varejista. Isto é, através de uma modelagem usando Análise de Hierarquias, classificar este fluxo, de acordo com semelhanças, com o objetivo de diminuir tempos nos processos de negociação, transporte e compras. Com a identificação desses processos, uma padronização pode ser implantada, trazendo redução dos prazos de reposição das mercadorias diminuindo a necessidade de um estoque maior.

Método de Abordagem: Árvore de Decisão
Tipo de Problema: Previsão

Exemplo 5: Índices de performance

Os padrões de consumo no varejo, com relação a escolha do melhor local para efetuar suas compras podem ser alterados ao longo do tempo. Muitos fatores podem influenciar o sucesso de uma loja no varejo, como sua localização, quem pode ser seus competidores, área de influencia e histórico de vendas. Um dos grandes desafios da administração nessa Área é o de quantificar esses fatores. Uma das grandes aplicações para a previsão no varejo está no desenvolvimento de modelos que, através de um histórico de venda e perfis geográficos, responda com índices de performance.

Aplicações utilizando a abordagem de *Árvores de Decisão* tem sido feitas com o objetivo de prever giro de produtos em lojas existentes ou ainda em planejamento.

Dessa forma, com a construção da Árvore de Decisão, decisões de investimento podem ser tomadas para a escolha de novos locais para a construção de lojas, ou investimentos adicionais em lojas existentes.

Esse tipo de problema no varejo também poderia ser abordado através de *técnicas estatísticas de regressão múltipla*, isolando as variáveis relevantes para previsão.

Método de Abordagem: Redes Neurais
Tipo de Problema: Previsão

Exemplo 6: Séries temporais de vendas

Um dos desafios para grandes cadeias varejistas esta na correta previsão de venda dos itens nas lojas para evitar a falta ou excesso de produtos em estoque. Atualmente o número de itens é bastante elevado e o padrão de consumo em cada loja é bastante diferente e heterogêneo. Utilizando um histórico de vendas de cada produto, em cada loja, aplicações utilizando a abordagem de *Redes Neurais* tem sido amplamente utilizadas. A previsão de produtos ou lojas possui uma complexidade elevada, pelos diversos fatores externos que influenciam as vendas, como descontos de preços, clima, atividades do competidor e alterações nos padrões dos clientes. Com a previsão de demanda fornecida pela aplicação em Redes Neurais, estratégias de administração de estoque e reposição dos produtos podem ser otimizadas, evitando assim o excesso de produtos (custo de capital) ou a falta deles (perda de vendas ou clientes). Esse mesmo tipo de problema no varejo também pode ser abordado através de *técnicas estatísticas* de previsão de vendas.

Método de Abordagem: Sistemas de Geoprocessamento
Tipos de Problema: Segmentação

Exemplo 7: Análise de Informações de Mercado

A abordagem de Visualização utilizando Sistemas de geoprocessamento tem sido amplamente utilizada no varejo para *segmentação* de regiões de mercado. Através da relação entre o perfil de cada consumidor e sua localização, geográfica, é possível observar em mapas digitais características de regiões para *segmentação*.
Dessa forma, pode-se detectar as regiões de maior contribuição em receita, maior demanda de classe ou item de produtos, maior homogeneidade de vendas ou sazonalidades.

Método de Abordagem: Clustering
Tipos de Problema: Segmentação

Exemplo 8: Identificação de Áreas de Atendimento

No varejo, uma das decisões operacionais em transporte está na área de atendimento de Lojas para entrega direta a clientes. Essa alocação deve ser feita com o objetivo de minimizar o tempo de entrega, a distância percorrida e o custo total da operação. De posse de informações de histórico de consumo de clientes, sua localização geográfica e a distância rodoviária entre os clientes e os postos de abastecimento, é possível utilizar a abordagem de *clustering* para determinar a origem/destino, para cada postos de abastecimento e clientes.

Método de Abordagem: Redes Neurais
Tipos de Problema: Reconstrução de Funções

Exemplo 9: Análise de Elasticidade de Preço

A elasticidade do preço é o fator entre a variação do preço de venda e a variação de demanda. Para variações de aumento de preço é de se esperar variações de diminuição de demanda. Aplicando *Redes Neurais*, é possível para um conjunto conhecido de variações no preço do produto e sua demanda efetiva após essa variação, *reconstruir uma função* que esteja abrangendo todo o domínio possível de variação de preço. Dessa forma, chega-se a uma relação quantitativa entre variação de preço e variação de demanda, para qualquer valor de preço entre o mínimo e o máximo, auxiliando a tomada de decisão no varejo a promoção de produtos.

Método de Abordagem: Redes Auto-organizadoras
Tipos de Problema:: Segmentação

Exemplo 10: Desenvolvimento de novos produtos

Para o desenvolvimento de novos produtos, muitas vezes pesquisas são feitas com os clientes com relação às necessidades ou importâncias dadas a diversos fatores quando da compra de um novo produto. A partir dessas informações, podemos modelar redes auto-organizadoras que segmentem os tipos de necessidades dos clientes em grupos homogêneos. Dessa forma, produtos ou serviços diferentes podem ser lançados satisfazendo um maior número de pessoas.

Financeira

Na área financeira, o uso de técnicas de Mineração de Dados são aplicadas com relações aos consumidores dos serviços bancários, com aplicações similares ao setor varejista, e nas previsões de índices financeiros. As aplicações lidam com os dados do sistema financeiro, que são altamente voláteis. A maioria das aplicações utilizam métodos como redes neurais e métodos estatísticos.

Método de Abordagem: Árvore de Decisão
Tipos de Problema: Claassificação
Exemplo 1: Aumento de Resposta em Mala-direta

Para o setor financeiro de serviços, como empresas de créditos, a taxa de resposta à oferta de novos produtos ou serviços tem sido aumentada através da utilização de *Árvores de Decisão* na construção de modelos de classificação de clientes. Através de características de clientes e de seu histórico de utilização de serviços bancários, um modelo utilizando *Árvores de Decisão* é desenvolvido com o objetivo de gerar uma classificação de potencial de resposta à novos serviços aos clientes. Isso pode servir de base para o lançamento ou desenvolvimento de novos serviços, diminuição de custos promocionais e aumento da fidelidade do cliente à empresa. A abordagem de Redes Neurais também é utilizada para este tipo de problema.

