UNIVERSIDADE REGIONAL DE BLUMENAU

CENTRO DE CIÊNCIAS EXATAS E NATURAIS

CURSO DE CIÊNCIAS DA COMPUTAÇÃO

(Bacharelado)

SISTEMA DE APOIO À DECISÃO PARA PREVISÕES GENÉRICAS UTILIZANDO TÉCNICAS DE DATA MINING

TRABALHO DE CONCLUSÃO DE CURSO SUBMETIDO À UNIVERSIDADE REGIONAL DE BLUMENAU PARA A OBTENÇÃO DOS CRÉDITOS NA DISCIPLINA COM NOME EQUIVALENTE NO CURSO DE CIÊNCIAS DA COMPUTAÇÃO — BACHARELADO

RICARDO KREMER

BLUMENAU, JUNHO/1999

1999/1-51

SISTEMA DE APOIO À DECISÃO PARA PREVISÕES GENÉRICAS UTILIZANDO TÉCNICAS DE DATA MINING

RICARDO KREMER

ESTE TRABALHO DE CONCLUSÃO DE CURSO, FOI JULGADO ADEQUADO PARA OBTENÇÃO DOS CRÉDITOS NA DISCIPLINA DE TRABALHO DE CONCLUSÃO DE CURSO OBRIGATÓRIA PARA OBTENÇÃO DO TÍTULO DE:

BACHAREL EM CIÊNCIAS DA COMPUTAÇÃO

_____

Prof. Maurício Capobianco Lopes — Orientador na FURB

_____

Prof. José Roque Voltolini da Silva — Coordenador do TCC

BANCA EXAMINADORA

_____

Prof. Maurício Capobianco Lopes

_____

Prof. Everaldo Artur Grahl

_____

Prof. Ricardo Guilherme Radünz

À FURB, AO MEU ORIENTADOR MAURÍCIO CAPOBIANCO LOPES E A TODOS QUE CONTRIBUÍRAM DIRETA OU INDIRETAMENTE PARA A REALIZAÇÃO DESTE TRABALHO.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

KDD - *Knowledge Discovery in Databases*

MBR - *Memory-Based Reasoning*

MPP - Processador maciçamente paralelo

OLAP - *On Line Analytic Processing*

OLTP - *On Line Transaction Processing*

SAD - Sistema de Apoio à Decisão

SAE - Sistema de Automação de Escritórios

SE - Sistema Especialista

SI - Sistema de Informação

SIE - Sistema de Informações Executivas

SIG - Sistema de Informações Gerenciais

SMP - Multiprocessador simétrico

SPT - Sistema de Processamento de Transações

RESUMO

O trabalho tem como objetivo principal auxiliar a tomada de decisões através de um Sistema de Apoio à Decisão utilizando técnicas de Data Mining, mais especificamente para efetuar previsões genéricas. Para auxiliar esta tarefa foi implementado um protótipo que permite ao usuário definir um modelo de previsão, onde o mesmo pode ser treinado para responder às variáveis de previsão com certa flexibilidade. Para a elaboração do protótipo, foram analisadas as características de Sistemas de Informação, bem como dos Data Warehouses e das técnicas de Data Mining. Também estudou-se na área de Inteligência Artificial as Redes Neurais, que fazem parte do Data Mining aplicado à previsão. Como conseqüência do desenvolvimento deste trabalho, verificou-se que a aplicação do Data Mining juntamente com as etapas do KDD foi muito eficiente. Foram realizados testes e foi possível desenvolver modelos de previsão onde colocou-se em prática o uso de Redes Neurais, a qual se mostrou bastante eficiente para o processo de previsão principalmente por sua grande capacidade de generalização.

ABSTRACT

The main purpose of this work is to help decision making through a Decision Support System using Data Mining techniques, specifically to do generic forecasts. To aid in this task, a prototype was implemented that allows the user to define a forecast model, where the same can be trained to answer the forecast variables with certain flexibility. For the elaboration of the prototype, the characteristics of Information Systems were analyzed, as well as the Data Warehouses and Data Mining techniques. In the area of Artificial Intelligence, the Neural Networks were also studied, once they are part of Data Mining applied to the forecast. As a consequence of the development of this work, it was confirmed that the use of data Mining along the stages of KDD was very eficient. Tests were conducted and it was possible to develop models of forecast and the use of Neural Networks was put to practice, which proved eficient enough for the process of forecast especially for its great capacitance of generalization.

1 INTRODUÇÃO

1.1 ORIGEM/MOTIVAÇÃO

A cada ano, companhias acumulam mais e mais informações em seus bancos de dados. Como conseqüência, estes bancos de dados passam a conter verdadeiros tesouros de informação sobre vários dos procedimentos dessas companhias. Toda esta informação pode ser usada para melhorar seus procedimentos, permitindo que a empresa detecte tendências e características disfarçadas, e reaja rapidamente a um evento que ainda pode estar por vir. Alguns exemplos disto são o crescimento dos mecanismos de leitura de preço nos supermercados, dos caixas eletrônicos, dos cartões de crédito, da televisão por assinatura, do *home shopping*, da transferência eletrônica de fundos, o processamento automático de pedidos, das bilheterias eletrônicas e outros.

Mas não se tem somente uma grande quantidade de dados sendo produzida; os dados também estão sendo extraídos cada vez mais dos sistemas de onde são gerados e alimentados em um *Data Warehouse*, tornando-se parte da memória da empresa [HAR98].

Segundo [INM97], um *Data Warehouse* "é um conjunto de dados baseado em assuntos, integrado, não-volátil, e variável em relação ao tempo, de apoio às decisões gerenciais".

Com toda esta informação disponível, seja em um *Data Warehouse* ou simplesmente em uma Base de Dados, tende a crescer cada vez mais a demanda por sistemas que forneçam informações precisas, que respondam às dúvidas da empresa e que proporcionem uma tomada de decisões mais acertada. Um tipo de sistema que possui estes aspectos é o Sistema de Apoio à Decisão (SAD).

Conforme [SPR91], os Sistemas de Apoio à Decisão "são sistemas computacionais que ajudam os responsáveis pela tomada de decisões a enfrentar problemas estruturais através de uma interação direta com modelos de dados e análises".

No entanto, apesar do enorme valor desses dados, a maioria das organizações é incapaz de aproveitar totalmente o que está armazenado em seus arquivos. Esta informação preciosa está na verdade implícita, escondida sob uma montanha de dados, e não pode ser descoberta

utilizando-se sistemas de gerenciamento de banco de dados convencionais; elas precisam de um significado. O significado permite a análise dos dados observando modelos, estabelecendo mecanismos e tendo novas idéias para fazer previsões sobre o futuro. Com o uso do *Data Mining* pode-se acrescentar significado a esses dados.

Conforme [HAR98], o *Data Mining*, do modo como é usado o termo, é a exploração e análise, por meios automáticos ou semi-automáticos, de grandes quantidades de dados para descobrir modelos e regras significativas.

O *Data Mining* pode ser aplicado à tarefas como classificação, estimativas, previsões, agrupamento por afinidades ou reunião. Algumas destas técnicas são melhor executadas através de "testes hipotéticos", onde os dados do passado são verificados para aprovar ou não idéias ou suposições obtidas em cima dos dados disponíveis. Além destes "testes hipotéticos" pode ser utilizado também a "descoberta de conhecimento", onde os dados falam por si próprios. Este processo de "descoberta de conhecimento" pode ser de duas formas: a direcionada e a não-direcionada. A forma direcionada tenta explicar ou categorizar alguns campos de dados, como receitas ou respostas. A descoberta de conhecimento não-direcionada tenta encontrar modelos ou similaridades entre grupos de registros sem um campo-alvo específico ou um conjunto de classes predefinidas [HAR98].

O que o *Data Mining* faz é encontrar modelos interessantes nos dados. Mas não é só isso, deve-se poder agir aos modelos, em última análise, transformando os dados em informações, as informações em ação, e a ação em valores.

Reconhecendo o *Data Mining* como uma forma de incorporar significado aos dados, propõe-se especificar e desenvolver um Sistema de Apoio à Decisões para previsões genéricas utilizando as técnicas de *Data Mining*.

A quantidade de informação armazenada em bancos de dados está explodindo, e ultrapassa a habilidade técnica e a capacidade humana na sua interpretação. De compras através de cartões de crédito a imagens pixel-a-pixel de galáxias, bancos de dados são medidos hoje em gigabytes e terabytes. A necessidade de transformar estes terabytes de dados em informações significativas é óbvia. Felizmente, técnicas computacionais foram

desenvolvidas para analisar os dados, ou ao menos ajudar o analista a encontrar ordem no caos, ou seja, conhecimento.

Data Mining é uma tecnologia usada para revelar informação estratégica escondida em grandes massas de dados. É usada em diversas áreas, como análise de riscos, marketing direcionado, controle de qualidade, análise de dados científicos, etc. *Data Mining* define o processo automatizado de captura e análise de enormes conjuntos de dados, para então extrair um significado. Esta tecnologia está sendo usado para descrever características do passado, assim como predizer tendências para o futuro. Sua utilização permite avanços tecnológicos e descobertas científicas, além de garantir uma vantagem competitiva invejável.

1.2 OBJETIVOS

O objetivo principal deste trabalho é auxiliar o processo de tomada de decisões de uma empresa, através de um Sistema de Apoio à Decisão utilizando técnicas de Data Mining, mais especificamente para efetuar previsões genéricas.

Os objetivos específicos são:

a) estudar as tarefas e técnicas que o *Data Mining* incorpora;
b) demonstrar o potencial do *Data Mining* para previsão, analisando as técnicas mais adequadas;
c) desenvolver um SAD que seja flexível para o usuário, de modo que auxilie na construção de modelos de previsão;
d) aplicar o SAD desenvolvido no Jogo de Empresas.

1.3 ORGANIZAÇÃO DO TEXTO

O trabalho foi dividido em seis capítulos, descritos a seguir.

O primeiro capítulo define os objetivos do trabalho, apresentando a justificativa para seu desenvolvimento.

O segundo capítulo apresenta uma visão geral sobre os SI, do qual o trabalho propõe-se a utilizar, mostrando conceitos, tipos, problemas e utilidades dos mesmos. Os SI são a base para o desenvolvimento de *Data Warehouses*.

O terceiro capítulo enfatiza o *Data Warehouse*, que é uma tecnologia que oferece apoio ao Data Mining para desempenhar suas tarefas. Neste capítulo serão apresentados seus conceitos, componentes e vantagens.

O quarto capítulo enfatiza os conceitos, técnicas e aplicações do *Data Mining*.

O quinto capítulo enfatiza as Redes Neurais: conceito, rede neural biológica, modelos, limitações, vantagens, desvantagens e aplicações.

O sexto capítulo apresenta a análise, as características, o desenvolvimento e a utilização do modelo criado.

O sétimo capítulo completa o trabalho, apresentando as conclusões, limitações e sugestões para serem implementadas e aprimoradas.

2 SISTEMAS DE INFORMAÇÃO

Este capítulo apresenta os Sistemas de Informação, que são sistemas que ajudam os empreendedores a compreender e agir melhor sobre as suas empresas. Nele serão descritos seu conceito e os tipos de Sistemas de Informação. Os SI são a base para a construção de um *Data Warehouse*.

2.1 CONCEITOS

Aumentar o capital intelectual de uma empresa é uma necessidade competitiva. As organizações que usam com eficácia a tecnologia de informações adquirem conhecimento e velocidade para alcançar uma esmagadora superioridade nos mercados em que atuam. [HAR98].

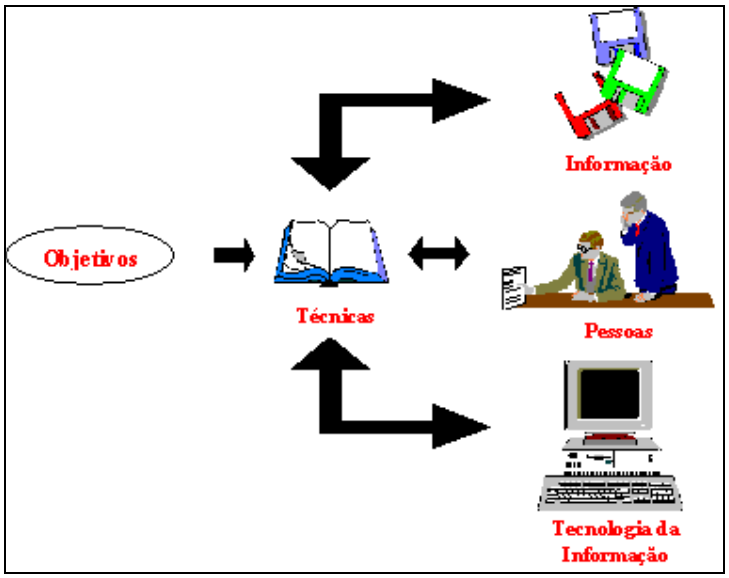
Atualmente, ainda existem empresas que possuem sistemas informatizados que servem somente para efetuar as transações operacionais e armazenar seus dados em uma base de dados. Este tipo de sistema pode ser caracterizado como um sistema de transações [OLI98].

Mas isso não é o suficiente para essas empresas, onde todos tem que ser rápidos o suficiente às oscilações e variáveis do mercado. Saem na frente as organizações cujas pessoas responsáveis pela tomada de decisões estratégicas conseguem fazer um trabalho correto e rápido. Para isso, os dados precisam ser correlacionados de tal forma que os executivos e analistas possam tomar decisões mais facilmente e trabalhar com cenários futuros. Tem-se estimado que uma pequena fração das informações está disponível na mão dos executivos e o outro tanto está nos computadores sendo cada vez mais acumulada [OLI98].

Desta forma, os executivos precisam de ferramentas que os auxiliem no gerenciamento de suas empresas, como na análise de dados e tomada de decisões. Essas ferramentas são chamadas de Sistemas de Informação, que são formas de se processar os dados de maneira ordenada para buscar informações com precisão e detectar tendências para a vitalidade da empresa. Os sistemas de informação tem um escopo diferente dos sistemas de transações; enquanto os dados operacionais estão focados em uma única área, os dados de informação precisam relacionar um grande número de áreas e um grande número de dados operacionais [OLI98].

Ao mesmo tempo, enquanto a tecnologia para a manipulação e apresentação de dados se expande, especialistas de tecnologia da informação concluíram que de todas as informações que são geradas nas empresas, apenas uma parte minúscula são dados realmente úteis [OLI98].

Segundo [ALT92], um Sistema de Informação é uma combinação das formas de trabalho, informações, pessoas, e tecnologias de informação organizadas para alcançar metas em uma Organização (figura 1).



Fonte: [ALT92]

Figura 1 - Elementos de um Sistema de Informação

Segundo [INM97], os Sistemas de Informações tem diversas possibilidades de utilização, tais como:

a) análise e investigação de tendências;
b) mensuração e rastreamento de indicadores de fatores críticos;
c) análise prospectiva;
d) monitoramento de problemas;
e) análise da concorrência.

Mas para se ter um Sistema de Informações que realmente dê as informações de forma prospectiva na hora em que se precisa, é necessário que haja um bom alicerce de dados para

os mesmos consultarem. É neste ponto da criação do alicerce de dados que fica localizada a parte mais difícil de se montar um Sistema de Informação que responda rápido aos requisitos de seu cliente [INM97].

Alguns estudos indicam que para cada U$ 9 gastos na preparação dos dados, é gasto U$ 1 para o Software e Hardware que compõem os Sistemas de Informação [INM97].

E esta montagem do alicerce fica mais difícil ainda quando tem-se a consciência de que a gerência a toda hora pode mudar de opinião sobre a informação que ela quer disponível [INM97].

