UNIVERSIDADE FEDERAL DO PARANÁ

DEPARTAMENTO DE CIÊNCIA E GESTÃO DA INFORMAÇÃO

LEANDRO DE SOUZA NETTO

A MINERAÇÃO DE DADOS E O APOIO À GESTÃO ORGANIZACIONAL

CURITIBA
2007

LEANDRO DE SOUZA NETTO

A MINERAÇÃO DE DADOS E O APOIO À GESTÃO ORGANIZACIONAL

Monografia apresentada à disciplina Pesquisa em Informação II, do curso de Gestão da Informação, Departamento de Ciência e Gestão da Informação, Setor Ciências Sociais Aplicadas da Universidade Federal do Paraná.

Orientadora: Profª. Denise Fukumi Tsunoda

CURITIBA
2007

LEANDRO DE SOUZA NETTO

A MINERAÇÃO DE DADOS E O APOIO À GESTÃO ORGANIZACIONAL

Professora Doutora Denise Fukumi Tsunoda
Orientadora

Professor Doutor Newton Correia de Castilho Junior
Banca

Professor Mestre Celso Yoshikazu Ishida – UFPR
Banca

Curitiba,___de_____de 2007.

DEDICATÓRIA

A todas as pessoas que consideram a busca pelo conhecimento a forma mais digna de se obter crescimento tanto profissional, pessoal como também espiritual e, principalmente, àquelas que fazem esta busca com muita paz, amor e responsabilidade.

RESUMO

Nas últimas décadas as organizações têm investido cada vez mais recursos financeiros em tecnologia e sistemas de informação para armazenar grandes volumes de dados. Porém, estes dados, na maioria das vezes, não são processados e explorados de forma estratégica, para fornecer suporte à tomada de decisão. A mineração de dados, embora seja uma tecnologia ainda pouco utilizada no Brasil, pode ser um poderoso recurso a ser aplicado às organizações para processar dados brutos e descobrir conhecimento para apoiar o processo decisório organizacional. Este trabalho é caracterizado como uma pesquisa exploratória, cujo objetivo principal é abordar as funcionalidades da mineração de dados, dentro do processo de descoberta de conhecimento em bancos de dados, como suporte tecnológico aos gestores da informação e à tomada de decisão em organizações. Primeiramente é feita uma abordagem teórica no que diz respeito a definições, aplicações, principais algoritmos e tarefas associados à mineração de dados. Em seguida são realizadas análises de estudos de caso selecionados, no âmbito nacional, com o intuito de se observar na prática como a mineração de dados vem sendo aplicada no apoio à gestão organizacional.

Palavras-chave: *data mining*; mineração de dados; gestão organizacional; tomada de decisão; bancos de dados; descoberta de conhecimento; processo decisório; gestão da informação.

LISTA DE QUADROS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

DSI	–	Distribuição Seletiva de Informações
CSV	–	*Comma-Separeted Values*
EUA	–	Estados Unidos da América
FURB	–	Fundação Universidade Regional de Blumenal
GI	–	Gestão da Informação
ID3	–	*Induction of Decision Trees*
IME	–	Instituto Militar de Engenharia
IPARDES	–	Instituto Paranaense de Desenvolvimento Econômico e Social
KDD	–	*Knowledge Discovery in Databases*
KNN	–	*K-Nearest Neighbor*
MD	–	Mineração de Dados
SRI	–	Sistema de Recuperação de Informações
SQL	–	Structured Query Language
UFPR	–	Universidade Federal do Paraná
UFSC	–	Universidade Federal de Santa Catarina
UTL	–	Universidade Técnica de Lisboa
UTP	–	Universidade Tuiuti do Paraná

SUMÁRIO

1 INTRODUÇÃO

A partir do início dos anos 90, as organizações passaram a fazer maciços investimentos na área de tecnologia e sistemas de informação. Complexos programas e enormes redes de computadores começaram a ser utilizados para criação e alimentação de bancos de dados com quantidades de dados e informações cada vez maiores.

A mineração de dados (termo traduzido do inglês *data mining*) visa buscar relações e padrões relevantes entre os dados armazenados num banco de dados, apresentando como resultado a transformação destes dados brutos em conhecimento passível de ser utilizado na gestão organizacional, prestando um suporte significativo à tomada de decisão.