Método de Abordagem: Redes Neurais
Tipos de Problema: Previsão

Exemplo 2: Retirada de dinheiro em caixas eletrônicos

No setor de Bancos de serviços, um dos problemas está na previsão de retirada de dinheiro pelos clientes para colocação em caixas automáticos. Geralmente os valores são superestimados para evitarem a falta na máquina e a conseqüente diminuição do serviço prestado aos clientes. De posse de dados históricos de operações de saque de dinheiro, utiliza-se uma abordagem de Redes Neurais para prever valores futuros, podendo chegar ao nível de detalhe de cada ponto de atendimento. Com a melhor previsão do montante de dinheiro a ser retirado pelos seus clientes em cada posto automático de atendimento, não só o nível de serviço é implementado como custos com transporte, movimentação e processamento físico dessas quantias são reduzidos. Normalmente, para problemas de previsão, *técnicas estatísticas* poderiam ser utilizadas para solucionar esse tipo de problema.

Método de Abordagem: Redes Neurais
Tipos de Problema: Classificação

Exemplo 3: Análise de Risco de Créditos

Empresas do setor financeiro que trabalham com empréstimos necessitam determinar o risco de crédito de seus clientes. Quanto maior o risco para o pagamento de empréstimos, menor deve ser o limite que a empresa deve permitir. Uma das aplicações está em utilizar a abordagem de Redes Neurais para classificar os clientes de acordo com o risco de crédito. Podemos ter diferentes níveis de classificação de risco e associar esses níveis a estratégias de marketing da empresa ou limitações de crédito. Usando dados históricos que possuem características de cada cliente, tipo de empréstimo e no passado o risco desse cliente, *Redes Neurais* são desenvolvidas com o objetivo de classificar novos

clientes. Ao avaliar a entrada de um novo cliente, automaticamente sua classificação é gerada a partir de seus dados pessoais, retornando o seu risco ao empréstimo solicitado. Com isso, serviços mais personalizados podem ser oferecidos para esses tipos de clientes e a taxa de inadimplência pode ser bastante reduzida.

Método de Abordagem: Redes Neurais
Tipos de Problema: Previsão

Exemplo 4: Previsão de Índices Financeiros

Uma das maiores aplicações da abordagem de *Redes Neurais* no setor financeiro está na previsão de Índices Financeiros em séries históricas. Podem ser considerados diversos índices como: preços de produtos, valores de ações, índices de mercado, etc. Consiste em obter a série temporal histórica do índice. De acordo com parâmetros de entrada definidos, *Redes Neurais* são utilizadas como modelos de previsão dos índices para valores futuros. As informações de previsão são utilizadas como apoio na decisão de compra e venda de títulos, além de detectar tendência em índices financeiros

Método de Abordagem: Analíticos de Visualização
Tipos de Problema: Classificação

Exemplo 5: Detecção de Fraudes

Fraudes no setor financeiro são comuns, ocorrendo principalmente com a transferência fraudulenta de dinheiro, clonagem de cartões, utilização indevida de senhas de terceiros,

etc. Através de uma detalhada base histórica contendo as movimentações de cada cliente, é possível utilizar métodos *Analíticos de Visualização* para monitorar as operações, com o objetivo de detectar fraudes. Usando gráficos do tipo plots, podemos verificar comportamentos que não se adequam ao perfil histórico. Qualquer variação fora de limites aceitáveis que for detectada pode ser investigada, evitando assim comportamento fraudulentos.

Método de Abordagem: Regras de Associação
Tipos de Problema: Previsão

Exemplo 6: Verificação de Preços

Algumas empresas de informação vendem serviços de monitoramento e informação de preços de produtos, ações, etc. para o mercado internacional. Muitos dos dados são atualizados em tempo real, o que pode gerar erros. Uma das aplicações de *Regras de Associação* no setor financeiro está na verificação desses valores, para evitar assim a comunicação de índices errados. De posse de informações históricas de preços, as regras de associação são utilizadas para elaborar previsões dos preços. Quando os preços são atualizados, eles são comparados com a previsão feita. Caso esteja fora de limites aceitáveis, um aviso de erro é acionado, e valor deve ser verificado pela empresa.

Método de Abordagem: Clustering

Tipos de Problema: Segmentação

Exemplo 7: Segmentação de Clientes

Um dos componentes de sucesso nos serviços financeiros está na relação entre a instituição e seus clientes. As necessidades individuais de cada cliente norteiam o planejamento estratégico da empresa e seus investimentos em serviços. Grupos homogêneos de clientes que possuem as mesmas necessidades por serviços podem ser identificados através da abordagem de *Clustering*. Nesse caso, não existe a necessidade de uma pré-classificação. Com os dados característicos de cada cliente e seu perfil de consumo de serviços o método de clustering traz como resposta os grupos homogêneos de acordo com alguma característica. Com o resultado dessa abordagem, estratégias de melhorias de serviço podem ser feitas para cada grupo de cliente, aumentando assim a taxa de satisfação pelos serviços financeiros testados.

Método de Abordagem: Sistema de Geoprocessamento

Tipos de Problema: Descrição

Exemplo 8: Monitoramento de Performances

Sistemas de Geoprocessamento tem sido usados para descrever ou monitorar a performance de instituições financeiras por regiões de atendimento. Podemos ter por exemplo, num mapa digitalizado, os valores de empréstimos feitos pelo banco em cada cidade onde ele atua. De forma rápida, é possível analisar as regiões onde se concentra esse tipo de serviço. Diferenças entre cidades de um mesmo estado podem ser observadas, através de descrições do tipo: " O valor de empréstimos da cidade X é 10 vezes maior que os das outras cidades desse mesmo estado. Através desse

monitoramento que pode ser atualizado em tempo real, podem ser tomadas decisões que otimizem as estratégias de marketing.

Método de Abordagem: Algoritmos Genéticos
Tipos de Problema: Previsão

Exemplo 9: Previsão de Lucro para Portfolio de Investimentos

Uma das aplicações de *Algoritmos Genéticos* na área financeira está na previsão e otimização do lucro em aplicações que possuem múltiplas combinações de investimento em diferentes produtos ou serviços. Através da modelagem do problema, os algoritmos variam os parâmetros de forma a otimizar o lucro, respondendo com o melhor perfil para o portfolio de investimentos.

Método de Abordagem: Técnicas Estatísticas
Tipos de Problema: Afinidade de Grupos

Exemplo 10: Determinação de Serviços associados

Técnicas estatísticas tem sido utilizadas no setor financeiro para condução de estudos semelhantes ao de market basket analysis no varejo. Com base em dados históricos, *modelos estatísticos* são desenvolvidos para detectar correlação entre a utilização de serviços bancários. Com essa informação, as instituições financeiras podem tornar estratégias de marketing que associem esses serviços, ou desenvolver pacotes de serviços que aumentem a satisfação do cliente sem o aumento de custos operacionais.