2.2 TIPOS DE SISTEMAS DE INFORMAÇÃO

Os principais tipos de Sistemas de Informação, segundo [ALT92], são os seguintes:

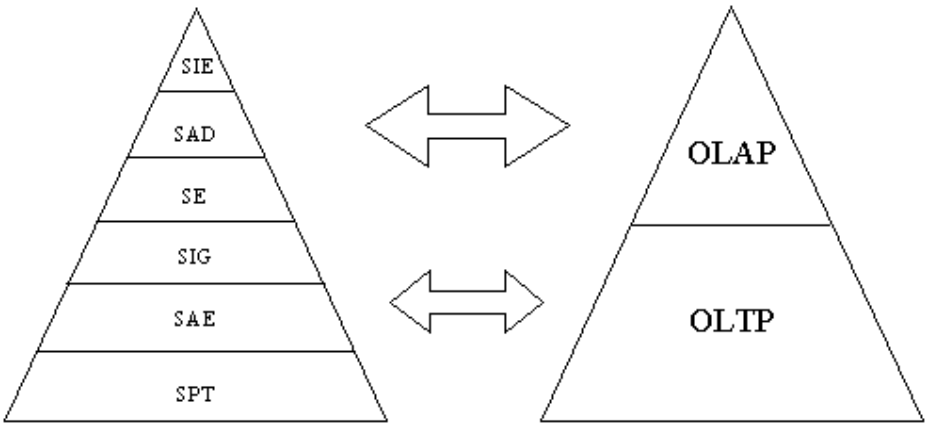
a) Sistema de Processamento de Transações (SPT): coletam e armazenam dados sobre transações e às vezes controlam decisões que são executadas como parte de uma transação. Uma transação é um evento empresarial que pode gerar ou modificar dados armazenados num Sistema de Informação. Ele foi o primeiro Sistema de Informação que surgiu e é freqüentemente encontrado. Por exemplo, quando pagamos uma conta com o Cartão de Crédito é o SPT que efetua a transação com a Central e valida o cartão. Enfim, ele grava as informações e assegura que as mesmas estão consistentes e disponíveis;
b) Sistema de Automação de Escritório (SAE): ajuda as pessoas a processar documentos e fornece ferramentas que tornam o trabalho no escritório mais eficiente e eficaz. Também pode definir a forma e o método para executar as tarefas diárias e dificilmente afeta as informações em si. Exemplos deste tipo de Sistema são editores de texto, planilhas de cálculo, softwares para correio eletrônico e outros. Todas as pessoas que tem em sua função tarefas como redigir textos, enviar mensagens, criar apresentações são usuárias de Sistemas de Automação de Escritórios.
c) Sistema de Informação Gerencial (SIG): converte os dados de uma transação do SPT em informação para gerenciar a organização e monitorar o desempenho da mesma. Ele enfatiza a monitoração do desempenho da empresa para efetuar as devidas comparações com as suas metas. As pessoas que o utilizam são os gerentes

e as que precisam monitorar seu próprio trabalho. Um exemplo disto são os relatórios que são tirados diariamente para acompanhar o Faturamento da empresa;

d) Sistemas Especialistas (SE): torna o conhecimento de especialistas disponível para outros, e ajuda a resolver problemas de áreas onde o conhecimento de especialistas é necessário. Ele pode guiar o processo de decisão e assegurar que os fatores chave serão considerados, e também pode ajudar uma empresa a tomar decisões consistentes. As pessoas que usam estes sistemas são aquelas que efetuam tarefas onde deveria existir um especialista. Um sistema especialista pode ser, por exemplo, um sistema onde médicos dizem os sintomas e é pesquisado em uma base de conhecimento os possíveis diagnósticos;
e) Sistema de Apoio à Decisão (SAD): ajuda as pessoas a tomar decisões, provendo informações, padrões, ou ferramentas para análise de informações. Ele pode prover métodos e formatos para porções de um processo de decisão. Os maiores usuários são os analistas, gerentes e outros profissionais. Os sistemas que disponibilizam gráficos 3D para comparativos são exemplos;
f) Sistema de Informações Executivas (SIE): fornece informações aos executivos de uma forma rápida e acessível, sem forçar os mesmos a pedir ajuda a especialistas em Análises de Informações. É utilizado para estruturar o planejamento da organização e o controle de processos, e pode eventualmente também ser utilizado para monitorar o desempenho da empresa. Um exemplo são os sistemas que fornecem comparativos simples e fáceis de Vendas x Estoque x Produção;

Esta forma de Sistemas de Informação que [ALT92] apresenta veio evoluindo e se transformando muito nos últimos anos, onde sua forma de apresentação mudou bastante. Antes existia uma pirâmide dividida em seis partes, na primeira camada os SPT, seguido do SAE, SIG, SE, SAD e por final o SIE.

Atualmente, segundo [MAC96] estas seis partes se transformaram em apenas duas, onde as linhas que separavam o segundo nível do sexto nível não fazem mais sentido. Estas duas camadas são a OLTP (*On Line Transaction Processing*) que fica na base da Pirâmide e a OLAP (*On Line Analytic Processing*) que fica no topo (figura 2).



Fonte: [MAC96]

Figura 2 – Evolução dos Sistemas de Informação

Conforme [MAC96], o motivo pelo qual houve a fusão entre estes grupos de sistemas reside nas mudanças por que passaram as organizações nos últimos anos. O SIE, por exemplo, voltava-se para a alta direção e tinha um aspecto mais informativo ao mesmo tempo que o SAD voltava-se para a gerência que tomava as decisões.

Da maneira como está hoje, as modificações na forma de gestão das empresas levaram as pessoas do topo a tomar mais decisões. Do mesmo modo, os gerentes que antes tomavam a maior parte das decisões tiveram seu número reduzido, consequentemente reduzindo a hierarquia e os funcionários que antes só obedeciam ordens agora podem dar sugestões para a mudança de processos.

Outro aspecto que ajudou na mudança dos Sistemas de Informação diz respeito a própria evolução tecnológica da informática. Muitas das tarefas que antes eram executadas em mainframes agora são executadas através de redes de micros, operando de forma Cliente/Servidor. Esta estrutura facilitou a montagem de Sistemas compartilhados voltados para um maior número de gerentes [MAC96].

Os sistemas baseados em OLTP são configurados e otimizados para prover respostas rápidas à transações individuais. Nestes sistemas, as transações devem ser realizadas rapidamente, e com grande confiança. Os dados são dinâmicos, mudando com grande freqüência. Já nos sistemas baseados em OLAP a velocidade das transações não é relevante,

pois os *Data Warehouses* podem armazenar os dados em forma estática, e são configurados e otimizados para suportar complexas decisões baseadas em dados históricos [OLI98].

O *Data Warehouse* é um banco de dados contendo dados extraídos do ambiente de produção da empresa, que foram selecionados e depurados, além de otimizados para processamento de consulta, e não para processamento de transações. Em geral, um *Data Warehouse* requer a consolidação de outros recursos de dados que não os armazenados em bancos de dados relacionais, incluindo informações provenientes de planilhas eletrônicas, etc. [DAL99].

A ferramenta OLAP é constituída de um conjunto de tecnologias especialmente projetadas para dar suporte ao processo decisório através de consultas, análises e cálculos mais sofisticados nos dados corporativos, estejam armazenados em um *Data Warehouse* ou não, realizados pelos seus usuários. Esta ferramenta está se tornando, cada vez mais a sucessora dos SIE e outros Sistemas do gênero com certas vantagens: ela não somente dá suporte ao processo decisório, como também às estratégias dos negócios [BIS99].

Segundo [HAR98], OLAP é essencial para a transformação do conteúdo do *Data Warehouse* em uma forma útil de informações que possam ser entregues a um grande número de usuários. Já [BIS99] afirma que o OLAP permite aos seus usuários ganharem perspicácia nas consultas e análises dos dados, através de um acesso consistente, interativo e rápido em uma grande variedade de possíveis visões dos dados.

No OLAP, as respostas não são automáticas. O processo é interativo, onde o usuário formula hipóteses, faz consultas, recebe informações, verifica um dado específico em profundidade e faz comparações. Este processo ajuda a sintetizar as informações sobre a empresa, através de comparações, visões personalizadas, análises estatísticas, previsões e simulações. Permite que os usuários se tornem exploradores de informações [BIS99].

A maioria das ferramentas OLAP são implementadas para ambientes multiusuário e arquitetura cliente/servidor, o que proporciona respostas rápidas e consistentes às consultas iterativas executadas pelos usuários, independentemente da complexidade da consulta.

Essa ferramenta pode ser utilizada em diversas situações, como no planejamento de orçamentos financeiros, análise e estimativa de vendas, pesquisa de mercado, análise de clientes, planejamento de produção, etc.

Segundo comenta [HAR98], o mercado de OLAP gira em torno de quatro habilidades diferentes, embora muitos combinem freqüentemente estas funções de análise:

a) consulta e relatórios: são aplicativos que proporcionam o tipo mais básico de análise de dados e são úteis para atender as solicitações dos usuários relativas a listas, contagens ou atualizações de status onde as exigências computacionais são relativamente simples;
b) análise multidimensional: são funções mais complexas que surgem da rápida seqüência de questionamentos do usuário. Permite uma visão conceitual multidimensional dos dados de uma empresa. A visão multidimensional dos dados é um conceito que pode parecer algo completamente abstrato e irreal; porém, é mais natural, mais fácil e intuitiva, permitindo a visão dos negócios da empresa em diferentes perspectivas. Os dados então podem ser analisados em várias dimensões, como região, produto, tempo e vendedor. Cada uma destas dimensões podem conter hierarquias, como por exemplo a dimensão tempo pode conter as hierarquias de ano, semestre, mês. [BIS99]. Os aplicativos de análise multidimensional permitem que os usuários entrem em qualquer dimensão de um *Data Warehouse* e naveguem para outras dimensões livremente;
c) análise estatística: é projetada para reduzir uma grande quantidade de dados a uma simples relação ou fórmula, como cálculos de média. Análises estatísticas mais sofisticadas incluem regressão, correlação, fatoração e agrupamentos. São geralmente utilizadas para gerar os tipos de modelos usados em aplicativos de previsões de vendas e segmentação de mercado;
d) *Data Mining*: usa muitas técnicas da análise estatística, mas ele acrescenta funções mais complexas como redes neurais para identificar modelos e relações em um conjunto de dados analisados. É particularmente útil para problemas de modelagem não-lineares com grandes números de variáveis.

3 DATA WAREHOUSE

Este capítulo apresenta o *Data Warehouse*, que é uma tecnologia que serve para o armazenamento e tratamento das informações das empresas de uma forma mais ordenada. Nele serão descritos os seus conceitos, componentes e vantagens. O *Data Warehouse* é a base para a construção de um *Data Mining*.

3.1 CONCEITOS

Em pouco tempo, novas tecnologias e conceitos para tratamento de informações estão surgindo e evoluindo para ajudar a resolver os problemas das empresas, onde através dessas tecnologias, pessoas terão um acesso mais rápido e global às informações já “lapidadas”. Uma destas ferramentas é o *Data Warehouse* [OLI98].

Segundo [OLI98], "O *Data Warehouse* é um banco de dados que armazena dados sobre as operações da empresa (vendas, compras, finanças, etc.) extraídos de uma fonte única ou múltipla, e transforma-os em informações úteis, oferecendo um enfoque histórico, para permitir um suporte efetivo à decisão".

Palma [PAL98] salienta que “Um armazém de dados se propõe a compatibilizar um número grande de sistemas desintegrados oriundos do legado a uma coleção igualmente diversa de tipos de estações de trabalho de usuário final”. Este acervo de dados, se explorado de forma inteligente, além de favorecer a tomada de decisões, propicia maior lucratividade nos negócios [INM97].

Um *Data Warehouse* é capaz de prover várias visões das informações para um grupo de usuários. Ele é capaz de derivar informações de dados que antes eram totalmente independentes um do outro [OLI98].

O *Data Warehouse* é um “depósito” de dados orientado a assunto, alimentado pelos vários sistemas transacionais da empresa, no qual se podem buscar informações para o conhecimento do negócio. Nele os dados estão agrupados e, dessa forma, é fácil a visualização dos mesmos. O *Data Warehouse* orientado a temas faz referência ao armazenamento de informações sobre temas específicos e importantes para o negócio da empresa. E também deve ser consistente, de forma a possuir uma uniformidade para viabilizar

uma melhor análise das informações. A implementação de um tema pode corresponder a um conjunto de tabelas relacionadas. Elas devem ter um elemento temporal e um identificador em comum, mas seus níveis de detalhe e sumarização são diferentes [DAL99].

Uma grande vantagem de um *Data Warehouse* é a de permitir a tomada de decisões baseadas em fatos. Na verdade, ele busca disponibilizar à organização o grande volume de dados que foram e estão sendo armazenados em bases de dados operacionais, espalhadas por toda a empresa [TAU98].

Os dados em um *Data Warehouse* não são atualizados de forma *on-line*, eles são gerados em blocos e gravados após a integração. Após este processo é que os dados ficarão disponíveis para os usuários.

O *Data Warehouse* não é um produto ou mesmo um conjunto de produtos, mas processos suportados por diversas tecnologias: ele coleta dados das várias aplicações operacionais; integra-os em um modelo lógico, por áreas de negócio; armazena as informações de tal maneira que possam ser recuperadas por usuários pouco técnicos; e entrega essas informações aos tomadores de decisão através de ferramentas de fácil uso, como geradores de relatórios e de consulta [TAU98].

O tamanho do *Data Warehouse* por si não é o fator determinante de seu sucesso. O seu uso como ferramenta de suporte a decisões é que é o principal fator. É absolutamente necessário que ele seja desenhado para acomodar as mudanças da visão de negócio, principalmente quando essas mudanças são cada vez mais rápidas [TAU98a].

Sua implementação permite análises de tendências e pode identificar relacionamentos muitas vezes desconhecidos ou simplesmente intuitivos. Nesses tempos de intensa e cruel competição, uma ferramenta que permita análises mais precisas e respostas mais rápidas da organização deve ser encarada como estratégica [TAU98].

3.2 COMPONENTES FUNCIONAIS DE UM *DATA WAREHOUSE*

Conforme Palma [PAL98], um armazém de dados é composto de três áreas funcionais distintas, cada uma das quais deve ser customizada para satisfazer as necessidades do negócio.

3.2.1 AQUISIÇÃO DE DADOS

O processo de aquisição de dados pode ser de sistemas legados ou de outras fontes quaisquer. Neste processo o dado é identificado, copiado, formatado e preparado para ser carregado no armazém.

Este processo geralmente é complexo, tedioso e caro; e gasta muito tempo efetuando basicamente três atividades [PAL98]:

a) catalogar os dados;
b) limpar e preparar os dados;
c) transportar os dados (de sua origem para o seu destino).

3.2.2 ARMAZENAMENTO DOS DADOS

Este processo pode ser administrado através de banco de dados relacionais ou orientados a objetos como o Unidata, Oracle, O2 e Jasmine. Neste contexto, deve-se utilizar um hardware ou software especializado, incluindo multiprocessador simétrico (SMP) ou processador maciçamente paralelo (MPP) [PAL98].

O SMP permite que os gerentes de armazéns de dados aumentem a capacidade dos seus sistemas sem sacrificar o ambiente existente. Máquinas com o SMP tem o seu sistema operacional UNIX ligeiramente modificado para trabalhar até com 16 processadores.

Já as máquinas com o MPP podem incorporar o uso de dúzias ou centenas de processadores.

3.2.3 ACESSO AOS DADOS

No processo de Acesso aos Dados, usuários de diferentes estações de trabalho tiram os dados do armazém com a ajuda de produtos de análise multidimensional, sistemas de redes neurais, ferramentas de mineração de dados ou outras ferramentas de análise.

Estes produtos podem ser divididos em seis categorias [PAL98]:

a) agentes inteligentes e agências: estes produtos se caracterizam por trabalhar e pensar pelo usuário. Eles permitem ao usuário pedir que o sistema inspecione

coisas, envie automaticamente relatórios e monitore o estado de várias funções do negócio empresarial;

b) facilidades de consulta e ambientes de gerenciamento de consultas: transformam um grande e complexo ambiente de armazém de dados, em uma amigável e bem-administrada estação de trabalho;

c) análise estatística: é o interesse na análise estatística tradicional com uma volta da popularidade dos pacotes estatísticos, como o SAS e o SPSS;

d) descoberta de dados: utilizando redes neurais, lógica nebulosa, árvores de decisão e outras ferramentas de matemática e estatística avançada, esses produtos permitem que os usuários peneirem quantidades volumosas de dados crus para descobrir aspectos novos, úteis sobre a companhia, suas operações e seus mercados;

e) OLAP: O processo on-line analítico ou ferramentas de planilha eletrônica multidimensionais representam uma nova geração de sistemas amigáveis de alto poder de solução. Esses sistemas permitem que as pessoas analisem a mesma informação de diversas perspectivas;

f) visualização de dados: essas ferramentas transformam simples números em excitantes apresentações visuais. Provavelmente, as ferramentas de visualização mais populares caem sob o título de sistemas de informação geográficos. Estes transformam dados sobre lojas, indivíduos ou qualquer outra coisa em mapas dinâmicos e de fácil compreensão.