Neste trabalho, primeiramente é feita uma abordagem dos principais conceitos teóricos que envolvem a mineração de dados (MD) e, em segundo lugar, são desenvolvidas análises de estudos de caso de MD, no âmbito nacional, sob o ponto de vista da funcionalidade desta tecnologia no apoio à gestão organizacional.

1.1 PROBLEMA

A maioria das organizações de grande porte utiliza complexos sistemas de informação e grandes estruturas computacionais para armazenar informações vitais, tanto para a gestão estratégica e tática, como também nos processos operacionais propriamente ditos.

Diante deste fato, os bancos de dados das grandes organizações têm se tornado um emaranhado de dados brutos armazenados em enormes quantidades. Porém, a maior parte das empresas não costuma aproveitar esses dados de forma significativa. Por mais que os dados fiquem armazenados de forma bruta nos bancos de dados, se não passarem por um processo de limpeza, preparação, mineração de dados e análise, muitas informações relevantes à gestão estratégica das organizações podem ser desconsideradas, cumprindo uma função meramente operacional dentro do contexto corporativo.

Embora a MD seja um recurso extremamente versátil, que utiliza um conjunto de ferramentas para transformar dados brutos em conhecimento, ainda é uma tecnologia pouco explorada nas organizações. E dentro do acirrado mercado que se apresenta atualmente, esta é uma tecnologia fundamental tanto para os gestores da informação como também para as organizações.

1.2 JUSTIFICATIVA

O dinamismo do capitalismo faz com que o mercado a cada dia fique mais competitivo, exigindo ações estratégicas mais apuradas e ousadas das organizações. Os cenários mudam rapidamente e, assim como há empresas que crescem num curto espaço de tempo, há também situações em que companhias já consolidadas no mercado sucumbem em apenas alguns anos, devido muitas vezes à carência de uma correta gestão estratégica.

Os profissionais de Gestão da Informação podem utilizar eficientes ferramentas tecnológicas para aperfeiçoar seu trabalho, na constante busca pela melhoria da gestão estratégica nas organizações, com o intuito de vencer as inúmeras dificuldades do mercado.

Neste contexto, a MD surgiu como um poderoso aliado das organizações no processamento de dados brutos, resultante em informação e conhecimento essenciais para apoiar processos decisórios corporativos. Atualmente, é fundamental que seja estudada a funcionalidade desta tecnologia, tanto para as organizações, como também para os profissionais e estudantes de GI, pois estes certamente atuam ou pretendem atuar na área estratégica e/ou tecnológica de uma organização.

1.3 OBJETIVOS

Os objetivos estão divididos em objetivo geral e objetivos específicos.

1.3.1 Objetivo Geral

Abordar as funcionalidades da mineração de dados, dentro do processo de descoberta de conhecimento em bancos de dados, como suporte tecnológico aos gestores da informação e à tomada de decisão em organizações.

1.3.2 Objetivos Específicos

Os objetivos específicos definidos para o projeto são:

a) abordar os principais conceitos teóricos relacionados à MD;
b) analisar estudos de caso relacionados à utilização de MD no âmbito nacional;
c) identificar principais resultados obtidos através da aplicação da MD em organizações;
d) verificar como a MD pode apoiar a gestão estratégica e a tomada de decisão em organizações.

1.4 ESTRUTURA DO PROJETO

Neste trabalho, primeiramente é feita uma abordagem dos principais aspectos teóricos ligados à MD: aplicações, conceito, formas de representação do conhecimento, principais algoritmos e tarefas (capítulo 3).

No capítulo 4 são analisados estudos de caso de aplicações de MD realizados em organizações e delimitados ao âmbito nacional de pesquisa.

Finalmente, no capítulo 5, são apresentadas as considerações finais baseadas na revisão de literatura e nas análises de estudos de caso.

2 METODOLOGIA

2.1 CARACTERIZAÇÃO DA PESQUISA

De acordo com Vieira (2002, p. 5) a pesquisa exploratória utiliza métodos bastante amplos e versáteis, que compreendem desde o levantamento de informações em fontes secundárias e estudos de caso selecionados, até a observação informal (a olho nu ou mecânica) e levantamento de experiência.

Desta forma, esta pesquisa foi caracterizada como exploratória em seus objetivos, pois além de discorrer sobre os principais aspectos teóricos relacionados à MD, procurou também analisar estudos de caso que relacionassem a MD, em termos práticos, com processos decisórios organizacionais.