Médica

A aplicação no setor médico de Mineração de Dados, tem aumentado sua utilização a cada ano. Essas aplicações se dão principalmente no auxílio do tratamento de doenças crônicas e problemas de administração hospitalar. No caso de doeças crônicas, é utilizado para detectar formas de classificar e identificar a partir dos sintomas dos pacientes a melhor terapia a ser utilizada. Já nos problemas de administração hospitalar, consistem em elaborar estratégias para reduzir o tempo de espera entre consultas e fraudes em planos de saúde.

Método de Abordagem: Técnicas Estatísticas

Tipos de Problema: Previsão

Exemplo 1: Previsão de Tempo de Sobrevivência para Doenças Crônicas

Técnicas Estatísticas de Regressão tem sido aplicadas para previsão de tempo de sobrevivência para algumas doenças no setor médico. O principal objetivo é prever, com base em características dos pacientes, o tempo de sobrevivência estimado. Caso esse tempo seja curto, tratamentos mais agressivos e que trazem resultados em menores tempos devem ser priorizados. Por outro lado, se o tempo estimado de sobrevivência for longo, técnicas de tratamento mais amenas que possuem resultados a prazos mais longos podem ser conduzidos. Para esse tipo de aplicação é necessário um histórico com

características dos pacientes, como idade, estado da doença, quando início do tratamento, etc. Essas são chamadas variáveis dependentes. Possuindo conhecimento do tempo de sobrevivência de cada paciente, é possível adequar funções estatísticas de regressão.

Método de Abordagem: Árvore de Decisão

Tipos de Problema: Classificação

Exemplo 2: Detecção de Fraudes

No setor médico temos o problema de auxílio financeiro para o tratamento de doenças, seja para consultas, internação ou compra de medicamentos. Em alguns países, o próprio estado financia programas de auxílio médico. Em outros países, temos o papel de convênios particulares. Essa aplicação consiste em detectar fraudes no pagamento desses auxílios médicos. Através da monitoração histórica de clientes e seu perfil de consumo de serviços médicos, é possível através da Abordagem de *Árvores de Decisão* modelar sistemas que classifiquem o perfil de consumo de cada tipo de cliente de acordo com suas características. A monitoração manual dos pagamentos e investigações extensivas de muitos casos suspeitos torna o processo caro e pouco efetivo. Por outro lado, com a utilização dessa abordagem de Mineração de Dados, podemos após a definição dos padrões, monitorar os novos dados. Qualquer indicação de ocorrência que aconteça fora dos padrões pré-estabelecidos deve ser verificada. Poderíamos também utilizar a abordagem de *Redes Neurais* para esse tipo de problema.

Método de Abordagem: Redes Neurais

Tipos de Problema: Previsão

Exemplo 3: Previsão de Doenças

Outra aplicação de Data Mining no setor médico está no auxílio de diagnóstico de doenças. O objetivo é que a partir de resposta a alguns parâmetros de entrada que representam sintomas no paciente, seja diagnosticado ou não algum tipo de doença. Isso é possível a partir dos dados de um conjunto de treinamento que possua as características dos sintomas dos pacientes e se houve ou não a manifestação da doença. Com esse conjunto histórico de dados, a abordagem de *Redes Neurais* pode ser utilizada para criar um modelo de previsão de diagnóstico. A partir dos dados de entrada de um novo paciente, com relação à presença ou não dos sintomas existentes, é possível diagnosticar a existência ou não das doenças.

Método de Abordagem: Redes Neurais
Tipos de Problema: Previsão

Exemplo 4: Previsão de Pedidos para Testes de Laboratório

Para pacientes que chegam em Salas de Emergência existe um tempo de espera valioso entre a chegada do médico, a finalização de testes de laboratório solicitados pelo médico e o recebimento dos resultados dos testes. Muitas vezes, pela urgência dos pedidos, são feitos testes desnecessários, ou testes adicionais precisam ser feitos após o recebimento dos primeiros testes. Através da utilização de *Redes Neurais*, pode-se desenvolver um sistema de auxílio na escolha de testes de laboratório a serem feitos, de acordo com diversas características dos pacientes que chegam na Sala de Emergência. Utilizando

dados históricos contendo informações dos pacientes com relação a características demográficas, sociais, e de sintomas dos pacientes, e quais testes foram requeridos e utilizados no passado. De posse dessas informações, uma *Rede Neural* é desenvolvida para previsão de novos pedidos. Com isso, quando novos pacientes chegam na sala de emergência, os dados de entrada são fornecidos e o modelo responde com os testes de laboratório necessários para diagnosticar a situação atual do paciente.

Os dados de entrada poderiam ser: idade, sexo, estado (crítico, sério, rotina), pressão sangüínea, temperatura, medicamentos tomados, dores abdominais, etc. A *Rede Neural* processa essas informações e responde quais tipos de testes laboratoriais devem ser pedidos.

Método de Abordagem: Regras de Associação

Tipos de Problema: Descrição

Exemplo 5: Auxílio no Monitoramento de Tratamento de Pacientes

Regras de Associação podem ser implementadas em aplicações de auxílio no monitoramento da eficácia no tratamento de pacientes. O objetivo é informar aos administradores se os tratamentos ministrados aos diversos tipos de pacientes estão tendo efeito, ou gastos desnecessários ou em tratamentos incorretos estão sendo feitos. De posse de dados históricos contendo informações como: sintomas do paciente, diagnóstico clínico, histórico de internações passadas, intervenções cirúrgicas, etc., e os tratamentos utilizados, é possível obter associações que questionem a eficácia de um determinado tratamento. Um exemplo de associação obtida nesse tipo de aplicação

poderia ser: "Pacientes com diabetes com alta pressão sangüínea e um alto nível de potássio no sangue não reage bem ao tratamento X".

Método de Abordagem: Sistema de Geoprocessamento

Tipos de Problema: Descrição

Exemplo 6: Estudo Demográfico de Atendimento Médico

Com a utilização de sistemas de Geoprocessamento, estudos demográficos de serviços médicos podem ser conduzidos. Através de base de dados demográficas, é possível traçar os potenciais pacientes e contrapor com a oferta de serviços médicos. Estatísticas podem ser geradas descrevendo relações, como por exemplo: "Na região X temos uma taxa de 1 leito hospitalar para cada 130 pacientes". Com isso, de posse de informações que avaliam a distribuição dos serviços, atitudes preventivas podem ser tomadas ou decisões de investimento podem ser melhores direcionadas.