3.3 DATA MARTS

Um *Data Mart* não é uma evolução de um *Data Warehouse*, mas sim parte da estratégia deste. Um *Data Mart* é um subconjunto de dados de um *Data Warehouse*, desenhado para suportar uma necessidade de negócio ou uma unidade organizacional específica [NIM98].

A idéia correta de um *Data Mart* é ele fazer parte da arquitetura *Data Warehouse*, sem perder a visão de conjunto. Essa visão de conjunto é decorrência de um bom projeto de *Data Warehouse*.

3.4 VANTAGENS

O *Data Warehouse* é feito sob medida para as necessidades do Analista de Sistemas de Informação, e por este motivo sua construção e definição são extremamente complicadas. Uma vez que ele está construído, a tarefa do Analista fica mais fácil do que quando ele não tinha este alicerce [INM97].

Conforme [INM97] relata, as vantagens para o Analista de Sistemas de Informação com o uso do *Data Warehouse* são:

a) não precisa procurar pela fonte definitiva de dados;
b) não precisa criar programas de extração especiais a partir dos sistemas existentes;
c) não precisa se preocupar com dados não integrados;
d) não precisa se preocupar com dados detalhados ou resumidos e a conexão entre os dois tipos;
e) não precisa se preocupar em encontrar um horizonte de tempo adequado;
f) não precisa se preocupar com a constante mudança de opinião por parte da gerência sobre o que precisa ser examinado a seguir;
g) dispõe de um rico suprimento de dados resumidos.

Para obter os dados necessários, o Analisa de SI pode obtê-las a partir do nível individual de processamento, do nível de processamento departamental (*data marts*), do nível resumido ou até mesmo do histórico de operações. Se o Analista partir da análise dos dados do nível individual para o nível de histórico ele terá uma visão prospectiva do processo, quer dizer, cada vez mais ele irá se aprofundando nas informações para a solução de problemas [INM97].

Os Sistemas de Informação tem as seguintes funções:

a) usar o *Data Warehouse* como o local onde os dados estão disponíveis de forma resumida;
b) usar a Construção do *Data Warehouse* para dispor de uma visão prospectiva;
c) usar os metadados do *Data Warehouse* para que o analista de SI possa planejar o modo como o Sistema de Informação será construído;
d) usar o conteúdo histórico do *Data Warehouse* para oferecer suporte a análise de tendências que a gerência deseja;

e) usar a integração dos dados que o *Data Warehouse* proporciona para se ter uma visão geral da corporação.

Resumindo, há uma afinidade muito forte entre o Analista de Sistema de Informações e o *Data Warehouse*, onde o *Data Warehouse* é o fundamento que o Analista necessita para um eficiente Sistema de Informação. Com um *Data Warehouse* bem projetado e com informações, o Analista pode tomar uma postura pró-ativa diante das necessidades da gerência fazendo análises em cima das informações, em vez de ter sempre uma postura reativa às mesmas necessidades.

O *Data Warehouse* é, sem dúvida, um conjunto de tecnologias com altíssimo potencial para as organizações. Entretanto, exige cuidados especiais para sua implementação. Além das disciplinas tradicionais de gerenciamento de projetos, o *Data Warehouse* deve ser desenhado com todos os objetivos do negócio em mente. Se os executivos não o usarem, certamente não terá tido sucesso. Por outro lado, seu uso poderá transformar radicalmente o próprio processo decisório da organização e possibilitar melhores e maiores condições de sobrevivência e crescimento nesse novo e cruel ambiente de negócios.

4 DATA MINING

A tecnologia tornou relativamente fácil o acúmulo de dados. A conseqüência é a ampliação do uso dos *Data Warehouses*, grandes repositórios de dados, agregados de forma organizada e eficiente, e em geral, de natureza histórica. Ao mesmo tempo, informação é valorizada como nunca antes na história, e os dados armazenados nos *Data Warehouses* são vasculhados por profissionais especializados, a procura de tendências e padrões.

Entretanto, a análise desses dados ainda é demorada, dispendiosa, pouco automatizada, e sujeita a erros, mal-entendidos e falta de acurácia. A automatização dos processos de análise de dados, com a utilização de softwares ligados diretamente à massa de informações, se tornou uma necessidade, já que o aproveitamento das informações já existentes, transformando-as em conhecimento, permite avanços sem paralelo na história do desenvolvimento dos bancos de dados [FIG98].

Este capítulo apresenta o *Data Mining*, que é a exploração e análise, por meios automáticos ou semi-automáticos, de uma grande quantidade de dados para descobrir padrões e regras significativos [BER97]. Nele serão descritas as etapas do Processo de KDD (*Knowledge Discovery in Databases* - KDD) e as tarefas que o Data Mining pode desempenhar.

4.1 PROSPECÇÃO DE CONHECIMENTO E *DATA MINING*

Considera-se uma hierarquia de complexidade: basicamente, se é atribuído algum significado especial a um dado, este se transforma em uma informação (ou fato). Se os especialistas elaboram uma norma (ou regra), a interpretação do confronto entre o fato e a regra constitui um conhecimento [FIG98].

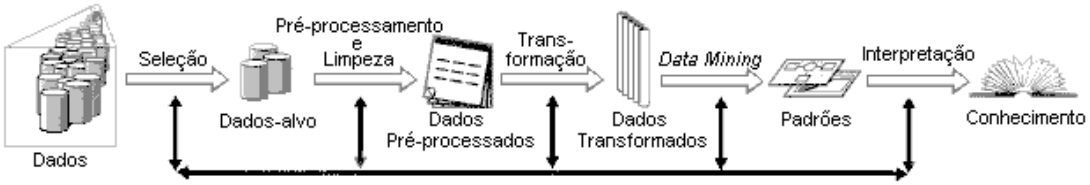
Prospecção de conhecimento em bases de dados (*Knowledge Discovery in Databases - KDD*) é um processo que envolve a automação da identificação e do reconhecimento de padrões em um banco de dados. Trata-se de uma pesquisa de fronteira, que começou a se expandir mais rapidamente nos últimos cinco anos. Sua principal característica é a extração não-trivial de informações a partir de uma base de dados de grande porte. Essas informações são necessariamente implícitas, previamente desconhecidas, e potencialmente úteis [FIG98].

Devido a essas características incomuns, todo o processo de KDD depende de uma nova geração de ferramentas e técnicas de análise de dados, e envolve diversas etapas. A principal, que forma o núcleo do processo, e que muitas vezes se confunde com ele, chama-se *Data Mining*, ou Mineração de Dados, também conhecido como processamento de padrões de dados, arqueologia de dados, ou colheita de informação (*information harvesting*).

O KDD compreende todo o processo de descoberta de dados, enquanto o Data Mining refere-se a aplicação de algoritmos para extração de padrões de dados, sem os passos adicionais do KDD e da análise dos resultados [AVI98].

4.2 AS ETAPAS DO PROCESSO DE KDD

O processo de KDD (figura 3) começa com o entendimento do domínio da aplicação e dos objetivos finais a serem atingidos. Em seguida, é feito um agrupamento organizado de uma massa de dados, alvo da prospecção. A etapa da limpeza dos dados (*data cleaning*) vem a seguir, através de um pré-processamento dos dados, visando adequá-los aos algoritmos. Isso se faz através da integração de dados heterogêneos, eliminação de incompletude dos dados, repetição de tuplas, problemas de tipagem, etc. Essa etapa pode tomar até 80% do tempo necessário para todo o processo, devido às bem conhecidas dificuldades de integração de bases de dados heterogêneas [FAY96].



Fonte: [FIG98]

Figura 3 - Os passos do processo de KDD

Os dados pré-processados devem ainda passar por uma transformação que os armazena adequadamente, visando facilitar o uso das técnicas de *Data Mining*. Nessa fase, o uso de *Data Warehouses* se expande consideravelmente, já que nessas estruturas as informações estão alocadas da maneira mais eficiente. Em *Data Warehouses*, os dados são não-voláteis, classificados por assunto, e de natureza histórica, tendendo portanto a se tornarem grandes

repositórios de dados extremamente organizados. Entretanto, em algumas aplicações de *Data Mining* mais específicas, ferramentas avançadas de representação de conhecimento podem descrever o conteúdo de um banco de dados por si só, usando esse mapeamento como uma meta-camada para os dados.

Prosseguindo no processo, chega-se à fase de *Data Mining* especificamente, que começa com a escolha dos algoritmos a serem aplicados. Essa escolha depende fundamentalmente do objetivo do processo de KDD: classificação, segmentação, agrupamento por afinidades, estimativas, etc. De modo geral, na fase de *Data Mining*, ferramentas especializadas procuram padrões nos dados. Essa busca pode ser efetuada automaticamente pelo sistema ou interativamente com um analista, responsável pela geração de hipóteses. Diversas ferramentas distintas, como redes neurais, indução de árvores de decisão, sistemas baseados em regras e programas estatísticos, tanto isoladamente quanto em combinação, podem ser então aplicadas ao problema. Em geral, o processo de busca é iterativo, de forma que os analistas revêem o resultado, formam um novo conjunto de questões para refinar a busca em um dado aspecto das descobertas, e realimentam o sistema com novos parâmetros. Ao final do processo, o sistema de *Data Mining* gera um relatório das descobertas, que passa então a ser interpretado pelos analistas de mineração. Somente após a interpretação das informações obtidas encontra-se o conhecimento.

Uma diferença significante entre *Data Mining* e outras ferramentas de análise está na maneira como exploram as interrelações entre os dados. As diversas ferramentas de análise disponíveis dispõem de um método baseado na verificação, isto é, o usuário constrói hipóteses sobre interrelações específicas e então verifica ou refuta, através do sistema. Esse modelo torna-se dependente da intuição e habilidade do analista em propor hipóteses interessantes, em manipular a complexidade do espaço de atributos, e em refinar a análise baseado nos resultados de consultas ao banco de dados potencialmente complexas. Já o processo de *Data Mining* fica responsável pela geração de hipóteses, garantindo mais rapidez, acurácia e completude aos resultados.

Estas etapas são interdependentes, pois os resultados de cada uma são a entrada da próxima etapa. Toda a abordagem é dirigida por resultados e cada estágio depende dos resultados do estágio anterior [HAR98]. Mas não existe uma ordem ou seqüência totalmente única para o andamento deste processo, porque isso depende das técnicas empregadas e dos

dados sobre os quais o KDD está sendo aplicado [AVI98]. A qualquer momento, por exemplo, pode-se voltar o processo de KDD para uma etapa anterior, desde que a técnica e os dados empregados permitam.

4.3 UTILIDADES DO DATA MINING.

O *Data Mining* pode desempenhar uma série limitada de tarefas dependendo das circunstâncias. Cada classe de aplicação em *Data Mining* tem como base um conjunto de algoritmos que serão usados na extração de relações relevantes dentro de uma massa de dados [HAR98]:

a) classificação;
b) estimativa;
c) agrupamento por afinidade;
d) previsão;
e) segmentação.

Cada uma destas propostas difere quanto à classe de problemas que o algoritmo será capaz de resolver.

4.3.1 CLASSIFICAÇÃO

Classificação é uma técnica que consiste na aplicação de um conjunto de exemplos pré-classificados para desenvolver um modelo capaz de classificar uma população maior de registros. Detecção de fraudes e aplicações de risco são exemplos de casos em que este tipo de análise é bastante apropriada. Em geral, algoritmos de classificação incluem árvores de decisão ou redes neurais, e começam com um treinamento a partir de transações-exemplo. O algoritmo classificador usa estes exemplos para determinar um conjunto de parâmetros, codificados em um modelo, que será mais tarde utilizado para a discriminação do restante dos dados.

Uma vez que o algoritmo classificador foi desenvolvido de forma eficiente, ele será usado de forma preditiva para classificar novos registros naquelas mesmas classes pré-definidas.

Alguns exemplos de Classificação são:

a) classificar pedidos de créditos como de baixo, médio e alto risco;
b) esclarecer pedidos de seguro fraudulentos;
c) atribuir palavras-chave a artigos jornalísticos.

4.3.2 ESTIMATIVA

Uma variação do problema de classificação envolve a geração de valores ao longo das dimensões dos dados: são os chamados algoritmos de estimativa. A estimativa lida com resultados contínuos, ao contrário da classificação que lida com resultados discretos. Fornecidos alguns dados, usa-se a estimativa para estipular um valor para alguma variável contínua desconhecida como receita, altura ou saldo de cartão de crédito.

Ao invés de um classificador binário determinar um risco "positivo" ou "negativo", a técnica gera valores de "escore", dentro de uma determinada margem. A abordagem de estimativa tem a grande vantagem de que os registros individuais podem ser agora ordenados por classificação, e as redes neurais são adequadas a esta tarefa.

Exemplos de Estimativa incluem:
a) estimar o número de filhos numa família;
b) estimar a renda total de uma família;
c) estimar o valor em tempo de vida de um cliente.

4.3.3 AGRUPAMENTO POR AFINIDADE

Este algoritmo identifica afinidades entre itens de um subconjunto de dados. Essas afinidades são expressas na forma de regras: "72% de todos os registros que contém os itens A, B, e C também contém D e E". A porcentagem de ocorrência (72 no caso) representa o fator de confiança da regra, e costuma ser usado para eliminar tendências fracas, mantendo apenas as regras mais fortes. Dependências funcionais podem ser vistas como regras de associação com fator de confiança igual a 100%.

Trata-se de um algoritmo tipicamente endereçado à análise de mercado, onde o objetivo é encontrar tendências dentro de um grande número de registros de compras, por exemplo, expressas como transações. Essas tendências podem ajudar a entender e explorar padrões de compra naturais, e pode ser usada para ajustar mostruários, modificar prateleiras

ou propagandas, e introduzir atividades promocionais específicas. Um exemplo mais distinto, onde essa mesma técnica pode ser utilizada, é o caso de um banco de dados escolar, relacionando alunos e disciplinas. Uma regra do tipo "84% dos alunos inscritos em 'Introdução ao Unix' também estão inscritos em 'Programação em C'" pode ser usada pela direção ou secretaria para planejar o currículo anual, ou alocar recursos como salas de aula e professores [FIG98].

4.3.4 PREVISÃO

A previsão é o mesmo que classificação ou estimativa, exceto pelo fato de que os registros são classificados de acordo com alguma atitude futura prevista. Em um trabalho de previsão, o único modo de confirmar a precisão da classificação é esperar para ver.

Essa tarefa é uma variante do problema de agrupamento por afinidades, onde as regras encontradas entre as relações podem ser usadas para identificar seqüências interessantes, que serão utilizadas para predizer acontecimentos subsequentes. Nesse caso, não apenas a coexistência de itens dentro de cada transação é importante, mas também a ordem em que aparecem, e o intervalo entre elas. Seqüências podem ser úteis para identificar padrões temporais, por exemplo entre compras em uma loja, ou utilização de cartões de crédito, ou ainda tratamentos médicos.

Exemplos de tarefas de previsão:

a) previsão de quais clientes sairão nos próximos seis meses;
b) previsão da quantia de dinheiro que um cliente utilizará caso seja oferecido a ele um certo limite de cartão de crédito.

4.3.5 SEGMENTAÇÃO

A segmentação é um processo de agrupamento de uma população heterogênea em vários subgrupos ou *clusters* mais homogêneos. O que a distingue da classificação é que segmentação não depende de classes pré-determinadas.

Essa segmentação é realizada automaticamente por algoritmos que identificam características em comum e particionam o espaço n-dimensional definido pelos atributos.

Os registros são agrupados de acordo com a semelhança e depende do usuário determinar qual o significado de cada segmento, caso exista algum. Muitas vezes a segmentação é uma das primeiras etapas dentro de um processo de *Data Mining*, já que identifica grupos de registros correlatos, que serão usados como ponto de partida para futuras explorações. O exemplo clássico é o de segmentação demográfica, que serve de início para uma determinação das características de um grupo social, visando desde hábitos de compras até utilização de meios de transporte.