No que diz respeito aos meios utilizados para pesquisa, o estudo de caso foi escolhido como o mais adequado, tendo em vista que pode promover um aprofundamento na análise das características do objeto de estudo. Deste modo, efetuou-se neste trabalho uma abordagem qualitativa, através da análise de estudos de caso selecionados, nos quais havia utilização da tecnologia de MD como suporte para o processo decisório em organizações.

2.2 PROCEDIMENTOS METODOLÓGICOS

A pesquisa contém um embasamento teórico que elucida aspectos como definições, aplicações, tarefas e algoritmos da MD, bem como a contextualiza dentro do processo de Descoberta de Conhecimento em Bancos de Dados (KDD).

Na parte prática da pesquisa há análises de estudos de caso em que houve a aplicação da MD no âmbito organizacional, para que se pudesse verificar como esta tecnologia foi aplicada do ponto de vista metodológico e quais os resultados obtidos com ela.

Por último, foi feita uma análise dos resultados encontrados, sob o ponto de vista organizacional, comprovando ou refutando a premissa de que a MD é uma

tecnologia fundamental para auxiliar os gestores da informação e um importante recurso para otimizar a visão estratégica e o processo decisório das organizações.

2.3 ESTUDO DE CASO

Segundo Danton (2002, p. 18), o estudo de caso parte de uma lógica dedutiva, sendo considerado uma unidade significativa do todo. Deve ter três fases: seleção e delimitação do caso, trabalho de campo e organização e redação do relatório.

Neste trabalho os estudos de caso, bem como toda a coleta de dados propriamente dita, representam etapa crucial da pesquisa, obedecendo a seguinte estrutura:

a) identificação do problema: qual o principal fator motivante que levou a organização a utilizar a tecnologia da MD;
b) metodologia adotada: a descrição de como a MD foi aplicada à organização, identificando algoritmos, tarefas e método de tratamento de dados;
c) resultados encontrados: o detalhamento das regras, padrões e/ou modelos descobertos no processo de MD;
d) análise de resultados: como essas relações foram interpretadas pela organização para consolidar o conhecimento e o suporte à tomada de decisão.

Para a seleção dos estudos de caso incorporados ao trabalho foi necessária a adoção dos seguintes critérios de análise:

a) localidade: experimentos realizados no âmbito nacional como delimitação do universo de pesquisa;
b) algoritmos utilizados: somente os estudos de caso em que houve a utilização de algoritmos clássicos de MD, abordados na parte teórica do trabalho, foram selecionados;

c) detalhamento da metodologia: foram escolhidos somente os artigos de MD que apresentaram um detalhamento envolvendo o problema, a metodologia, os resultados encontrados e a caracterização de suporte à gestão organizacional.

A maioria dos estudos de caso analisados não atendia a estes critérios e, para não comprometer os objetivos da pesquisa, não foi incorporada ao trabalho. Por esta razão optou-se por uma abordagem qualitativa.

Finalmente, depois de realizadas as seleções e análises de estudos de caso, foi feita a interpretação dos resultados juntamente com as considerações finais do trabalho.

3 REVISÃO DE LITERATURA

3.1 O PROCESSO DE KDD

Para que o conhecimento possa ser descoberto em meio a dados brutos em um banco de dados há um longo processo compreendido por várias fases. Como a MD representa uma das etapas fundamentais desta tecnologia, muitos autores utilizam o termo para denominar o processo global de descoberta de conhecimento em bancos de dados.

Entretanto, neste trabalho este processo global é chamado de KDD - Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Databases*). A MD será abordada como a fase principal dentro deste processo, na qual muitos recursos podem ser aplicados com o intuito de se extrair padrões e relações não explícitas dos dados, com a finalidade de apoiar a gestão organizacional.

Segundo Fayyad (1996), o processo de KDD concretiza-se através de sete etapas: limpeza, integração, seleção, transformação, mineração de dados, avaliação dos resultados e apresentação do conhecimento.

Entretanto, a quantidade de etapas, a ordem em que ocorrem e o nome atribuído a cada uma delas, varia muito de acordo com o autor analisado. Em alguns casos, a limpeza e a integração são simplesmente chamadas de pré-processamento. A seleção muitas vezes é realizada após a transformação dos dados ou, então, ambas as fases são reunidas numa única etapa. Porém, o que é fundamental no processo de KDD, é que grandes quantidades de dados brutos precisam ser selecionadas e preparadas para que possibilitem a aplicação das diversas tarefas de mineração na busca por relações e padrões relevantes. A partir dessas relevâncias identificadas nos dados, deve ser feita uma análise para que haja a consolidação da descoberta de conhecimento.