Telecomunicações

Nas empresas de telecomunicações, os estudos usando técnicas de Mineração de Dados, se concentram na detecção de fraudes (celulares, cartões de ligação, e venda ilegal de serviços), além de estudos de segmentação de clientes com o objetivo de fidelisar os clientes. Em países como o Brasil, onde o número de operadoras de telefonia é grande, existe uma forte preocupação em manter os clientes fiéis à empresa. Isso é

necessário pelo serviço prestado, que é indiferente à qual empresa o presta. Aplicações são desenvolvidas para oferecer serviços adicionais direcionados e melhor administrar os grupos de clientes.

Método de Abordagem: Árvores de Decisão

Tipos de Problema: Classificação

Exemplo 1: Detecção de Fraudes

Um dos maiores problemas no setor de telecomunicações está na fraude de seus sistemas. Atualmente, esforços tem sido conduzidos para a construção de modelos que identifiquem a fraude desses sistemas em tempo real, antecipando prejuízos a empresa ou cobranças indevidas a seus clientes. Baseado em informações históricas de fraudes anteriores, modelos utilizando a abordagem de *Árvores de Decisão* são desenvolvidos com o objetivo de classificar os padrões dos serviços telefônicos na probabilidade de ocorrência ou não de fraudes. Com isso, padrões fraudulentos podem ser identificados e evitados previamente. Esse tipo de problema poderia ser tratado também com a abordagem de *Redes Neurais*.

Método de Abordagem: Árvores de Decisão

Tipos de Problema: Segmentação

Exemplo 2: Segmentação de Clientes

Para obter uma maior taxa de retenção de clientes, as empresas de telecomunicações possuem uma estratégia de oferecer serviços adicionais, agregando valor na sua relação com o cliente. No entanto, para otimizar suas estratégias de marketing, necessita de uma categorização do seus clientes com relação a potencialidade de compra de novos serviços. Utilizando a abordagem de *Árvores de decisão*, os clientes são segmentados em categorias potenciais para cada tipo de serviço a partir de diversas informações do cliente, de seus serviços contratados atuais, informações demográficas, e histórico anterior. Essa aplicação ocorre de forma semelhante que no setor varejista. Após a modelagem, podemos identificar os clientes potenciais para determinado serviço, direcionando as estratégias para esse s grupos de clientes. Isso aumenta a taxa de resposta à oferta de novos serviços e fortalece a ligação entre o cliente e a empresa.

Método de Abordagem: Técnicas Estatísticas

Tipos de Problema: Estimação

Exemplo 3: Estimação de Índice de Fidelidade dos Clientes

Uma das preocupações no setor de telecomunicações está na perda de clientes atuais para outras prestadoras de serviços concorrentes. Isso pode ocorrer pelas diferenças entre os valores das taxas cobradas ou os serviços agregados. Uma interessante aplicação está na estimação de um índice que informe a probabilidade do cliente trocar de prestadora de serviço. Aplicando *Técnicas Estatísticas* de correlação, visualização e clusterização, em bancos de dados que possuam histórico de clientes e sua eventual

escolha por outra operadora, é possível desenvolver modelos para estimar a taxa de fidelização dos clientes.

Dessa forma, a partir das características de um novo cliente e de sua performance atual no consumo de serviços, um índice é gerado relacionando suas características com a probabilidade de mudança de operadora. De posse dessa informação, as empresas podem tornar estratégias de marketing preditivas ou corretivas, com o objetivo de reter os consumidores antigos.

Método de Abordagem: Métodos Analíticos de Visualização
Tipos de Problema: Descrição

Exemplo 4: Identificação de Causas de Perda de Clientes

Para melhora compreender as causas de perdas de clientes, as empresas de telecomunicações tem usado métodos de visualização para determinação dos fatores que causas a perda de fidelidade. Com informações como: características demográficas do cliente, localização geográfica, tipo de contrato, utilização de serviços, padrões de uso de serviço, número de ligações ao serviço de atendimento e se esse cliente foi perdido ou não, estatísticas são visualizadas e analisadas. Descrições como: " Clientes na região X que utilizam freqüentemente nosso serviço telefônico durante o dia estão deixando nossa empresa" ou " Clientes no estado X está trocando de companhia 3 vezes mais que no estado Y" são geradas. Com a identificação desses tipos de relações, podendo ser até em tempo real, a empresa pode tomar medida para reter esse cliente, evitando que troquem de operadora telefônica..

Outras Áreas

Com os exemplos mencionados anteriormente, não temos a pretensão de ter atingido todos os casos já utilizados em Mineração de Dados, mas de ser um ponto de partida de classificação sobre suas maiores aplicações.

Durante a pesquisa algumas outras aplicações foram obtidas em outras áreas do conhecimento. Algumas aplicações tem sido usadas para auxílio nas detecção de crimes. A partir da descrição de suspeitos, *Árvores de Decisão* ou R*edes Neurais* tem sido utilizadas para indicar o maior provável do crime.

Na área esportiva, o uso de *Regras de Associação* para obtenção de relações descritivas tem sido amplamente utilizado. Com a massiva coleta de estatísticas sobre jogadores e outros índices de performance do time, descrições são geradas que podem guiar a troca de jogadores ou mudanças de estratégias durante o jogo.

Na área de meio ambiente, determinando o risco relacionados a usinas nucleares, determinação do impacto ambiental de instalação de fábricas em uma determinada região, estudo de difusão de poluentes.

Na Internet, fazendo a identificação de determinados padrões de *home page*, busca e agrupamento de documentos.

E no nosso caso estamos propondo o uso de técnicas de Mineração na imputação de dados.

Exemplos de Ferramentas de *Data Mining*

O mercado atual oferece diversas ferramentas de *Data Mining* , onde cada uma é orientada para realizar determinadas tarefas de "mineração", utilizando diferentes algoritmos para isto (redes neurais, algoritmos genéticos, árvores de decisão, entre outros). Nosso conjunto de pesquisa, foi baseado na análise das informações dos fabricantes e a indiscutível contribuição da Internet na análise de páginas de provedores de software, páginas acadêmicas sobre o assunto, páginas de organizações dedicadas ao estudo do tema e diversos estudos de caso. Todo material relevante foi inserido nas referências.