4.4 TÉCNICAS DE DATA MINING

Muitas das técnicas usadas em ferramentas atuais de *Data Mining* se originaram na pesquisa em inteligência artificial da década de 80 e princípio da década de 90. Entretanto, somente agora essas técnicas passaram a ser utilizadas em sistemas de banco de dados de grande escala, devido a confluência de diversos fatores que aumentaram o valor líquido da informação, dentre os quais se destacam [FIG98]:

a) a expansão e difusão de sistemas transacionais volumosos: nos últimos 15 ou 20 anos, computadores estão sendo usados para capturar e armazenar informações detalhadas de processos transacionais intensivos, como vendas, telecomunicações, bancos e operações com cartões de crédito. Os SGBDs saltaram de algumas centenas de transações por minuto para mais de 10.000/min, com exceções que chegam a 30.000. Esse crescimento da capacidade de processamento é acompanhado de uma redução equivalente do custo por processamento, que ajuda a disseminar a tecnologia e integrá-la ao mercado, gerando uma proliferação ainda maior de sistemas de transações geradores de informação.
b) informação como vantagem competitiva: a necessidade da informação resulta na proliferação de *Data Warehouses* que integram múltiplos sistemas operacionais para suporte a decisão, muitas vezes incluindo dados de fontes externas, como registros demográficos.
c) a difusão de tecnologia de informação escalável: a busca da interoperabilidade levou à recente adoção da sistemas de informação escaláveis, incluindo SGBDs, ferramentas analíticas e troca de informações via serviços de Internet/Intranet.

Por outro lado, a quantidade de dados brutos armazenados em *Data Warehouses* corporativos está crescendo rapidamente, tornando o "espaço de decisão" muito extenso e complexo para os atuais sistemas de suporte a decisão.

[FIG98] explica que por causa desta grande quantidade de dados brutos, todo o processo de KDD atual ainda requer pré/pós-processamentos dos dados, necessários para assegurar o melhor aproveitamento da aplicação e a consistência dos resultados. Atividades de pré-processamento incluem a seleção apropriada de subconjuntos de dados, por razões de desempenho, assim como complexas transformações de dados que servem de ponte para o chamado "gap representacional", separação entre os dados e seu significado real. Pós-processamento envolve a subseleção de resultados volumosos e a aplicação de técnicas de visualização para auxiliar o entendimento. Essas atividades são críticas para contornar alguns problemas de implementação, tais como:

a) alta suscetibilidade a dados "sujos": as ferramentas de *Data Mining* via de regra não possuem uma estrutura dotada de semântica, orientada a aplicação, e como tal, tomam todos os dados factualmente. Torna-se necessário tomar precauções para assegurar que os dados analisados são "limpos", o que pode significar uma exaustiva análise dos atributos que alimentam os algoritmos. Entretanto, um bom processo de "limpeza de dados" (data cleaning), utilizado na passagem dos dados para um *Data Warehouse* certamente beneficia o processo de *Data Mining*.
b) inabilidade para "explicar" resultados em termos humanos: mesmo em aplicações utilizando árvores de decisão e regras de indução, que são capazes de gerar informação sobre os atributos utilizados, o volume e formato da informação encontrada pode ser inútil sem um processamento adicional.
c) "gap" representacional: a maior parte das fontes de dados das aplicações de *Data Mining* atuais está armazenada em grandes sistemas relacionais, e seus dados estão em geral normalizados, com os atributos espalhados em múltiplas tabelas. Além disso, a maioria das ferramentas é restrita em termos dos tipos de dados com as quais podem operar, tornando-se necessário categorizar variáveis ou remapeá-las.

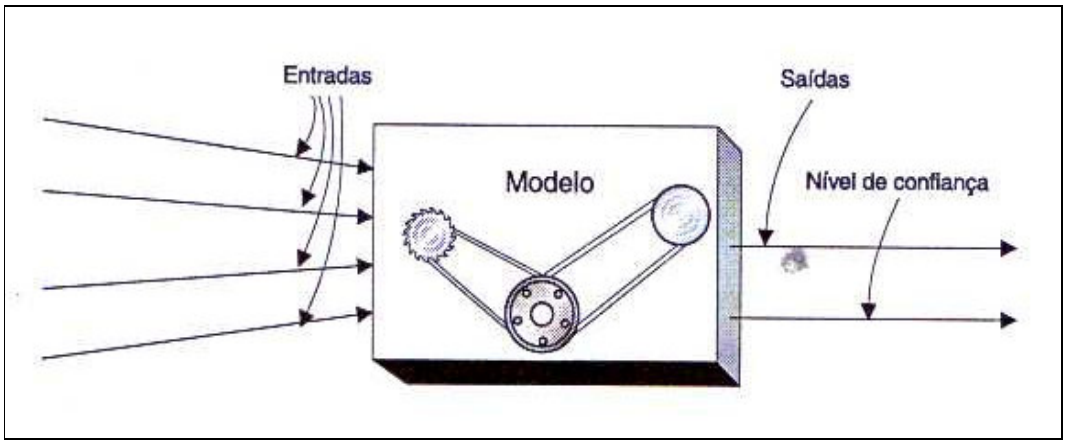
4.4.1 MODELOS

Conforme salienta [HAR98], um modelo produz um ou mais valores a partir de um dado conjunto de entradas. A análise dos dados é, com freqüência, o processo de construção de um modelo apropriado para os dados (figura 4). Um exemplo disso é uma regressão linear, onde é construída sobre um modelo em linha com a seguinte forma:

$$aX + bY + c = 0$$

Onde a, b, c são os parâmetros e X e Y são as variáveis. Para um dado valor de X, estima-se o valor de Y. Este tipo de modelo é um dos mais simples existentes.

O fato de um modelo existir não significa que proporcionará resultados precisos. Existem bons e maus modelos e, medir seus resultados é um passo crítico em seu uso e desenvolvimento [HAR98].



Fonte: [BER97]

Figura 4 - Modelo recebe entradas e produz informações.

Um modelo de classificação apanha um novo registro e atribui ao mesmo uma classificação existente. Um modelo de previsão é semelhante a um modelo de classificação, exceto por não ser limitado a um conjunto de número de classes. Um modelo de agrupamento toma vários registros e retorna um número menor de grupos. Esses grupos podem então ser aplicados a novos registros, criando um modelo de classificação. Um modelo de séries temporais é como um modelo de classificação ou de previsão, exceto por incluir dados tomados com o decorrer do tempo [BER97].

Na criação dos modelos, a entrada é geralmente especificada claramente. Geralmente, preparar os dados de sistemas para preencher o domínio de um modelo – chamado de depuração de dados ou *data scrubbing* – é mais desafiador do que a própria criação do modelo. Os dados que alimentarão o modelo podem afetar a escolha da técnica. Para problemas físicos, com muitas variáveis contínuas de entrada, as técnicas de regressão estatísticas normalmente funcionam muito bem. Quando as entradas tem muitas variáveis de categorias, as árvores de decisão funcionam melhor. Quando a relação entre as entradas e a saída de dados é difícil de ser estabelecida, as redes neurais são as melhores opções.

Freqüentemente a saída de dados de um modelo é especificada em primeiro lugar e geralmente é uma categoria ou uma variável contínua.

Segundo [BER97], para criar um modelo para *Data Mining*, deve-se ter em mente o seguinte:

a) um dos perigos no uso de modelos é o excesso ou a carência de dados;
b) tanto o *Data Mining* direto quanto o indireto usam modelos, mas de maneira diversa;
c) alguns modelos expõem sua finalidade melhor que outros;
d) alguns modelos são mais fáceis de aplicar que outros.

4.4.2 TÉCNICAS E TAREFAS

Cada técnica de *Data Mining* possui tarefas onde elas são melhores aplicáveis.

4.4.2.1 ANÁLISE DE SELEÇÃO ESTATÍSTICA

A análise de seleção estatística é uma forma de agrupamento usada para encontrar grupos de itens que tendem a ocorrer em conjunto em uma seleção estatística. Como técnica de agrupamento, ela é útil quando se deseja saber quais itens ocorrem ao mesmo tempo ou em uma seqüência particular [HAR98].

4.4.2.2 MBR

O MBR (*Memory-Based Reasoning* – raciocínio baseado em memória) é uma técnica de *Data Mining* dirigida que usa exemplos conhecidos como modelo para fazer previsões

sobre exemplos desconhecidos. O MBR procura os vizinhos mais próximos nos exemplos conhecidos e combina seus valores para atribuir valores de classificação ou de previsão [BER97].

Os elementos-chave no MBR são a função de distância usada para encontrar os vizinhos mais próximos e a função de combinação, que combina valores dos vizinhos mais próximos para fazer uma previsão. Uma vantagem do MBR é sua habilidade de aprender sobre novas classificações simplesmente introduzindo novos exemplos no banco de dados. Uma vez encontrada a função de distância e a função de combinação corretas tendem a permanecer muito estáveis, mesmo com a incorporação de novos exemplos para novas categorias nos dados conhecidos. Aliás, esta é uma característica que diferencia o MBR da maior parte das outras técnicas de *Data Mining*.

4.4.2.3 ALGORITMOS GENÉTICOS

Os algoritmos genéticos aplicam a mecânica da genética e seleção natural à pesquisa usada para encontrar os melhores conjuntos de parâmetros que descrevem uma função de previsão. Eles são utilizados no *Data Mining* dirigido e são semelhantes à estatística, em que a forma do modelo precisa ser conhecida em profundidade. Os algoritmos genéticos usam os operadores seleção, cruzamento e mutação para desenvolver sucessivas gerações de soluções. Com a evolução do algoritmo, somente os mais previsíveis sobrevivem, até as funções convergirem em uma solução ideal [BER97].

Esta técnica é apropriada para resolver os mesmos tipos de problemas que as outras técnicas de *Data Mining*, mas ela também pode ser usada para aprimorar MBRs e redes neurais.

4.4.2.4 DETECÇÃO DE AGRUPAMENTOS

Esta técnica constitui-se na construção de modelos para encontrar dados semelhantes, e estas reuniões por semelhança são chamadas de grupos (*clusters*). É uma forma de *Data Mining* não-direcionado, onde a meta é encontrar similaridades não conhecidas anteriormente. Existem muitas técnicas para encontrar grupos, incluindo métodos geométricos, estatísticos e redes neurais [HAR98].

4.4.2.5 ANÁLISE DE VÍNCULOS

A análise de vínculos segue as relações entre registros para desenvolver modelos baseados em padrões nas relações. Esse é um aplicativo de construção de teoria gráfica de *Data Mining*. Esta técnica não é muito compatível com a tecnologia de banco de dados relacionais e sua maior área de aplicação é a área policial, onde pistas são ligadas entre si para solucionar os crimes. As poucas ferramentas que existem, enfocam mais a visualização de vínculos que a análise de padrões [HAR98].

4.4.2.6 ÁRVORES DE DECISÃO E INDUÇÃO DE REGRAS

As árvores de decisão são usadas para o *Data Mining* dirigido, mais especificamente a classificação. Esta técnica divide os registros do conjunto de dados de treinamento em subconjuntos separados, cada um descrito por uma regra simples em um ou mais campos [HAR98].

Uma grande vantagem nesta técnica é que o modelo é bem explicável, já que tem a forma de regras explícitas. Isto permite às pessoas avaliarem os resultados, identificando os atributos-chave do processo.

4.4.2.7 REDES NEURAIS ARTIFICIAIS

As redes neurais são modelos simples de interconexões neurais no cérebro, adaptados para o uso em computadores e são, provavelmente, a técnica de *Data Mining* mais utilizada. Elas aprendem com um conjunto de dados de treinamento, generalizando modelos para classificação e previsão. Esta técnica pode também ser aplicada ao *Data Mining* não-dirigido (na forma de redes Kohonen e estruturas relacionadas) e às previsões em séries temporais [HAR98].

Uma das principais vantagens na utilização desta técnica é a sua variedade de aplicação. Elas são interessantes porque detectam padrões nos dados de forma análoga ao pensamento humano. Mas existem duas desvantagens em seu uso:

a) a dificuldade de interpretar os modelos produzidos por elas;
b) a sensibilidade ao formato dos dados que a alimentam, pois representações de dados diferentes podem produzir resultados diversos.

4.5 ESCOLHA DA TÉCNICA

No trabalho de [HAR98] está descrito que a escolha da técnicas de *Data Mining* dependerá da tarefa específica a ser executada e dos dados disponíveis para análise (tabela 1).

	Classificação	Estimativa	Previsão	Agrupamento por afinidade	Segmentação
Estatística padrão	4	4	4	4	4
Análise de seleção estatística			4	4	4
MBR	4		4	4	4
Algoritmos genéticos	4		4		
Detecção de grupos					4
Análise de vínculos	4		4	4	
Árvores de decisão	4		4		4
Redes neurais	4	4	4		4

Fonte: [HAR98]

Tabela 1 – Quais técnicas para cada tarefa

5 REDES NEURAIS

Este capítulo apresenta as redes neurais artificiais, que são umas das técnicas utilizadas para implementar algumas tarefas de Data Mining. Nele será descrita a definição das redes neurais artificiais e apresentada uma semelhança com a rede neural biológica humana. Também será demonstrada sua estrutura e componentes, sua aprendizagem, revocação, vantagens e desvantagens.

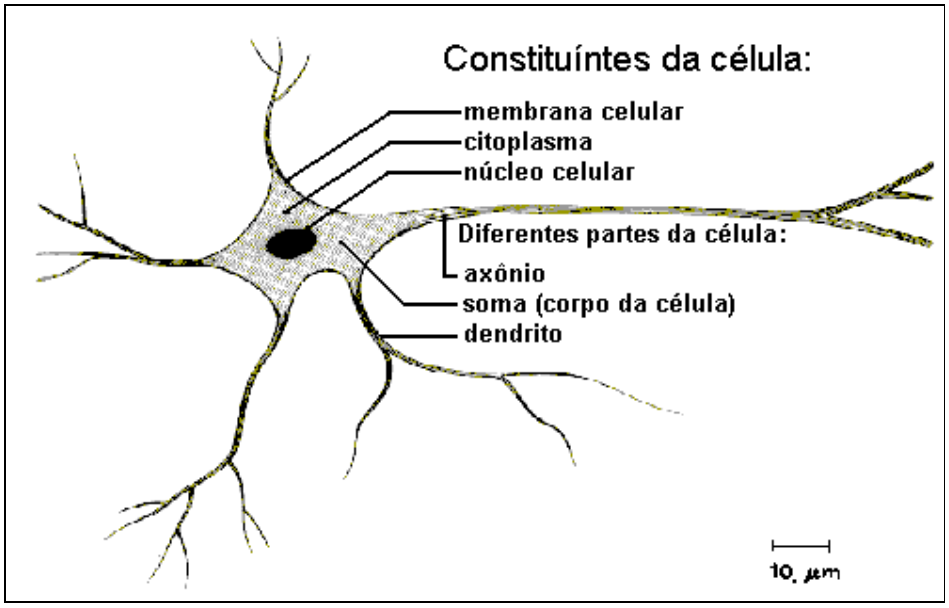
As redes neurais artificiais são muito utilizadas para efetuar a tarefa de previsão em Data Mining, principalmente com o modelo de rede *Feedforward/Backpropagation* que é apresentado neste capítulo. E a compreensão do funcionamento básico de uma rede neural artificial é de extrema importância para a aplicação de um Data Mining com o uso da mesma.

5.1 REDE NEURAL BIOLÓGICA

O sistema nervoso é formado por um conjunto extremamente complexo de células, os neurônios. (figura 5) Eles têm um papel essencial na determinação do funcionamento e comportamento do corpo humano e do raciocínio. Os neurônios são formados pelos dendritos, que são um conjunto de terminais de entrada, pelo corpo central, e pelos axônios que são longos terminais de saída [LOE96].