Com objetivos didáticos, são elucidadas neste trabalho cinco etapas fundamentais dentro do KDD:

a) seleção: nesta etapa é feito o agrupamento do conjunto de dados que pretende ser minerado. Esta seleção deve ser feita com base na definição do problema a ser resolvido pelo KDD e também no entendimento do

domínio de aplicação onde será descoberto o conhecimento necessário para apoiar a tomada de decisão. Nesta fase é possível também incorporar dados externos ao conjunto selecionado, com o intuito de se obter mais dados relevantes à solução do problema;

b) pré-processamento: após o conjunto de dados ter sido selecionado, é bem possível que nele sejam encontrados registros duplicados, alimentados incorretamente ou até mesmo contendo a ausência de dados, caracterizando um conjunto de dados heterogêneo. Desta forma, são necessárias várias operações de limpeza de ruídos e pré-processamento, tais como: padronização do valor dos atributos, remoção de registros duplicados e tratamento de valores ausentes. Segundo Adriaans (2006) as etapas de seleção e pré-processamento podem corresponder a até 80% do tempo gasto em todo o processo de KDD;

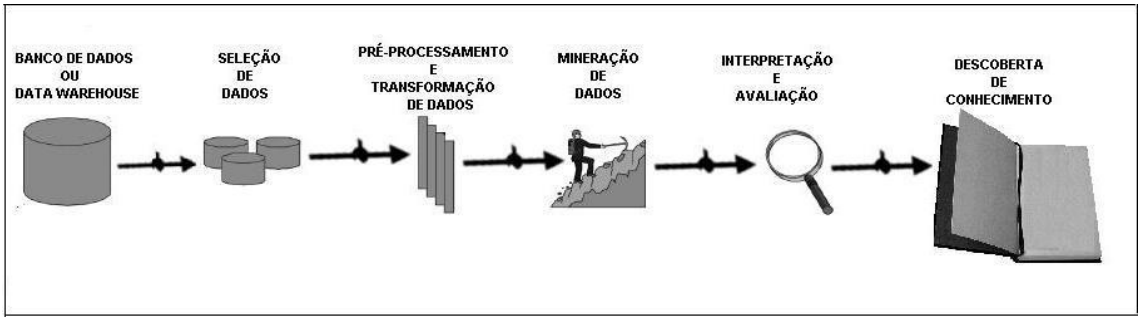
c) transformação: depois de pré-processados os dados precisam ser transformados para que fiquem armazenados da forma mais adequada para a aplicação das técnicas de mineração. Dentre as transformações mais comuns, há a atribuição de faixas nominais para dados numéricos, na qual valores contínuos são transformados em discretos como “grande”, “médio” e “pequeno”;

d) mineração de dados: nesta etapa os dados são processados através de tarefas e algoritmos de mineração, com a finalidade de serem identificados padrões válidos e relações relevantes dentro do conjunto de dados. Primeiramente, deve ser escolhida a tarefa de mineração conforme o tipo de conhecimento que espera-se que seja descoberto a partir dos dados. Em seguida, é necessário um algoritmo que atenda a tarefa de mineração eleita e que possa representar satisfatoriamente os padrões a serem encontrados. Essas técnicas podem muitas vezes ser combinadas para que sejam obtidos resultados melhores;

e) interpretação / avaliação: após identificados os padrões válidos e concluída a etapa de mineração, os resultados precisam ser analisados e interpretados pelos analistas e usuários envolvidos no processo. A finalidade dessa interpretação é verificar a validade ou, em alguns casos, a irrelevância dos padrões encontrados. Nos casos de não validação do conhecimento, é preciso que todas as etapas anteriores do KDD sejam

repetidas e aperfeiçoadas para que se encontrem resultados relevantes. A partir do momento que o conhecimento é validado, pode então servir de apoio à tomada de decisão.

FIGURA 1 – ETAPAS DO KDD



FONTE: O AUTOR

3.2 MINERAÇÃO DE DADOS

Nesta seção serão apresentados os principais aspectos teóricos vinculados à MD no que diz respeito a definições, aplicações e tarefas.