Agrupamos os software em grupos de acordo com suas características comuns, com referências de seus fornecedores e breve resumo de suas funções:

Podemos dividir em 7 grupos principais:

- **Suítes**: são pacotes que auxiliam todo o processo de Descoberta do Conhecimento, desde interfaces para acesso a banco de dados, manipulação de dados, desenvolvimento do modelo e auxílio na análise dos resultados. Tais pacotes podem utilizar diferentes abordagens.
- **Software de Classificação**: Esse grupo de software possui característica de gerar modelos de classificação a partir de um banco de dados. Podem utilizar diferentes abordagens como classificação múltipla, árvores de decisão, regras de associação, redes neurais, e estatísticas bayesiana.
- **Software de Clustering**: Utilizados para a determinação de *clusters* e subgrupos do conjunto de dados.

- **Software de Estatística, estimação e Regressão**: Auxiliam na condução de tarefas usuais de estatística, estimação de parâmetros e determinação de curvas de regressão. Pode se dividir em aplicações de estimação não-linear e software de estatística clássica e regressão.
- **Software de Análise de Relacionamento e Associações**: Utilizados para a determinação de regras de associações e construção de redes de relacionamentos.
- **Software de Padrões Seqüenciais**: Aplicações utilizadas para a descoberta de padrões seqüenciais no banco de dados.
- **Software de Visualização**: Os aplicativos para visualização podem ser subdivididos em Aplicativos para descoberta e Aplicativos para visualização científica e estatística.
- **Software de Transformação de Dados e Limpeza**: Aplicativos utilizadas para tratamento e limpeza dos dados.

TABELA 2 - Exemplos de ferramentas de *Data Mining* disponíveis no mercado

Software	Tipo	Fornecedor	Internet
Alice d´Isoft	Árvore de Decisão, Suite	Isoft	www.alice.fr/products/alice.html
Analytica	Estatística Bayesiana	Lumina Decision Systems	www.lumina.com/
AT-Sigma Data Chopper	Estatística Bayesiana	AT-Sigma	www.atsigma.com/
Data Desk	Estatística Clássica e Regressão, Visualização Científica e Estatítica	Data Description Inc.	www.datadesk.com/
DataDetective	Suite	Sentient Machine Research	www.smr.nl/
Data Manager	Transformação de Dados e Limpeza	Joe Spanicek	www.ggmate.com/DataManagerSoftware/
DataMind DataCruncher		DataMind	www.datamindcorp.com/

DBMiner	Suite	DBMiner Technology Inc.	www.dbminer.com/
Decision House	Classificação Múltipla, Árvore de Decisão	Quadstone Inc.	www.quadstone.com/
DEXTR's Data EXTRactor	Transformação de Dados e Limpeza	DogHouse Enterprises	www.dataextractor.com/
Enterprise Miner	Suite, Padrões Seqüênciais	SAS Institute	www.sas.com/
Ergo	Estatítica Bayesiana	Noetic Systems	www.noeticsystems.com/
Genio	Transformação de Dados e Limpeza	Hummingbird	www.leologic.com/
Hugin	Estatítica Bayesiana	Hugin Expert	www.hugin.com/
IBM Intelligent Miner for Data	Suite, Análise de Relacionamentos e Associações, Padrões Seqüênciais	IBM	www.software.ibm.com/data/iminer/
IDIS Data Mining Suite	Suite	IDIS	www.dataminig.com/dmsuite.htm
IDL	Visualização Científica e Estatística	Research Systems Inc.	www.rsinc.com/

Continua

Software	Tipo	Fornecedor	Internet
IND v2.0	Árvore de Decisão	Cosmic	ic-www.arc.nasa.gov/fia/projects/bayes-group/group/ind/IND-program.htm
JWAVE	Visualização para Descoberta	Visual Numerics	www.vni.com/index.html
K-wiz	Suite	K.wiz Solutions Limited	www.kwiz-solutions.com/
KATE-tools	Árvore de Decisão	Acknosoft	www.acknosoft.com
Kepler	Suite	Dialogis Software	www.dialogis.de/
Knowledge SEEKER	Árvore de Decisão	Angoss	www.angoss.com/
Knowledge Studio	Suite, Classificação Múltipla	Angoss	www.angoss.com/
Magnify PATTERN	Suite	Magnify	www.magnify.com/productsframe.htm
MARS	Estimação não-linear	Salford Systems	www.salford-systems.com/products-mars.html

Mathematica	Visualização Científica e Estatística	Wolfram Research	www.wolfram.com/
MATLAB	Redes Neurais, Estatítica Clássica e Regressão	Math Works Inc.	www.mathworks.com
MineSet	Suite, Análise de Relacionamentos e Associações	Silicon Graphics	www.sgi.com/software/mineset
ModelQuest	Classificação Múltipla, Redes Neurais	Abtech	www.abtech.com
NeoVista Decision Series	Suite, Padrões Seqüênciais	Neovista	www.neovista.com/
Netica	Estatítica Bayesiana	Norsys	www.norsys.com/netica.html
Net-ID	Padrões Seqüênciais	Net-ID Inc.	www.netid.com/
NETMAP	Visualização para Descoberta	Alta Analytics	www.altaanalytics.com

Continua

Software	Tipo	Fornecedor	Internet
NeuralWorks Professional II/Plus	Redes Neurais	NeuralWare Inc.	www.neuralware.com
NeuroSolutions	Redes Neurais	NeuroDimension	www.nd.com/products/nsv3.htm
NewView	Transformação de Dados e Limpeza	SPSS	www.spss.com/software/spss/newview/
Partek	Suite	Partek	www.kdnuggets.com/sift/partek.html
Pilot Decision Support Suite		Pilot	www.pilotsw.com/
Polyanalyst	Suite	Megaputer Intellingence	www.megaputer.com
Preclass	Árvore de Decisão	Prevision Inc.	www.prevision.com
Previa	Classificação	Elseware	www.elseware.fr
Classpad	Múltipla, Estimação não-linear		
Proforma	Redes Neurais	Neural Innovation Ltd	www.neural.co.uk/
PRW	Suite, Redes Neurais	Única	www.única-usa.com/
Pilot Discovery Series	Suite	Pilot Software	www.pilotsw.com/products/discover.htm

PV-Wave	Visualização Científica e Estatística	Visual Numerics	www.vni.com/
PVE	Visualização Científica e Estatística	IBM	www.ibm.com/News/950203/pve-01.html
Relational Tools	Transformação de Dados e Limpeza	Princenton Softech	www.princetonsoftech.com/products/os-relationaltools.htm
SAS	Estatítica Clássica e Regressão	SAS	www.sas.com/
SOMine	Clustering, Visualização para Descoberta	Eudaptics	www.eudaptics.co.at/
Sphinx Vision	Visualização para Descoberta	ASOC	www.asoc.com/
Spotfire	Visualização para Descoberta	Spotfire Inc.	www.ivee.com