Os neurônios se comunicam através de sinapses. Sinapse é a região onde dois neurônios entram em contato e através da qual os impulsos nervosos são transmitidos entre eles. Os impulsos recebidos por um neurônio A, em um determinado momento, são processados, e atingindo um dado limiar de ação, o neurônio A dispara, produzindo uma substância neurotransmissora que flui do corpo celular para o axônio, que pode estar conectado a um dendrito de um outro neurônio B. O neurotransmissor pode diminuir ou aumentar a polaridade da membrana pós-sináptica, inibindo ou excitando a geração dos pulsos no neurônio B. Este processo depende de vários fatores, como a geometria da sinapse e o tipo de neurotransmissor.

Em média, cada neurônio forma entre mil e dez mil sinapses. O cérebro humano possui cerca de 10 E11 neurônios, e o número de sinapses é de mais de 10 E14, possibilitando a formação de redes muito complexas.



Fonte: [VAL97]

Figura 5 - Constituintes da célula neuronal - esquema.

5.2 REDES NEURAIS ARTIFICIAIS

Segundo Loesch [LOE96], redes neurais artificiais são sistemas computacionais de implementação em hardware ou software, que imitam as habilidades computacionais do sistema nervoso biológico, usando um grande número de simples neurônios artificiais interconectados.

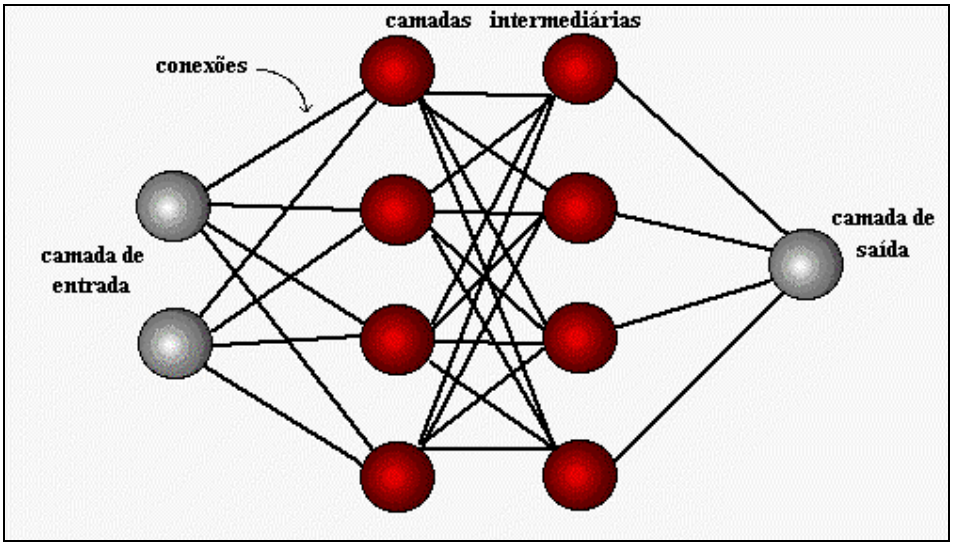
Redes neurais artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência [WIL95]. Uma grande rede neural artificial pode ter centenas ou milhares de unidades de processamento; já o cérebro de um mamífero pode ter muitos bilhões de neurônios.

Uma rede neural artificial é composta por várias unidades de processamento, cujo funcionamento é bastante simples. Essas unidades, geralmente são conectadas por canais de comunicação que estão associados a determinado peso. As unidades fazem operações apenas sobre seus dados locais, que são entradas recebidas pelas suas conexões. O comportamento

inteligente de uma rede neural artificial vem das interações entre as unidades de processamento da rede [LOE96].

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 [LOE96].

Arquiteturas neurais são tipicamente organizadas em camadas, com unidades que podem estar conectadas às unidades da camada posterior (figura 6).



Fonte: [VAL97]

Figura 6 - Organização das camadas.

Usualmente as camadas são classificadas em três grupos:

a) Camada de Entrada: onde os padrões são apresentados à rede;
b) Camadas Intermediárias ou Escondidas: onde é feita a maior parte do processamento, através das conexões ponderadas; podem ser consideradas como extratoras de características;
c) Camada de Saída: onde o resultado final é concluído e apresentado.

Uma rede neural é especificada, principalmente pela sua topologia, pelas características dos nós e pelas regras de treinamento. A seguir, serão analisados os processos de aprendizado.

5.2.1 PROCESSOS DE APRENDIZADO

A propriedade mais importante das redes neurais é a habilidade de aprender de seu ambiente e com isso melhorar seu desempenho. Isso é feito através de um processo iterativo de ajustes aplicado a seus pesos, o treinamento (figura 7). O aprendizado ocorre quando a rede neural atinge uma solução generalizada para uma classe de problemas [LOE96].

Denomina-se algoritmo de aprendizado a um conjunto de regras bem definidas para a solução de um problema de aprendizado. Existem muitos tipos de algoritmos de aprendizado específicos para determinados modelos de redes neurais, estes algoritmos diferem entre si principalmente pelo modo como os pesos são modificados.

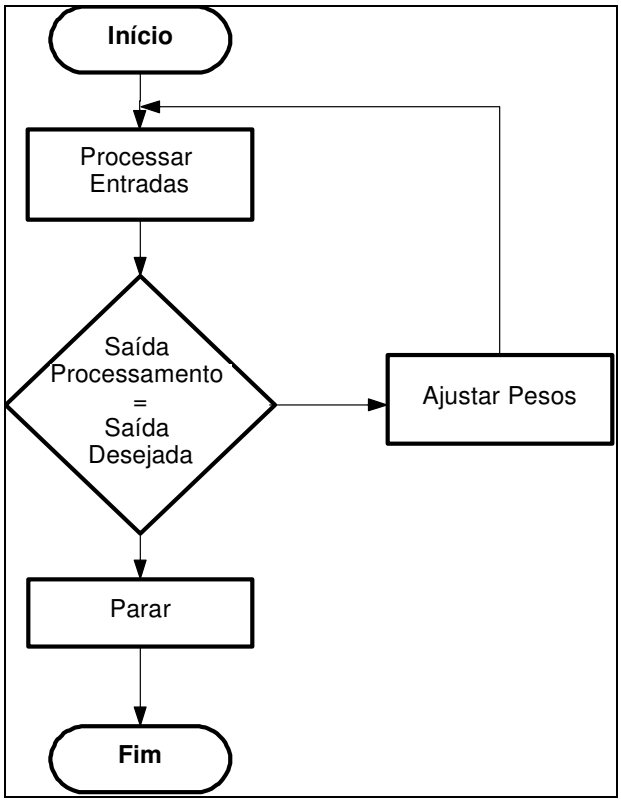


Figura 7 – Fluxograma do algoritmo de aprendizagem de uma Rede Neural.

Outro fator importante é a maneira pela qual uma rede neural se relaciona com o ambiente. Nesse contexto existem os seguintes paradigmas de aprendizado:

a) Aprendizado Supervisionado: A maioria absoluta das aplicações existentes compõem-se de redes neurais com aprendizado supervisionado, que pode ser

considerado como a capacidade que a rede possui de modificar o seu desempenho a partir da comparação entre a resposta obtida e a resposta desejada. A partir das entradas fornecidas, os pesos das conexões são ajustados por cálculo até obter-se as saídas desejadas [VAL97]. Como um exemplo de arquitetura de rede com aprendizado supervisionado pode ser citada a *BackPropagation*;

b) Aprendizado Não Supervisionado (auto-organização): estas redes têm a capacidade de determinar uma correlação entre os possíveis padrões de entrada que são particularmente úteis nos problemas em que as entradas variam com o tempo de forma conhecida. Pode-se considerar este mecanismo de aprendizado como sendo a capacidade que a rede possui de abstrair correlações entre os estímulos de modo a obter as respostas desejadas. Somente as entradas são apresentadas à rede. A mesma é auto-organizada e ajusta-se às entradas fornecidas. Como um exemplo de arquitetura de rede com aprendizado não-supervisionado pode ser citada a *Kohonen* [VAL97].

Denomina-se ciclo uma apresentação de todos os N pares (entrada e saída) do conjunto de treinamento no processo de aprendizado. A correção dos pesos num ciclo pode ser executado de dois modos:

a) Modo Padrão: A correção dos pesos acontece a cada apresentação à rede de um exemplo do conjunto de treinamento. Cada correção de pesos baseia-se somente no erro do exemplo apresentado naquela iteração. Assim, em cada ciclo ocorrem N correções.

b) Modo Batch: Apenas uma correção é feita por ciclo. Todos os exemplos do conjunto de treinamento são apresentados à rede, seu erro médio é calculado e a partir deste erro fazem-se as correções dos pesos.

5.2.2 REVOCAÇÃO

Após o modelo de rede ter sido submetido a um período de aprendizagem ou treinamento, aplica-se um processo denominado de revocação. A revocação consiste em aplicar os resultados obtidos com o treinamento da rede (valor estabelecidos para as conexões) em aplicações cujas variáveis de entrada, podem ou não ser iguais (ou bastante próximas) às oferecidas quando do aprendizado [VAL97].

Revocação é o processo onde se verifica a efetividade da aprendizagem. Através deste processo a rede deverá reconhecer ou não novos padrões de entrada que lhe forem apresentados.

5.2.3 MODELOS DE REDES NEURAIS ARTIFICIAIS

Muitos são os modelos de redes neurais existentes. Para cada aplicação pode-se definir qual o melhor modelo e estrutura. Loesch [LOE96] apresenta alguns modelos de redes neurais bem como sua aplicação básica e ano de publicação (tabela 2):

Modelo	Aplicações Básicas	Ano
Adaline/Madaline	Filtragem de sinal adaptativo, equalização adaptativa	1960
Adaptative Resonance Theory (ART)	Reconhecimento de Padrões	1983
Backpropagation Perceptron	Reconhecimento de padrões, filtragem de sinal, controle robótico, compressão de dados, segmentação de sinal, etc.	1974-1986
BAM – Memória Associativa Bidirecional	Heteroassociativa (memória endereçada por conteúdo)	1987
Boltzmann Machine, Cauchy Machine	Reconhecimento de padrões (imagens, sons, radar), otimização	1984
Brain-State-in-a-Box (BSB)	Revocação autoassociativa	1977
Hopfield	Evocação autoassociativa, otimização	1982
Neocognitron	Reconhecimento de caracteres manuscritos / imagens	1975
Quantização de Vetor de aprendizagem	Revocação autoassociativa (complementação de uma padrão a outro parcial apresentado), compressão de dados	1981
Recurrent	Controle robótico, reconhecimento de fala, previsão do elemento seqüencial	1987
Redes de funções de base radial	Classificação, mapeamento	1987
Redes de ligações funcionais	Classificação, mapeamento	1988
Time-Delay	Reconhecimento de fala	1987

Tabela 2 - Modelos de redes neurais

5.2.3.1 MODELO FEEDFORWARD / BACKPROPAGATION

O modelo *Feedforward* com aprendizado *Backpropagation*, surgiu por meados da década de 80 e constitui, segundo pesquisadores, a mais difundida e largamente usada entre

todas arquiteturas e modelos de redes neurais conhecidas [WIL95]. O que ele trouxe de diferente em relação aos modelos existentes até então foram as múltiplas camadas, com a possibilidade de valores de entradas e saídas contínuos (ex. 0.0001 à 0.9999).

Este modelo utiliza-se de valores contínuos (ex. 0.0001 à 0.9999), o que difere do modelo *Perceptron* que utiliza valores discretos (ex. 0 ou 1) [LOE96].

Uma aplicação modelada nesta topologia, necessita de padrões de entrada e saída, para a qual a rede converge e se estabiliza, constituindo o que denomina-se de treinamento ou aprendizado da rede. Este modelo utiliza-se de aprendizado supervisionado, ou seja, a cada padrão de entrada está associado a uma saída desejada [VAL97].

Os elementos de processamento das camadas ocultas dão ao modelo a capacidade de abstração e generalização, ou seja, é capaz de classificar um padrão complexo mesmo quando este não pertenceu ao conjunto de treinamento. A rede é portanto imune a pequenas falhas.

O treinamento deste modelo de rede consiste em ajustar os pesos de conexões das camadas para que o conjunto de entradas atinja o conjunto de saídas desejadas.

A revocação é feita diante dos pesos das conexões armazenadas em arquivos. Nesta etapa, um novo padrão de entrada é oferecido (como par de entrada) à rede, ela computa e gera uma saída [VAL97].

Sua rápida operacionalização, apresentando capacidade de generalização, robustez e abstração são suas principais vantagens. O fato de requerer um longo tempo de treinamento, em função da necessidade de estabilização e convergência da rede é sua desvantagem, segundo [LOE96].

6 DESENVOLVIMENTO DO SAD

Para o desenvolvimento do SAD, adotou-se algumas etapas da metodologia de prototipação fundamental. Esta, segundo [MEL90], é uma metodologia no qual tanto os analistas quanto os usuários sabem que o produto final da prototipação será o próprio sistema, já na sua forma aperfeiçoada. A metodologia de prototipação fundamental é dividida em oito etapas, mais foi feita uma adaptação nas mesmas e foram utilizadas apenas quatro (figura 8).

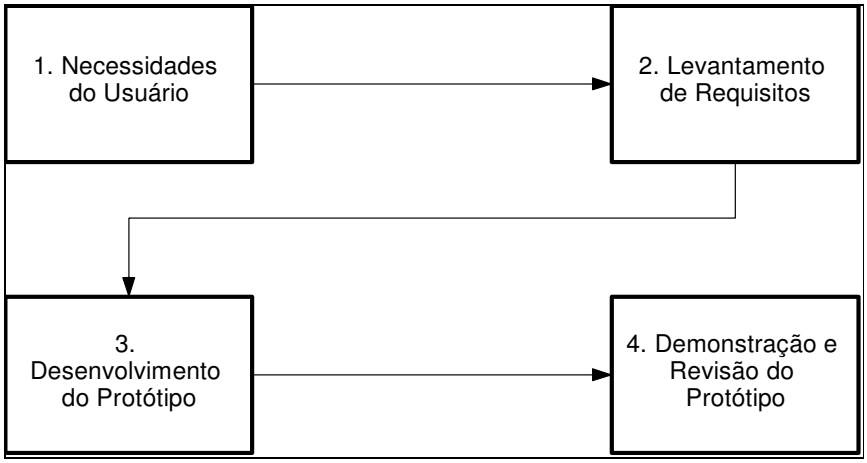


Figura 8 – Adaptação das etapas da metodologia de prototipação fundamental.

6.1 NECESSIDADES DO USUÁRIO

Cada vez mais executivos tem a necessidade de analisar o desempenho de suas empresas, bem como prever e agir sobre os resultados que estão por vir. E uma das maiores necessidades destes executivos é saber prever quais serão os impactos que sua empresa sofrerá se o mesmo tomar essa ou aquela decisão.

Ao mesmo tempo em que eles tem a necessidade de obter informações sobre seus negócios, a quantidade de informações que está sendo acumulada nos sistemas de sua empresa e que não está sendo devidamente utilizada chega a ser algo exorbitante.

Sente-se, então, a necessidade de alguma tecnologia que possa incorporar um maior significado aos dados, estabelecendo padrões para que se possam efetuar previsões, classificar os dados, e enfim utilizar estes dados que estão de certa forma guardados inutilmente.

Decidiu-se então desenvolver um protótipo de Sistema de Apoio à Decisão que utilize a técnica de *Data Mining* (item 4). O *Data Mining* pode ser caracterizado pela execução de várias tarefas, mas neste caso escolheu-se implementar a tarefa de previsão. A tarefa de previsão é algo que os executivos solicitam muito e, na maior parte das vezes elas ajudam a tomar decisões, uma vez que indicam o que as decisões corretas e incorretas podem acarretar.

Este SAD deve ser um sistema parametrizável, onde o usuário tenha a liberdade de escolher as variáveis que ele deseja prever. Fazendo-o assim, o sistema será flexível o suficiente para o usuário prever qualquer tipo de informação que desejar.

6.2 LEVANTAMENTO DE REQUISITOS

Conforme os objetivos estabelecidos anteriormente e levando em conta as necessidades dos usuários, decidiu-se desenvolver um Sistema de Apoio à Decisão para Previsão genérica utilizando a técnica de *Data Mining*.

Baseando-se nos estudos de *Data Mining* (tabela 1) para efetuar a tarefa de previsão, decidiu-se utilizar o modelo de rede neural *Feedforward*, com aprendizado *Backpropagation*. A escolha deste modelo, deu-se principalmente pela grande capacidade de generalização e na sua rápida operacionalização. Segundo [BER97], este modelo de rede é muito utilizado para a previsão por causa destas características.