3.2.1 Definições

O termo MD é também conhecido pela expressão em inglês "*data mining"*. Segundo o Dicionário Houaiss da Língua Portuguesa (Houaiss, 2001, p. 1926), o termo mineração refere-se ao ato ou efeito de minerar, ou ainda, trabalho de extração do minério. Dado pode ser definido como o elemento inicial de qualquer ato de conhecimento e pode ser apresentado de forma direta e imediata à consciência, podendo servir de base ou pressuposto no processo cognitivo como um todo (Houaiss, 2001, p. 903).

Na exploração de minas, o material nobre, como o ouro por exemplo, é o objetivo primordial. No caso da MD as minas podem ser relacionadas aos bancos de dados e, o ouro (ou outro material nobre), ao conhecimento descoberto dentro desses bancos. Neste contexto podemos citar a definição de Han e Kamber (2001): "*data mining* é o processo de descoberta de conhecimento interessante a partir de

grandes quantidades de dados armazenados tanto em bancos de dados e *data warehouses* quanto em qualquer outro repositório de informação". Nesta definição há dois pontos importantes: o primeiro é que o termo MD se confunde com a própria descoberta de conhecimento em banco de dados propriamente dita. O segundo, é a menção do termo *data warehouse* que diz respeito a bancos de dados otimizados, projetados para o armazenamento de grandes volumes de dados, cujo propósito é fornecer suporte à tomada de decisão nas organizações. Ao contrário dos bancos de dados convencionais, que costumam armazenar dados operacionais de transações diárias, por exemplo, o *data warehouse* armazena dados analíticos, não-voláteis, ideais para dar suporte ao processo de tomada de decisão na gestão organizacional.

Berry e Linoff (1997) definem: "*data mining* é a exploração e análise, por meio automático ou semi-automático, de grandes quantidades de dados, com o objetivo de revelar regras e padrões significativos". Essa definição difere da anterior tendo em vista que, ao invés de vincular a MD diretamente com a descoberta de conhecimento, relaciona-a apenas às regras e padrões significativos obtidos nesse processo. Desta forma, para Berry e Linoff, a MD é apenas uma das etapas da descoberta de conhecimento em repositórios de dados.

Possivelmente a definição mais simples e clara para a MD seja a apresentada por Fayyad et al (1996): "[...] o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis".

Neste trabalho, considera-se a MD como sendo etapa crucial do KDD, levando em consideração a definição de Fayyad como a mais adequada para o escopo da pesquisa.

3.2.2 Aplicações

A MD ainda não é um recurso muito difundido nas empresas brasileiras. A maioria das organizações de pequeno e médio porte desconhecem esta tecnologia e desta forma ignoram os benefícios que ela pode proporcionar à gestão organizacional.

Mesmo tendo uma história recente, já existem casos clássicos de aplicações bem-sucedidas da MD. De acordo com Campos (2005), uma das maiores redes de varejo dos Estados Unidos, há alguns anos, ao buscar relações entre a venda de

produtos e os dias da semana, encontrou uma associação interessante: às sextas-feiras o número de vendas de fraldas aumentava na mesma proporção da venda de cervejas. A partir desta relação, descobriu-se que os compradores de fraldas tinham o mesmo perfil dos compradores de cervejas: homens que ao adquirir fraldas para seus filhos aproveitavam para levar também a tradicional cerveja do fim de semana. Ao tomar ciência disto, a empresa decidiu manter alguns fardos de cerveja próximos às fraldas e isso fez com que a venda destes dois produtos aumentasse.

Embora a mineração de dados seja mais comum em grandes multinacionais, a sua aplicabilidade estende-se a organizações de todos os portes e áreas de atuação. Carvalho (2005, p. 5-6) cita alguns exemplos importantes da utilização desta tecnologia em diferentes contextos. O governo dos EUA identifica, através da MD, padrões de transferência de fundos internacionais com características de lavagem de dinheiro, o que facilita muito as investigações e o combate à criminalidade.

Na medicina já é possível a criação e manutenção de grandes bancos de dados com informações sobre doenças, sintomas, resultados de exames e perfis de pacientes. A mineração destes dados permite, por exemplo, a descoberta de relações entre tipos de doença e certos perfis profissionais ou hábitos pessoais, auxiliando o diagnóstico e até mesmo facilitando a compreensão e o tratamento das doenças.