Continua

Software	Tipo	Fornecedor	Internet
SPSS	Suite	SPSS	www.spss.com/software/spss/newview/
SPSS AnswerTree	Árvore de Decisão	SPSS	www.psss.com/software/spss/AnswerTree/
SRA KDD Toolset	Suite, Padrões Seqüênciais	SRA International	www.knowledgediscovery.com
SPSS SigmaStat	Estatística Clássica e Regressão	SPSS	www.spss.com/
STATlab	Estatística Clássica e Regressão, Visualização Científica e Estatítica	SLP Infoware	www.slp.fr
STORM	Suite	Elseware	www.elseware.fr/
SuperQuery	Regras de Associação	AZMY Thinkware	www.azmy.com/
TETRAD II	Estatística Bayesiana	Carnegie Mellon University	www.erlbaum.com/558.htm
TGS 3-D-MasteSuite	Visualização Científica e Estatística	TGS	www.tgs.com/
Thinx	Visualização para Descoberta	Thinx	www.thinx.com/
VDI Discovery for Developers	Visualização para Descoberta	Visible Decisions	www.vdi.com/

Visual Insights	Visualização para Descoberta	Visual Insights	www.lucent.com/visualinsights/
VisualMine	Visualização para Descoberta	AIS	www.visualmine.com/
VRCharts	Visualização para Descoberta	AlterVue Systems	www.vrcharts.com/
WINROSA	Regras de Associação	MIT Management Intelligenter Technologien	www.mitgmbh.de
WizWhy	Regras de Associação	WizSoft	www.wizsoft.com/
XpertRule Miner	Suite, Árvore de Decisão, Análises de Relacionamentos e Associações, Transformação de Dados e Limpeza	Attar Software	www.attar.com/
Zoom'n View	Suite	Skygate International	www.skygate.dk

MINISTÉRIO DA ECONOMIA, FAZENDA E PLANEJAMENTO

FUNDAÇÃO INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA — IBGE

X RECENSEAMENTO GERAL DO BRASIL - 1991

CENSO DEMOGRÁFICO

CD 1.02 - QUESTIONÁRIO DA AMOSTRA

1

1 MUNICÍPIO	2 FAIXA	3 Nº NA FAIXA

PARA USO DO ÓRGÃO CENTRAL

4 DISTRITO	5 SUBDISTRITO	6 Nº DO SETOR	7 QUARTEIRÃO	8 FACE	9 Nº NO CD 1.07	10 Nº NO CD 1.03	PESSOAS RESIDENTES 11 Masculino	PESSOAS RESIDENTES 12 Feminino	13 INFORMANTE	14 QUESTIONÁRIO SUPLEMENTAR
										Não tem ☐ Tem ☐ É ☐

LOCALIDADE ---------- LOGRADOURO ---------- Nº ---------- DEPENDÊNCIA ----------

NOME DO INFORMANTE ---------- ASSINATURA DO INFORMANTE ----------

2 CARACTERÍSTICAS DO DOMICÍLIO

1 ESPÉCIE

Particular
- 1 Permanente
- 2 Improvisado

3 Coletivo

(Os quesitos seguintes só serão preenchidos para o domicílio particular permanente)

2 LOCALIZAÇÃO

Casa
- 1 Isolada ou de condomínio
- 2 Em conjunto residencial popular
- 3 Em aglomerado subnormal

Apartamento
- 4 Isolado ou de condomínio
- 5 Em conjunto residencial popular
- 6 Em aglomerado subnormal

7 Cômodo(s)

3 PAREDES

- 1 Alvenaria
- 2 Madeira aparelhada
- 3 Taipa não revestida
- 4 Material aproveitado
- 5 Palha
- 6 Outro

4 COBERTURA

- 1 Laje de concreto
- 2 Telha de barro
- 3 Telha de cimento-amianto
- 4 Zinco
- 5 Madeira aparelhada
- 6 Palha
- 7 Material aproveitado
- 8 Outro

5 ABASTECIMENTO DE ÁGUA

Com canalização interna
- 1 Rede geral
- 2 Poço ou nascente
- 3 Outra forma

Sem canalização interna
- 4 Rede geral
- 5 Poço ou nascente
- 6 Outra forma

INSTALAÇÃO SANITÁRIA

6 ESCOADOURO

Fossa séptica
- 1 Rede geral
- 2 Ligada à rede pluvial
- 3 Sem escoadouro
- 4 Fossa rudimentar
- 5 Vala negra
- 6 Outra
- 7 Não sabe
- 0 Não tem

7 BANHEIRO
- Só do domicílio
- 2 Comum a mais de um
- 0 Não tem

8 CONDIÇÃO DE OCUPAÇÃO

Própria
- 1 A construção e o terreno
- 2 Só a construção
- 3 Alugado

Cedido
- 4 Por empregador
- 5 Por parentesco
- 6 Outra

9 ALUGUEL MENSAL

- 0 Não paga
- Nº de dígitos
- Cr$

10 COMBUSTÍVEL USADO PARA COZINHAR
- 1 Gás canalizado
- 2 Só gás de botijão
- 3 Só lenha
- 4 Gás de botijão e lenha
- 5 Carvão
- 6 Outro
- 0 Não sabe ou não cozinha

11 TOTAL DE CÔMODOS

Número de cômodos

(Quando o número de cômodos for inferior a 10, registrar 0 no primeiro campo)

12 CÔMODOS SERVINDO DE DORMITÓRIO
- 1 1 cômodo
- 2 2 cômodos
- 3 3 cômodos
- 4 4 cômodos
- 5 5 cômodos
- 6 6 cômodos
- 7 7 cômodos
- 8 8 cômodos
- 0 9 cômodos ou mais