Tendo como base esses fatos, elaborou-se um SAD onde o usuário pode definir os parâmetros da rede neural, como suas iterações, variáveis de entrada e de saída, taxa de erros, etc, e, consequentemente, pode treiná-la e testá-la, para mais tarde tomar suas decisões com maior segurança.

Levando em consideração que uma das necessidades do usuário é a facilidade de uso do sistema, deve-se desenvolver uma interface amigável para o sistema. Tendo esse quesito como chave, decidiu-se desenvolver o sistema em cima da plataforma PC no ambiente operacional Windows. No desenvolvimento da aplicação, optou-se pelo ambiente Delphi 3, que permite um rápido desenvolvimento de aplicações para o ambiente Windows e suporta muito bem manipulação de dados com banco de dados.

Em relação ao banco de dados que irá armazenar as informações, optou-se pelo Sybase SQL Anywhere que, apesar de não ser um banco de dados de alta performance, foi escolhido pela sua flexibilidade e praticidade na utilização. Esse banco de dados pode rodar tanto em modo *standalone*, quanto em modo cliente/servidor; dependendo somente de sua configuração.

A partir daqui será apresentada a especificação do sistema onde foram utilizados o Diagrama de Contexto, o Diagrama de Fluxo de Dados, o Dicionário de Dados e o Modelo Entidade-Relacionamento. Para a especificação foram utilizadas as ferramentas PowerDesigner ProcessAnalyst e DataArchitect da Sybase.

6.2.1 DIAGRAMA DE CONTEXTO

O Diagrama de Contexto do SAD está apresentado na figura 9.

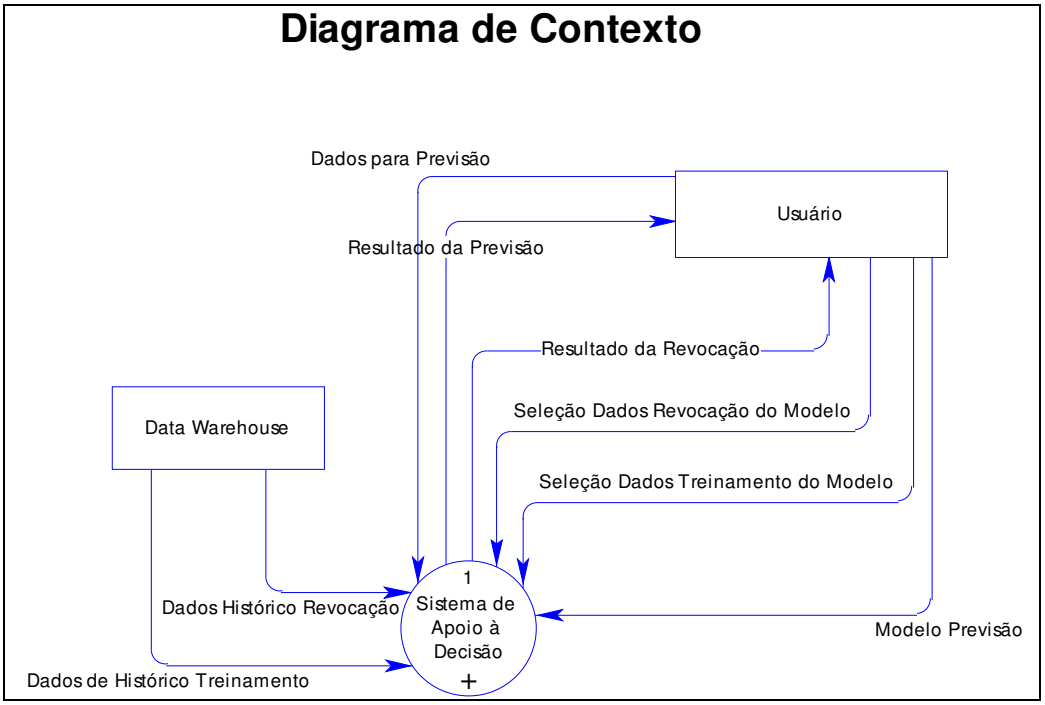


Figura 9 – Diagrama de Contexto do Sistema de Apoio à Decisão.

O Sistema irá interagir com o Usuário que fará a definição do Modelo de Previsão, o seu treinamento, revocação e, por fim, as previsões em cima do modelo definido. O *Data*

Warehouse fornecerá para o sistema as variáveis de entrada e saída e seus respectivos dados para o treinamento e revocação do modelo definido.

6.2.2 DIAGRAMA DE FLUXO DE DADOS

O DFD do sistema está apresentado na figura 10.

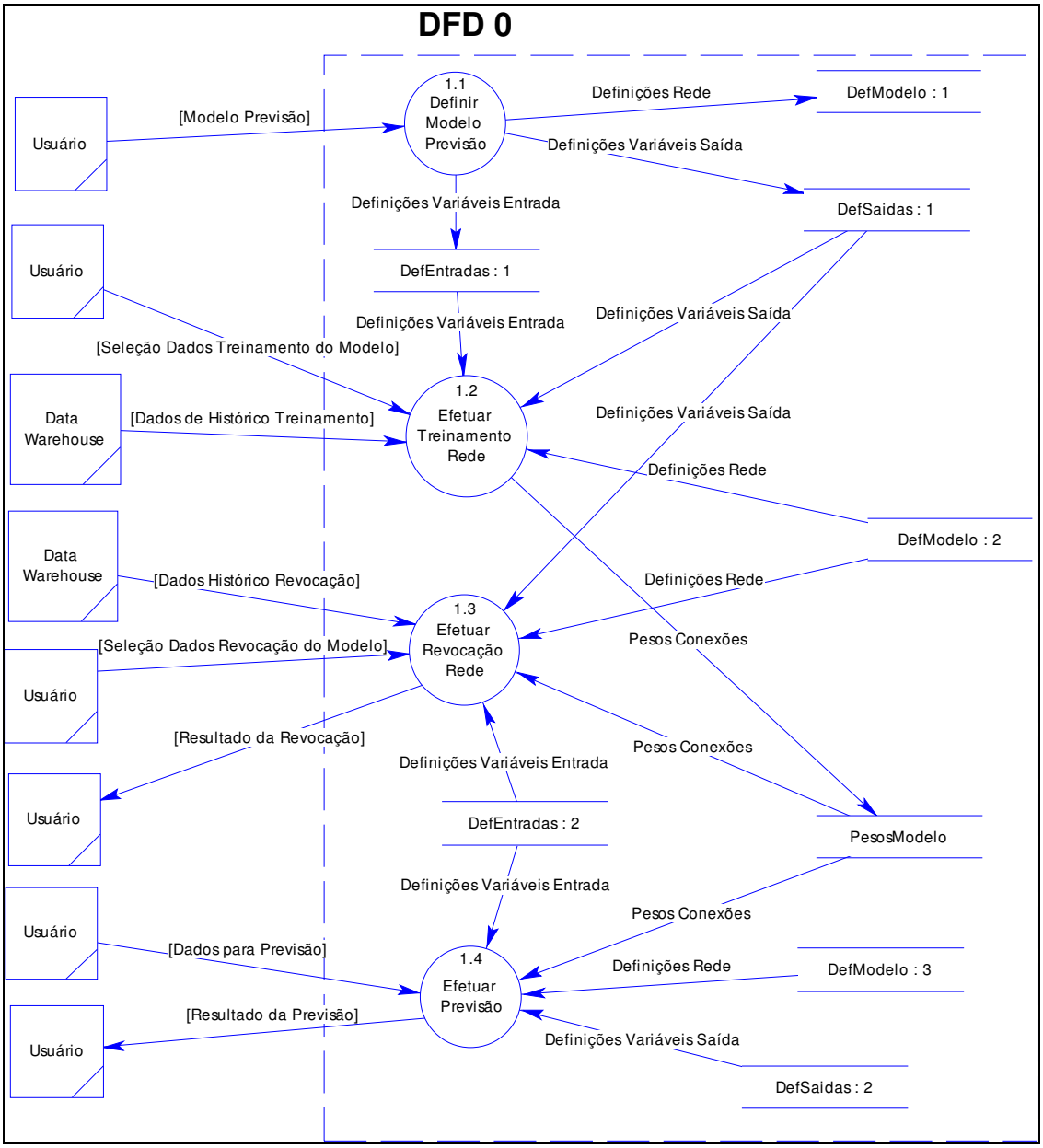


Figura 10 – DFD0 do Sistema de Apoio à Decisão.

Descreve-se a seguir os processos do DFD0:

a) definir o modelo previsão: neste passo o usuário está definindo o Modelo de Previsão que ele deseja efetuar. Um Modelo de Previsão é um conjunto de informações que irão influenciar diretamente na funcionalidade do sistema. Neste momento o usuário definirá informações como o nome do modelo, o número de iterações da rede para o treinamento, as variáveis de entrada e de saída com as respectivas regras de pré-processamento, a indicação de retreinamento da rede, a taxa de aprendizado e a taxa de erros;
b) efetuar o treinamento da rede: neste passo, após ter definido todas as informações do modelo de previsão, o usuário deverá treinar a rede selecionando os dados do *Data Warehouse* conforme desejar. Este passo é de extrema importância para o correto funcionamento do Modelo de Previsão, pois dependendo dos dados que forem utilizados para o treinamento da rede, seu comportamento pode ser totalmente diferente;
c) efetuar a revocação da rede: o passo de revocação consiste em verificar se a rede está respondendo conforme a aprendizagem à ela aplicada. O usuário então seleciona os dados a partir do *Data Warehouse* para efetuar a revocação. Se for verificado que a mesma ainda não está respondendo conforme o esperado, é o momento de efetuar um retreinamento;
d) efetuar previsão: esta etapa é caracterizada pela utilização do modelo treinado e revocado para efetuar finalmente a previsão dos dados. Neste momento o usuário entra com os dados das variáveis de entrada e a partir daí o SAD processa a rede neural para obter a resposta e mostrar o resultado para o mesmo.

6.2.3 DICIONÁRIO DE DADOS

Conforme especificado no Diagrama de Contexto, é apresentado parte do Dicionário de Dados:

a) modelo de previsão: nome do modelo, número de iterações da rede, variáveis de entrada, regras de pré-processamento das variáveis de entrada, variáveis de saída, regras de pré-processamento das variáveis de saída, indicação de retreinamento, taxa de aprendizado da rede, taxa de erros e comentário do Modelo;

b) dados de histórico para treinamento: dados do *Data Warehouse* para o usuário selecionar durante o treinamento;

c) seleção de dados para treinamento da rede: seleção de dados do *Data Warehouse* para treinamento (conforme definição de variáveis de entrada e de saída);

d) dados de histórico para revocação: dados do *Data Warehouse* para o usuário selecionar durante a revocação;

e) seleção de dados para revocação da rede: seleção de dados do *Data Warehouse* para revocação (conforme definição de variáveis de entrada e saída);

f) resultado da revocação: dados resultantes do processamento da rede neural (conforme definição de variáveis de saída);

g) dados de previsão: dados de entrada para serem previstos (conforme definição de variáveis de entrada);

h) resultado da previsão: dados resultantes da previsão (conforme definição de variáveis de saída).

6.2.4 MODELO ENTIDADE RELACIONAMENTO

O Modelo Entidade Relacionamento do SAD está apresentado na figura 11.

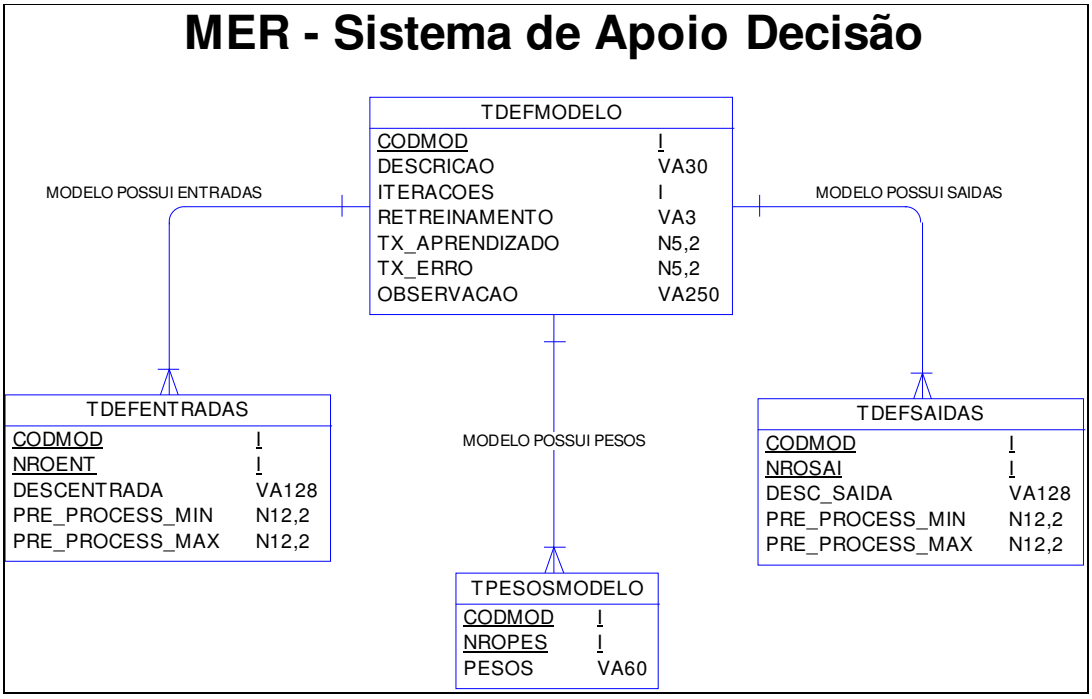


Figura 11 – MER – Sistema de Apoio à Decisão.

6.3 DESENVOLVIMENTO DO PROTÓTIPO

Levando em conta os objetivos propostos por este trabalho, construiu-se um Sistema de Apoio à Decisão que fosse flexível e de fácil utilização.

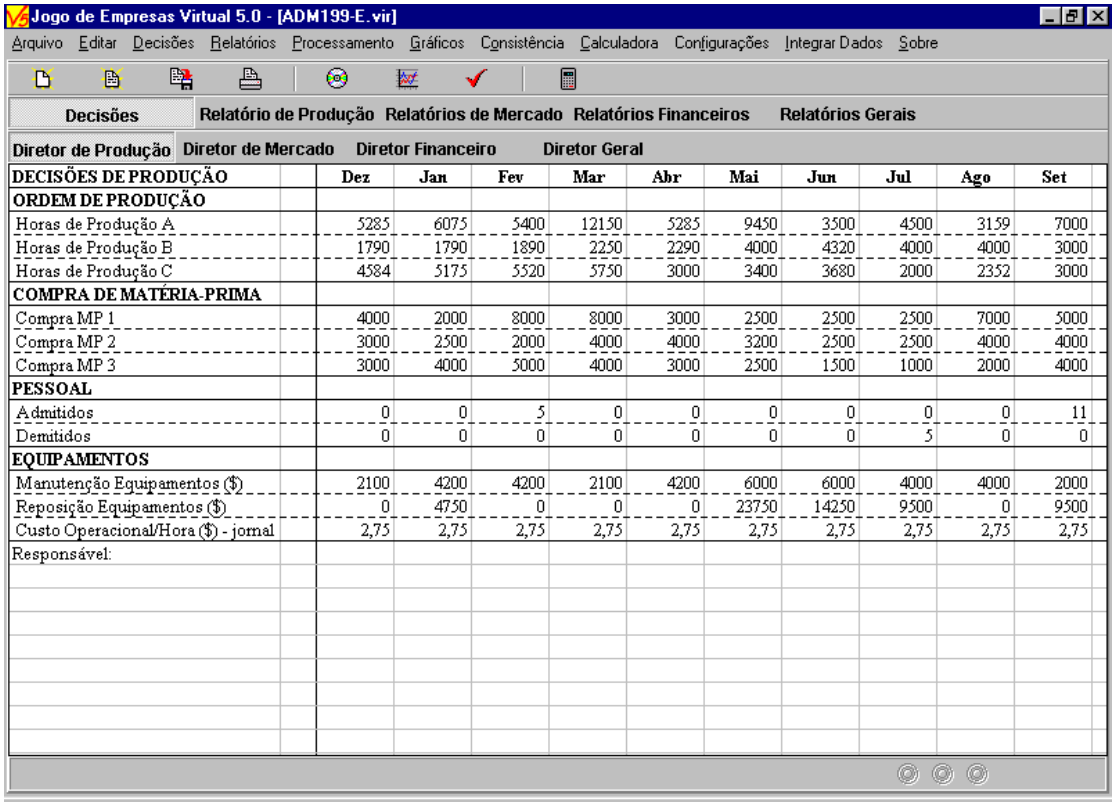
Aproveitando a flexibilidade da linguagem escolhida, resolveu-se utilizar um componente de Rede Neural que foi desenvolvido para a mesma. Este componente foi desenvolvido por [VAL97] e permitiu uma maior rapidez no desenvolvimento da aplicação.