13 BANHEIROS
- 1 1 banheiro
- 2 2 banheiros
- 3 3 banheiros
- 4 4 banheiros
- 5 5 banheiros ou mais
- 0 Não tem

14 DESTINO DO LIXO

Coletado
- 1 Diretamente
- 2 Indiretamente
- 3 Queimado

Jogado
- 4 Enterrado
- 5 Terreno baldio
- 6 Rio, lago ou mar
- 7 Outro

15 NESTE DOMICÍLIO RESIDE CRIANÇA COM MENOS DE 2 ANOS, INCLUSIVE ALGUMA RECÉM-NASCIDA?
- 1 Sim
- 0 Não

16 FILTRO DE ÁGUA
- 1 Tem
- 0 Não tem

17 TELEFONE
- 1 1 linha
- 2 2 ou mais linhas
- 0 Não tem

18 AUTOMÓVEL PARTICULAR
- 1 1 carro
- 2 2 carros
- 3 3 ou mais carros
- 0 Não tem

19 AUTOMÓVEL PARA TRABALHO
- 1 Próprio
- 2 Cedido
- 0 Não tem

20 RÁDIO
- 1 Tem
- 0 Não tem

21 ILUMINAÇÃO

Elétrica
- 1 Com medidor
- 2 Sem medidor
- 3 Óleo ou querosene
- 4 Outra

(Os quesitos seguintes só serão preenchidos quando houver iluminação elétrica)

22 GELADEIRA
- 1 1 porta
- 2 Mais de 1 porta
- 0 Não tem

23 TELEVISÃO PRETO E BRANCO
- 1 Tem
- 0 Não tem

24 TELEVISÃO EM CORES
- 1 1 aparelho
- 2 2 aparelhos
- 3 3 ou mais aparelhos
- 0 Não tem

25 FREEZER
- 1 Tem
- 0 Não tem

26 MÁQUINA DE LAVAR ROUPA
- 1 Tem
- 0 Não tem

27 ASPIRADOR DE PÓ
- 1 Tem
- 0 Não tem

3 1.ª PESSOA — NOME

01 Sexo — 1 Masculino; 2 Feminino

02 Parentesco ou relação com o Chefe do domicílio — 01 Chefe; 20 Individual

03 Parentesco ou relação com o Chefe da família — 01 Chefe; 20 Individual

04 Família a que pertence — 1 Única; 2 Domicílio coletivo; Convivente: 3 1ª; 4 2ª; 5 3ª; 6 4ª; 7 5ª

05 Se a mãe reside no domicílio, indique o número de ordem em que foi recenseada. Se não reside, indique se está viva, falecida ou não sabe — Nº da pessoa (Reside); 70 Está viva; 80 Falecida; 90 Não sabe (Não reside)

06 Mês e ano de nascimento (se não souber o mês e/ou o ano preencher a questão seguinte) — Mês; Ano

07 Idade presumida (se inferior a 1 ano, o número de meses) — 1 ano ou mais (20); Menos de 1 ano (400)

08 Faixa de idade — 1 Menos de 5 anos; 2 De 5 a 9 anos; 3 10 anos ou mais

09 Raça ou cor (amarela só para as pessoas de origem oriental) — 1 Branca; 2 Preta; 3 Amarela; 4 Parda; 5 Indígena

10 Religião ou culto

11 Deficiência física ou mental — 1 Cegueira; 2 Surdez; 3 Paralisia de um dos lados; 4 Paralisia das pernas; 5 Paralisia total; 6 Falta de membro(s) ou parte dele(s); 7 Deficiência mental; 8 Mais de uma; 0 Nenhuma das enumeradas

12 Neste Município morou — 1 Só na zona urbana; 2 Só na zona rural; 3 Nas zonas urbana e rural

13 Se no Quesito 12 assinalou o retângulo 3, indique há quantos anos se deu a última mudança — Nº de anos; 98 Nunca mudou

14 Nasceu neste Município (se assinalou o retângulo 1, passe para o Quesito 23. Se assinalou o 2, passe para o Quesito 17. Se assinalou o 3, responda aos Quesitos seguintes) — 1 Sim e sempre morou neste; 2 Sim, mas já morou em outro; 3 Não nasceu

15 Se Naturalizado Brasileiro ou Estrangeiro, indique o ano em que teve residência no País — 100 Brasileiro nato; 2 Naturalizado brasileiro; 3 Estrangeiro

16 Unidade da Federação ou País estrangeiro de nascimento

17|18 Há quantos anos mora sem interrupção — Nesta Unidade da Federação (Nº de anos); Neste Município (Nº de anos)

PARA A PESSOA QUE NO QUESITO 17/18 RESPONDEU MENOS DE 10 ANOS, NO CAMPO — NESTE MUNICÍPIO

19 Indique a sigla da UF e o nome do Município ou do País estrangeiro em que morava antes de mudar-se para este Município — Sigla; Nome

20 Na localidade indicada no Quesito 19, residia — 1 Na zona urbana; 2 Na zona rural

OS QUESITOS SEGUINTES SÓ SERÃO PREENCHIDOS PARA A PESSOA DE 5 ANOS OU MAIS (nascida antes de 01/09/1986)

21 Indique a sigla da UF e o nome do Município ou do País estrangeiro em que residia em 01/09/1986 (se residia no Município, assinale o retângulo 7 — Neste) — Sigla; nome; 7 Neste

22 Na localidade indicada no Quesito 21, antes de mudar, residia — 1 Na zona urbana; 2 Na zona rural

23 Sabe ler e escrever ou Não sabe — 1 Sabe ler e escrever; 2 Não sabe

PARA A PESSOA QUE FREQÜENTA ESCOLA (Se não freqüenta nenhum assinale os retângulos 0)

24 Série que freqüenta (se não freqüenta curso seriado, assinale o retângulo 0 — Nenhuma) — 1 1ª série; 2 2ª série; 3 3ª série; 4 4ª série; 5 5ª série; 6 6ª série; 7 7ª série; 8 8ª série; 0 Nenhuma

25 Grau da série que freqüenta (se não freqüenta curso seriado, assinale o retângulo 0 — Nenhum) — 1 1º grau; 2 2º grau; 3 Superior; 4 1º grau; 5 2º grau; 0 Nenhum (Supletivo seriado)

26 Para a pessoa que freqüenta curso não seriado (se freqüenta curso seriado, assinale o retângulo 0 — Nenhum) — 1 Pré-escolar; 2 Curso de alfabetização de adultos; 3 1º grau; 4 2º grau (Supletivo não seriado); 5 Pré-vestibular; 6 Alfabetização ou Doutorado; 0 Nenhum

PARA A PESSOA QUE NÃO FREQÜENTA ESCOLA (se freqüenta assinale os retângulos 0)

27 Última série que concluiu com aprovação (Para a pessoa que concluiu cursos de alfabetização de adultos, assinale o declarado, assinale o retângulo 0 — Nenhuma) — 1 1ª série; 2 2ª série; 3 3ª série; 4 4ª série; 5 5ª série; 6 6ª série; 7 7ª série; 8 8ª série; 9 Nunca freqüentou; 0 Nenhuma

28 Grau da última série que concluiu com aprovação — 1 Curso de alfabetização de adultos; 2 Primário ou elementar; 3 Ginasial ou médio 1º ciclo; 4 1º grau; 5 2º grau; 6 Colegial ou médio 2º ciclo; 7 Superior; 8 Mestrado ou Doutorado; 0 Nenhum

OS QUESITOS SEGUINTES SÓ SERÃO PREENCHIDOS PARA A PESSOA DE 10 ANOS OU MAIS (nascida antes de 01/09/1981)

29 Espécie do curso mais elevado concluído com aprovação

30 Vive ou viveu em companhia de cônjuge — esposo(a), companheiro(a), etc. (Se assinalou o retângulo 2 — Não, passe para os Quesitos 35 ou 45, conforme o caso) — 1 Sim; 2 Não

31 Se no Quesito 30 assinalou o retângulo 1 — Sim, indique o mês e o ano em que contraiu a primeira união — Mês; Ano; 13 Não lembra

32 Se vive em companhia de cônjuge — esposo(a), companheiro(a), etc., indique a natureza da união — 1 Casamento civil e religioso; 2 Só casamento civil; 3 Só casamento religioso; 4 União consensual

33 Se não vive em companhia de cônjuge — esposo(a), companheiro(a), etc., indique se é — 5 Separado(a) não judicialmente; 6 Desquitado(a) ou separado(a) judicialmente; 7 Divorciado(a); 8 Viúvo(a)

34 Se respondeu o Quesito 32, indique o mês e o ano em que passou a viver com o cônjuge atual. Se respondeu o Quesito 33, indique o mês e o ano em que passou a viver a situação indicada — Mês; Ano; 13 Não lembra

OS QUESITOS 35 A 44 SÓ SERÃO PREENCHIDOS PARA AS MULHERES (ao teve filho(s) só do sexo errado registrar no campo correspondente ao outro sexo — 00)

35|36 Filhos(as) tidos(as) que moram no domicílio — 7 Não tem; Homem; Mulher

37|38 Filhos(as) tidos(as) que moram em outro domicílio — 7 Não tem; 8 Não sabe; Homem; Mulher

39|40 Filhos(as) tidos(as) nascidos(as) vivos(as) que já morreram (não incluir tidos(as) nascidos(as) mortos(as)) — 7 Não tem; 8 Não sabe; Homem; Mulher

41|42 Filhos(as) tidos(as) nascidos(as) mortos(as) (gestação não inferior a 1 mês sem nenhum sinal de vida ao nascer). Não incluir abortos — 7 Não teve; 8 Não sabe; Homem; Mulher

43|44 Sexo, mês e ano de nascimento do último tido(a) nascido(a) vivo(a) — 7 Não tem; Mês e ano; 1 Homem; Mês e ano; 2 Mulher

45 Trabalhou em todos ou em parte dos últimos 12 meses (01/09/1980 a 31/08/1981). Se assinalou o retângulo 3 — Não trabalhou, passe para o Quesito 55 — 1 Habitualmente; 2 Eventualmente (Trabalhou); 3 Não trabalhou

46 Qual a ocupação, profissão, cargo, função etc., que exerceu nos últimos 12 meses ou em parte deles (se exerceu duas ou mais ocupações, registre a que considera principal. Se mudou de ocupação, com a mesma definição, registre a ocupação atual)

47 Setor ou ramo de atividade, negócio, organização, instituição, etc., em que exerceu a ocupação declarada no Quesito 46

48 Tipo de estabelecimento, negócio, instituição, etc., onde exerceu a ocupação declarada no Quesito 46

49 Relação de trabalho ou a posição da ocupação no estabelecimento, negócio, instituição, etc., declarada no Quesito 48 — 01 Trabalhador agrícola volante; 02 Empregado; 03 Conta própria; 04 Empresário; 05 Conta própria (Trabalhador doméstico); 06 Empregado do setor privado; 07 Servidor público; 08 De empresa estatal (Empregado do setor público); 09 Conta própria; 10 Empregador; 11 Sem remuneração (Parceiro ou meeiro)

50 Tem carteira de trabalho assinada — 1 Sim; 2 Não sabe; 3 Não tem; 4 Não é empregado

51 Quantas pessoas trabalham no estabelecimento, negócio, instituição, etc. declarado no Quesito 48 — 1 1 ou 2; 2 3 ou 4; 3 5 a 9; 4 10 ou mais (Nº de empregados); 5 Trabalha sozinho; 6 Com sócio ou não remunerado; 7 Trabalhador doméstico; 8 Não sabe

52 Local de trabalho onde exerceu a ocupação declarada no Quesito 46 — 1 Sem local exclusivo; 2 Com local exclusivo; 3 Com equipamento pesado; 4 Com equipamento leve ou sem (No domicílio); 5 Propriedade agropecuária; 6 Empresa ou firma; 7 Em casa de clientes ou patrão; 8 Outro (Via pública)

53 É contribuinte do Instituto de Previdência Pública — 1 Sim; 2 Não sabe; 3 Não é

54|55 Horas habitualmente trabalhadas por semana (quando exercer só uma ocupação, registrar 00 no campo Em outras ocupações) — Na ocupação do Quesito 46; Em outras ocupações

56 Rendimento bruto do mês de agosto de 1981 da ocupação declarada no Quesito 46 — 0 Não tem; Nº de dígitos; Cr$,00

57 Rendimento bruto do mês de agosto de 1981 de outra(s) ocupação(ões) — 0 Não tem; Nº de dígitos; Cr$,00

58 Se no Quesito 45 assinalou Não trabalhou, indique a situação ou ocupação que tem, observada a ordem enumerada. Se Trabalhou, não preencha este quesito e passe para o Quesito 59. — 1 Já trabalhou; 2 Nunca trabalhou (Procurando trabalho); 3 Aposentado; 4 Pensionista; 5 Vive de rendas; 6 Doente; 7 Estudante; 8 Doente ou inválido; 9 Afazeres domésticos; 0 Sem ocupação

59 Indique se é aposentado e/ou pensionista — 0 Não é; 1 Aposentado; 2 Pensionista; 3 Aposentado e pensionista

60 Rendimento mensal bruto do mês de agosto de 1981, proveniente de aposentadoria e/ou pensão — 0 Não tem; Nº de dígitos; Cr$,00

61 Rendimento bruto do mês de agosto de 1981, ou média mensal de outros rendimentos que não se enquadrem em nenhuma das categorias anteriormente relacionadas — aluguéis, arrendamentos, doações, emprego de capital, etc. — 0 Não tem; Nº de dígitos; Cr$,00