6.3.1 AQUISIÇÃO DOS DADOS

Para treinar o modelo de previsão, é necessário que hajam informações concisas e integradas. Para isto utilizou-se o JE Virtual, que é um jogo que tem por objetivo reproduzir parcialmente e de forma simplificada uma situação que poderia ser real, de uma ou mais empresas em que se pretende estudar e conhecer as relações de causa e efeito que as caracterizam. Essa técnica caracteriza-se por oferecer um aumento de conhecimento, desenvolvimento de habilidades e a fixação de atitudes [MAR87].

O JE Virtual é um jogo que possui vários cenários, onde existem uma grande quantidade de variáveis de decisão e de resultados (figura 12). Estas variáveis foram então analisadas e lapidadas, de modo que fossem apenas catalogadas as que tivessem algum tipo de relacionamento de decisão.

O JE Virtual possui seus dados em arquivos no formato Excel, o que dificulta a manipulação dos mesmos para o treinamento da rede. Então decidiu-se que a fonte externa dos dados seria um *Data Warehouse* baseado nas informações catalogadas no JE Virtual. Para isto foi feita uma rotina de integração de dados (figura 12 – Menu Integrar Dados), a qual obtém os dados dos arquivos no formato do Excel e os armazena no *Data Warehouse*.



DECISÕES DE PRODUÇÃO	Dez	Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set
ORDEM DE PRODUÇÃO										
Horas de Produção A	5285	6075	5400	12150	5285	9450	3500	4500	3159	7000
Horas de Produção B	1790	1790	1890	2250	2290	4000	4320	4000	4000	3000
Horas de Produção C	4584	5175	5520	5750	3000	3400	3680	2000	2352	3000
COMPRA DE MATÉRIA-PRIMA										
Compra MP 1	4000	2000	8000	8000	3000	2500	2500	2500	7000	5000
Compra MP 2	3000	2500	2000	4000	4000	3200	2500	2500	4000	4000
Compra MP 3	3000	4000	5000	4000	3000	2500	1500	1000	2000	4000
PESSOAL										
Admitidos	0	0	5	0	0	0	0	0	0	11
Demitidos	0	0	0	0	0	0	0	5	0	0
EQUIPAMENTOS										
Manutenção Equipamentos ($)	2100	4200	4200	2100	4200	6000	6000	4000	4000	2000
Reposição Equipamentos ($)	0	4750	0	0	0	23750	14250	9500	0	9500
Custo Operacional/Hora ($) - jornal	2,75	2,75	2,75	2,75	2,75	2,75	2,75	2,75	2,75	2,75
Responsável:										

Figura 12 – Janela do JE Virtual

6.3.2 ARMAZENAMENTO DOS DADOS

Este *Data Warehouse* foi implementado em um banco de dados relacional, mais precisamente o Sybase SQL Anywhere em uma máquina *standalone*. A configuração do banco de dados foi feita desta forma porque o SAD desenvolvido não tem como objetivo funcionar para um ambiente multi-usuário.

A modelagem deste *Data Warehouse* se deu a partir das variáveis catalogadas no JE Virtual (figura 13); de forma que pudesse armazenar os dados do mesmo de maneira histórica.

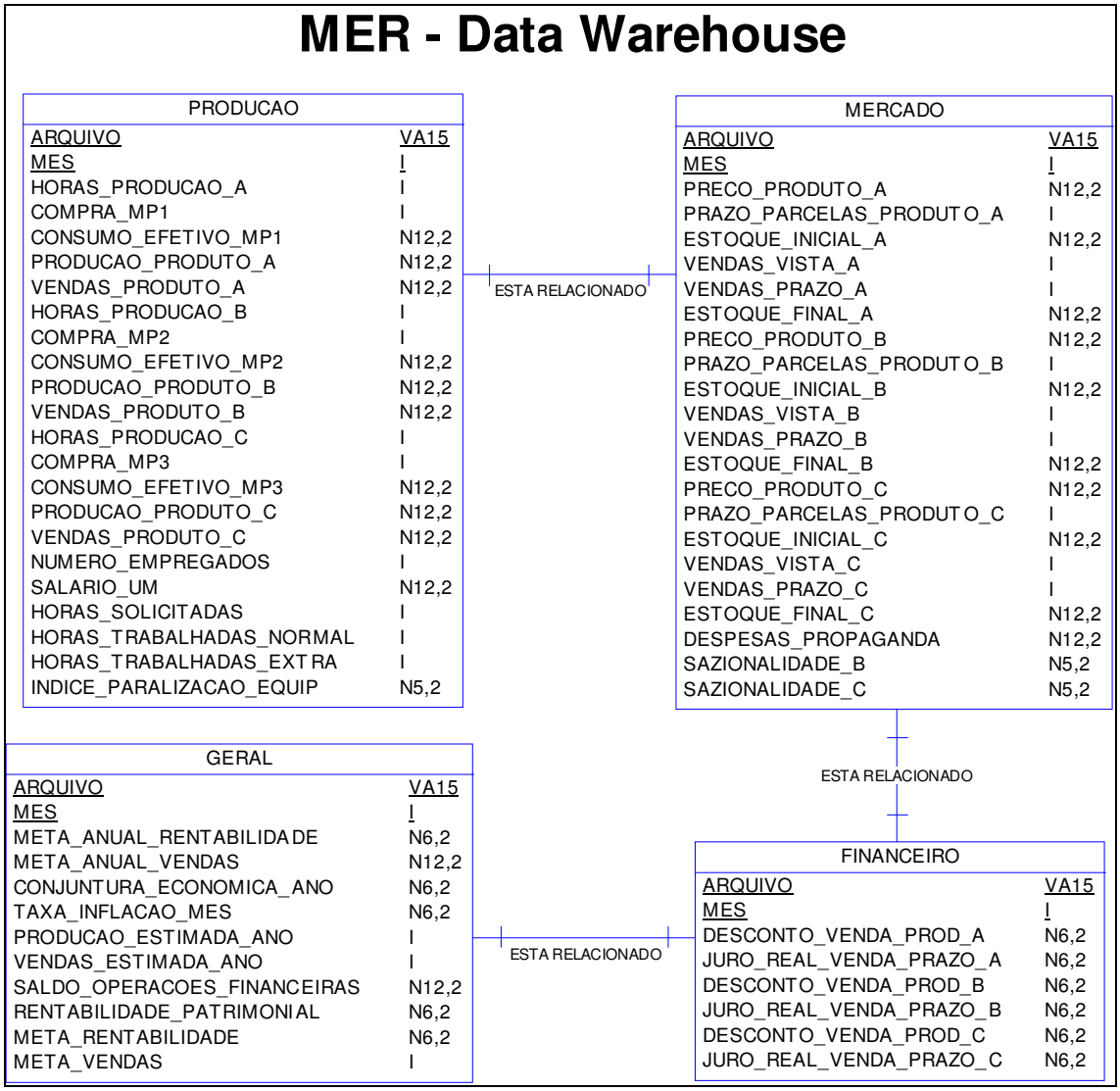


Figura 13 – MER - *Data Warehouse* do JE Virtual.

6.3.3 ACESSO AOS DADOS

Para o usuário usar as informações coletadas e armazenadas no *Data Warehouse* para tomar decisões, é necessário utilizar o Sistema de Apoio à Decisão que foi modelado. O protótipo possibilita ao usuário seguir as etapas de KDD que o *Data Mining* incorpora.

A seguir, faz-se uma analogia entre os processos do KDD e os passos que o sistema proporciona ao usuário.

6.3.3.1 DOMÍNIO DA APLICAÇÃO

Esta etapa do KDD é muito importante na aplicação do *Data Mining*, pois é onde o usuário deve analisar qual o conhecimento que ele deseja adquirir e quais os passos que ele deve seguir para chegar a esse resultado.

O usuário então deve estudar as variáveis que possui para chegar a um modelo de previsão, com suas respectivas variáveis de entrada e saída. É baseado nestas informações que o usuário definirá o modelo de previsão do SAD.

O modelo de previsão é primeiramente definido na tela principal do sistema (figura 14), descrevendo o código do modelo e sua descrição. O sistema permite ao usuário cadastrar quantos modelos de previsão o usuário necessitar, tornando assim esta ferramenta muito flexível e genérica para o processo de tomada de decisão. A partir da tela principal do sistema pode-se chamar outras telas, onde são executadas as próximas etapas do processo de KDD.

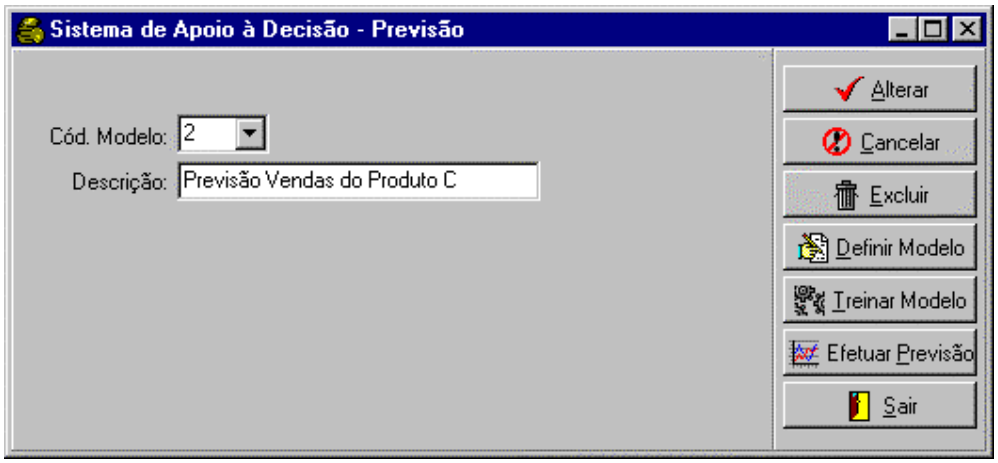


Figura 14 – Tela principal do sistema

Clicando-se no botão Definir Modelo (figura 14), pode-se definir o modelo de previsão. Ainda deve-se informar as variáveis de entrada e de saída com suas respectivas regras de pré-processamento, o número de iterações, a indicação de retreinamento e a taxa de erros da rede neural artificial. Estas informações podem ser informadas na tela de definições do modelo de previsão (figura 15).

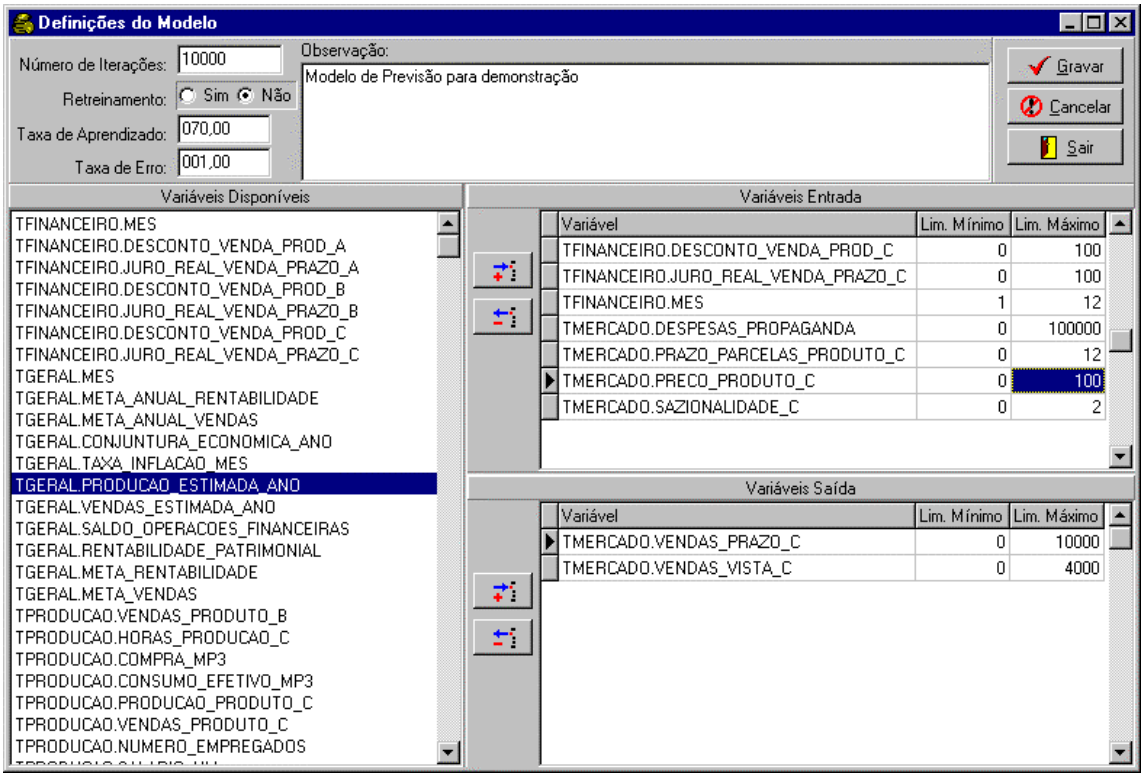


Figura 15 – Tela de definições do modelo de previsão

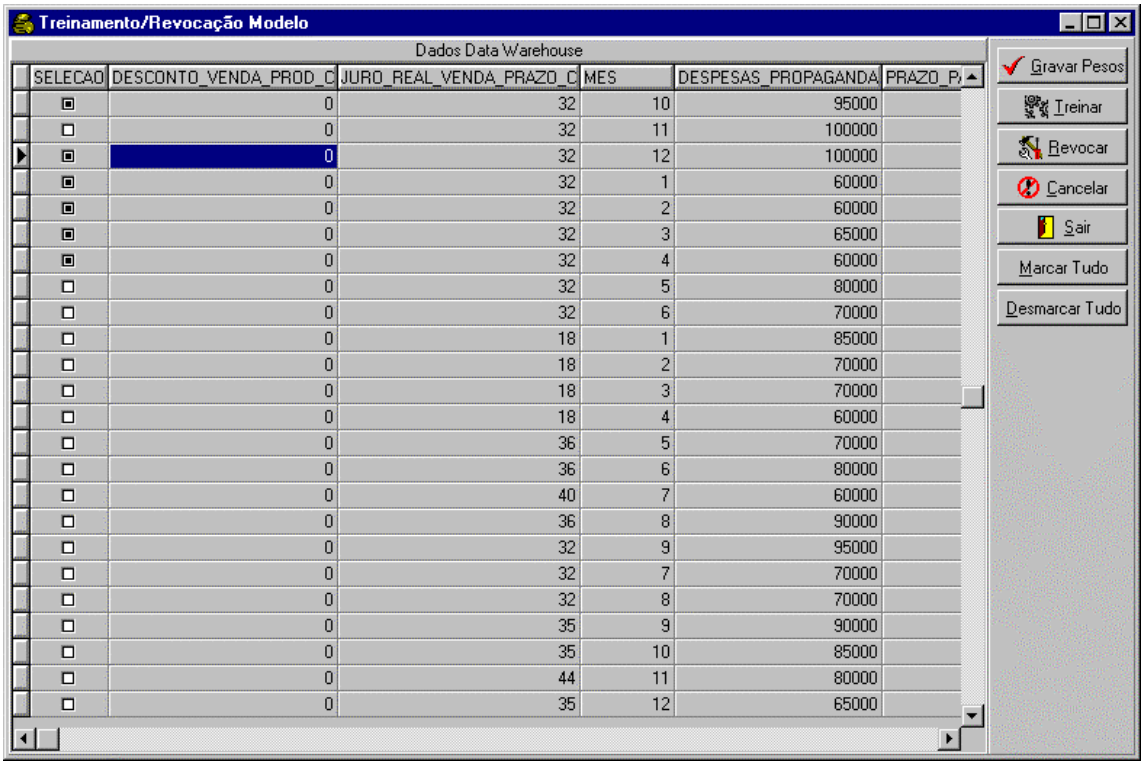
É importante salientar que as variáveis de entrada e saída são disponibilizadas a partir da definição do *Data Warehouse*.

6.3.3.2 SELEÇÃO DOS DADOS

Esta etapa é caracterizada pela seleção do conjunto de informações que serão utilizados no processo de *Data Mining*.

Neste ponto do processo, o usuário deve utilizar o domínio que possui sobre os dados do *Data Warehouse*. Levando isso em consideração, deve-se avaliar quais são as informações que ele deseja utilizar para o treinamento do modelo de previsão. Os dados que forem usados para o treinamento do modelo influenciarão diretamente na resposta do mesmo.

A tela de Treinamento e Revocação do modelo apresenta todo o conjunto de informações que foram extraídas do *Data Warehouse*, que é a fonte externa de dados do sistema (figura 16). É nesta tela que são selecionadas as informações que fazem parte do conjunto de treinamento.



SELECAO	DESCONTO_VENDA_PROD_C	JURO_REAL_VENDA_PRAZO_C	MES	DESPESAS_PROPAGANDA	PRAZO_P
☑	0	32	10	95000	
☐	0	32	11	100000	
☑	0	32	12	100000	
☑	0	32	1	60000	
☑	0	32	2	60000	
☑	0	32	3	65000	
☑	0	32	4	60000	
☐	0	32	5	80000	
☐	0	32	6	70000	
☐	0	18	1	85000	
☐	0	18	2	70000	
☐	0	18	3	70000	
☐	0	18	4	60000	
☐	0	36	5	70000	
☐	0	36	6	80000	
☐	0	40	7	60000	
☐	0	36	8	90000	
☐	0	32	9	95000	
☐	0	32	7	70000	
☐	0	32	8	70000	
☐	0	35	9	90000	
☐	0	35	10	85000	
☐	0	44	11	80000	
☐	0	35	12	65000	

Figura 16 – Tela de Treinamento e Revocação do Modelo

6.3.3.3 PRÉ-PROCESSAMENTO E LIMPEZA

A etapa de pré-processamento visa adequar as informações aos algoritmos de *Data Mining*. Os algoritmos de *Data Mining* na maior parte das vezes requerem os dados formatados para o seu processamento. As redes neurais que são usadas pelo SAD, necessitam que os dados de entrada e de saída sejam contínuos de 0,0001 à 0,0009.

Para efetuar a formatação dos dados selecionados para os dados requisitados pela rede neural, deve-se efetuar o pré-processamento. Os valores que serão processados pela rede são calculados em relação proporcional à definição do limite mínimo e máximo de pré-processamento para cada entrada ou saída de dados.

Os valores máximo e mínimo para o pré-processamento são definidos na tela de definições do modelo (figura 15).

6.3.3.4 DATA MINING

O *Data Mining* é a etapa onde se incorpora um algoritmo para o processamento dos dados, e este algoritmo determinará os padrões dos dados que resultam no conhecimento.

No caso do desenvolvimento deste protótipo de SAD para a previsão, decidiu-se utilizar a técnica de redes neurais.

Desta forma, o usuário deverá treinar o modelo com os dados da seleção. Durante o treinamento a rede ajustará os pesos entre suas conexões, a fim de encontrar o melhor padrão para atender ao conjunto de treinamento.

O treinamento e a revocação do modelo são efetuados através da tela de Treinamento e Revocação do modelo (figura 16).

Após ter sido feito o treinamento, o usuário poderá revocar o modelo para aprovar o aprendizado do mesmo. Se for julgado que o mesmo ainda não está devidamente ajustado, deverão ser repetidos os passos de seleção e treinamento até que o aprendizado esteja concluído. Para o usuário verificar como se processou a Revocação, o SAD mostrará a tela de Resultado da Revocação (figura 17).

Figura 17 – Tela de apresentação do resultado da Revocação

A partir do momento em que a rede estiver devidamente treinada e com seus pesos ajustados, é que pode-se efetuar a previsão como sendo o próximo passo do KDD.

6.3.3.5 INTERPRETAÇÃO DO CONHECIMENTO

Após a etapa de *Data Mining* estar concluída e os padrões do modelo de previsão estarem estabelecidos, conforme o andamento da etapa anterior, pode-se efetuar a previsão dos dados e então verificar se o modelo processado está condizente com o que foi descrito inicialmente.

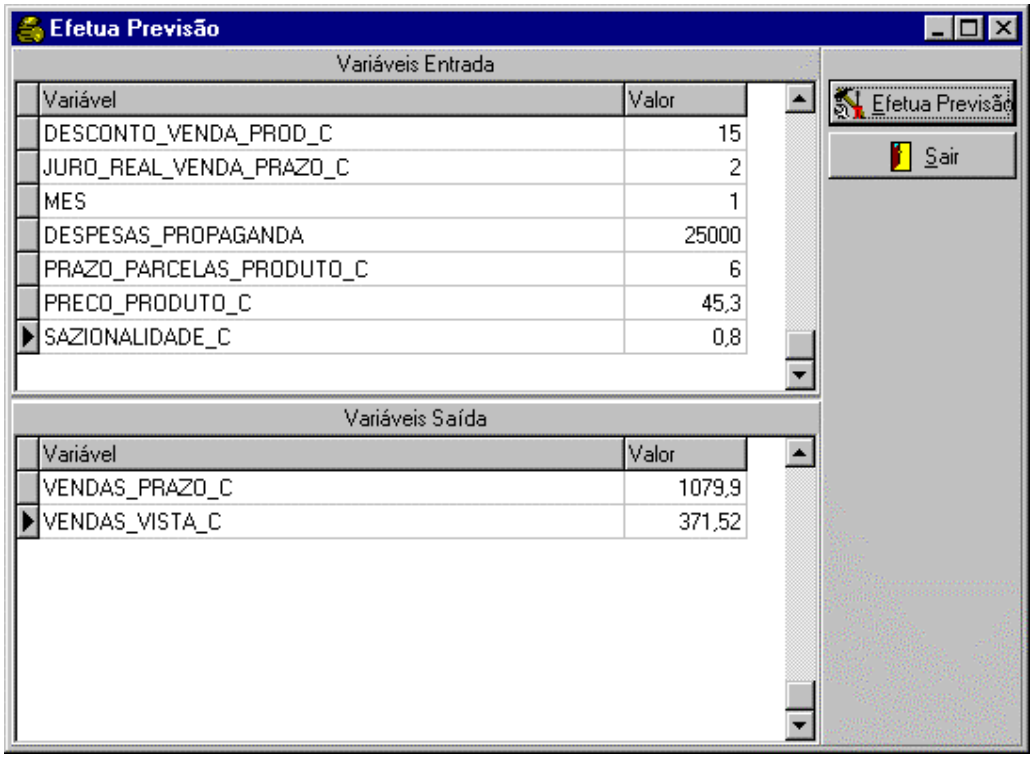


Figura 18 – Tela que efetua a previsão dos dados

Para efetuar a previsão, deve-se utilizar a tela do sistema que Efetua Previsões (figura 18). Nesta tela é informado um valor para cada variável de entrada e após isto deve-se efetuar a previsão, onde o sistema retorna os valores para as variáveis de saída.

7 CONCLUSÕES E SUGESTÕES

Este capítulo apresenta as conclusões, limitações e sugestões referentes ao trabalho desenvolvido.

7.1 CONCLUSÕES

Os sistemas tradicionais não proporcionam aos administradores de empresa praticamente nenhum tipo de ferramenta no qual os auxilie na tomada de decisões. Partindo dessa premissa, foi estudada a tecnologia de *Data Mining* que tem por finalidade adquirir conhecimento através da interpretação dos dados.

Foram estudados os seus conceitos e suas potencialidades e verificou-se que diferentemente das aplicações convencionais de bases de dados, que geralmente devolvem ao usuário informações baseadas em resultados de linguagens de consulta, o *Data Mining* devolve informações que são induzidas dos dados. Desta forma, informações que não existem podem ser previstas, com uma certa medida de acerto e exatidão. Além disso, padrões e tendências podem ser encontradas nos dados, o que pode levar à tomada de decisões mais adequadas e facilitar o trabalho de análise dos dados.

Neste trabalho foi ilustrado o uso de *Data Mining* com Redes Neurais empregado em um Sistema de Apoio à Decisão para construir modelos de Previsão genéricos. Tendo isso como base, verificou-se que a utilização do *Data Mining* juntamente com as etapas de KDD se mostrou bastante eficiente.

Foram realizados testes com os dados que foram integrados do JE Virtual para o *Data Warehouse* e o sistema se mostrou muito flexível para a definição de modelos de previsão ao mesmo tempo em que a utilização de Redes Neurais mostraram a sua grande capacidade de generalização para os problemas apresentados nos testes.

Mas no decorrer destes testes, verificou-se algumas desvantagens no uso de Redes Neurais:

a) aprendizado lento: o processo de aprendizado é muito lento;
b) conhecimento não é explícito: o conhecimento gerado não está representado na forma de regras e conceitos de padrões, e sim implicitamente na própria rede;

c) treinamento complicado: não é fácil estabelecer as regras de pré-processamento e escolher os dados corretos para obter um ótimo resultado com os modelos desenvolvidos. Isto requer um bom conhecimento de redes neurais e principalmente dos dados com que se está trabalhando.

Partindo deste princípio, verificou-se que a aplicação de *Data Mining* com Redes Neurais em Sistemas de Apoio à Decisão para Previsões genéricas pode ajudar em muito o processo de tomada de decisão não estruturada dentro de uma corporação; principalmente se forem seguidos os passos do KDD. Esta ajuda pode ocorrer principalmente se for levada em conta a enorme quantidade de dados que estão disponíveis nestas corporações.

Durante a construção do modelo, foram utilizadas algumas etapas/fases da metodologia de prototipação fundamental, as quais auxiliaram em muito no desenvolvimento do projeto. A linguagem Delphi ajudou muito pela facilidade de aprendizado que ela proporciona sobre novos recursos e o banco de dados Sybase SQL Anywhere também demonstrou que é um software de extrema facilidade de uso e confiável.

Encontrou-se grande dificuldade em encontrar material bibliográfico relativo ao Data Mining. Mesmo sendo efetuado um pedido para a compra de material para estudo, o mesmo material chegou somente na etapa final do desenvolvimento do trabalho.

Considera-se que o objetivo principal do trabalho, o desenvolvimento de um SAD para efetuar previsões genéricas utilizando *Data Mining*, foi atingido.

7.2 LIMITAÇÕES

O protótipo construído apresenta as seguintes limitações:

a) a fonte de dados para definição das variáveis de entrada e de saída é fixa, desta forma não permitindo ao usuário escolher uma variável fora do escopo apresentado;
b) as regras de pré-processamento são limitadas à faixa de valores (limite mínimo e máximo), sendo esta uma forma muito simples para se efetuar um pré-processamento.

7.3 SUGESTÕES

Sugere-se o estudo do Data Mining aplicando outras tarefas e técnicas para a tomada de decisões, como o uso de Árvores de Decisão para efetuar classificações.

Em relação a incorporação de uma fonte de dados externa (*Data Warehouse*) no Sistema de Apoio à Decisão, lembra-se que esta fonte externa neste caso é fixa. Deste modo, podem ser implementados outros sistemas onde esta fonte de dados seja flexível ao ponto de o usuário escolher de onde os dados virão.

Um outro item importante na questão da origem dos dados que poderia ser implementado, seria um acesso a dados que fosse além do Sybase SQL Anywhere. Sugere-se implementar acesso também à outros bancos como Oracle, Microsoft SQL Server, Sybase Server, Informix, etc.

Analisando o nível dos usuários que podem utilizar o sistema, uma outra sugestão seria construir uma interface voltada mais para os executivos, utilizando uma maior quantidade de recursos gráficos.

REFERÊNCIAS BIBLIOGRÁFICAS

[ALT92] ALTER, Steven. **Information systems: a management perspective.** USA : Addison-Wesley Publishing, 1992.

[AVI98] ÁVILA, Bráulio Coelho. **Data Mining.** VI Escola Regional de Informática da SBC – Regional Sul. Blumenau, 1998. p. 87-106.

[BER97] BERRY, Michael J. A.; LINOFF, Gordon. **Data mining techniques.** USA : Wiley Computer Publishing, 1997.

[BIS99] BISPO, Carlos Alberto F.; CAZARINI, Edson Walmir. Análises sofisticadas com o on-line analytical processing. **Developers' Magazine.** Rio de Janeiro, v 1, n. 32, abr. 1999.

[DAL99] DALFOVO, Oscar; GRIPA, Robson. Data warehouse: usando a técnica de cubo de decisão. **Developers' Magazine.** Rio de Janeiro, v 1, n. 32, abr. 1999.

[FAY96] FAYYAD, Usama M... [et all]. **Advances in knowledge discovery and data mining.** Mento Park : AAAI : MIT, 1996.

[FIG98] FIGUEIRA, Rafael Medeiros Andrade. **Miner: um software de inferência de dependências funcionais.** Rio de Janeiro, 1998. Trabalho de Conclusão de Curso – Instituto de Matemática, Universidade Federal do Rio de Janeiro.

[HAR98] HARRISON, Thomas H. **Intranet data warehouse.** São Paulo : Berkeley Brasil, 1998.

[INM97] INMON, William H. **Como construir o data warehouse.** Rio de Janeiro : Campus, 1997.

[LOE96] LOESCH, Claudio; SARI, Solange Teresinha. **Redes neurais artificiais : fundamentos e modelos.** Blumenau: FURB, 1996.

[MAC96] MACHADO, Carlos. Como dar o tiro certo na hora de decidir. **Exame Informática.** São Paulo, v. 11, n. 120, p. 27-29, mar. 1996.

[MAR87] MARTINELLI, Dante P. A **Utilização dos jogos de empresas no ensino da administração.** São Paulo, 1987. Dissertação (Mestrado em Administração) - Departamento de Contabilidade, USP.

[NIM98] NIMER, Fernando. Analisando o retorno sobre o investimento de data warehouse. **Developers' Magazine.** Rio de Janeiro, v 1, n. 18, fev. 1998.

[OLI98] OLIVEIRA, Adelize Generini de. **Data warehouse: conceitos e soluções.** Florianópolis : Advanced, 1998.

[PAL98] PALMA, Sérgio. Os componentes funcionais de um data warehouse. **Developers' Magazine.** Rio de Janeiro, v 1, n. 18, fev. 1998.

[SPR91] SPRAGUE, R. H., WATSON, H. J. **Sistemas de apoio à decisão: colocando a teoria em prática.** Rio de Janeiro : Campus, 1991.

[TAU98] TAURION, Cezar. Data warehouse: Vale a pena gastar milhões investindo em um? **Developers' Magazine.** Rio de Janeiro, v 1, n. 18, fev. 1998.

[TAU98a] TAURION, Cezar. O data warehouse será útil para a sua organização? **Developers' Magazine.** Rio de Janeiro, v 1, n. 18, fev. 1998.

[VAL97] VALDAMERI, Alexander Roberto. **Redes neurais aplicadas ao sistema de informação do jogo de empresas virtual.** Blumenau, 1997. Trabalho de Conclusão de Curso – Centro de Ciências Exatas e Naturais, Universidade Regional de Blumenau.

[WIL95] WILHELM, Pedro Paulo Hugo; LOPES, Maurício Capobianco, et al. Sistema inteligente de apoio à decisão. **Revista de Negócios.** Blumenau, v 1, n. 1, dez. 1995.

[WIL97] WILHELM, Pedro Paulo Hugo. **Uma nova perspectiva de aproveitamento e uso dos jogos de empresas.** Florianópolis, 1997. Tese (Doutorado em Engenharia de Produção) - Centro Tecnológico, UFSC.