Na área financeira, bancos utilizam as técnicas de MD para identificar clientes com menor risco de se tornar inadimplentes nos empréstimos financeiros. Ao obter informações como idade, tempo de serviço e faixa de renda do cliente, por exemplo, as instituições financeiras se utilizam da MD para encontrar regras objetivas e claras sobre bons e maus pagadores, podendo aplicar estas regras para aumentar a taxa de sucesso das operações de empréstimo.

Desta forma, percebe-se que a MD pode ser aplicada em inúmeros contextos, independentemente do porte e ramo de atuação da organização, sendo uma ferramenta extremamente versátil na identificação de padrões e relações em repositórios de dados.

3.2.3 Principais Tarefas

A MD tem por objetivo principal identificar modelos de dados que se enquadram dentro de duas categorias: previsão e descrição. Esses modelos são obtidos através das chamadas tarefas. A seguir será apresentado um detalhamento a respeito das principais tarefas aplicadas na MD, bem como os métodos mais utilizados em cada uma delas:

a) classificação: é uma tarefa preditiva, ou seja, gera modelos que permitem que o comportamento dos dados possa ser previsto, considerada por alguns autores a tarefa mais comum da MD. Segundo Fayyad (1996) a classificação mapeia um item de dado em uma de várias classes previamente definidas. Em outras palavras, a classificação procura associar cada registro de um banco de dados a uma classe predefinida. Uma vez definida a classe de um registro, esta tarefa pode prever a classe de novos registros. Os métodos mais utilizados na tarefa de classificação são: regras determinísticas, árvores de decisão, regras probabilísticas, redes neurais e método bayesiano (probabilidade de características);
b) associação: esta tarefa é descritiva e procura identificar regras ou padrões significativos nos conjuntos de dados analisados. Dentre as principais funções desta tarefa, Carvalho (2005, p. 22) relata a determinação da probabilidade razoável de fatos que ocorrem (co-ocorrência) simultaneamente e também o cálculo das chances de dois ou mais itens estarem presentes juntos (correlacionados). Os algoritmos de associação são aplicados para se identificar regras e padrões em um conjunto de dados, como por exemplo: Apriori (AGRAWAL & SEIKANT, 1994) e IGART (DOMINGUES & REZENDE, 2004);
c) previsão: de acordo com Campos (2005) “esta função de mineração prediz os possíveis valores de alguns dados perdidos ou a distribuição de valores de certos atributos em um conjunto de objetos. De acordo com as principais aplicações desta tarefa Carvalho (2005, p. 21) menciona que a previsão pode: “determinar se o índice Bovespa subirá ou descerá amanhã, quanto o valor de uma dada ação da bolsa variará no próximo pregão, qual será a

população de uma certa cidade daqui a dez anos, entre outras". Dos principais métodos utilizados podem ser citadas as redes neurais e os algoritmos genéticos;

d) agrupamento: tarefa descritiva cuja finalidade é identificar coleções de dados semelhantes, pertencentes a uma mesma categoria ou classe. Fayyad (1996) ressalta que, diferentemente da classificação, em que as classes são predeterminadas, na tarefa de agrupamento as classes são estabelecidas a partir dos dados analisados. São identificados conjuntos de agrupamentos naturais que podem ser isolados em si ou constituídos por uma representação hierárquica. Dentre os principais métodos desta tarefa há as redes neurais e algoritmos genéticos.

3.3 FORMAS DE REPRESENTAÇÃO DO CONHECIMENTO

Nesta seção são apresentadas as principais formas de se representar o conhecimento por meio da MD. São feitas abordagens, respectivamente, sobre árvores de decisão, regras de classificação, regras de associação, regras de exceção e também análise de agrupamento.

3.3.1 Árvores de decisão

As árvores de decisão geram modelos de dados que permitem a classificação e predição de amostras desconhecidas com base na característica destes modelos. A partir da geração destas árvores, podem-se classificar as amostras desconhecidas sem necessariamente testar todos os valores dos seus atributos. As árvores de decisão são construídas classificando-se cada item de dado em uma de várias classes previamente definidas, agrupando todo o conjunto de classes numa estrutura arbórea que contém nós e folhas. E, pelo fato de haver a necessidade de se conhecer as classes de cada registro do conjunto de dados, os algoritmos de classificação por árvores de decisão são considerados algoritmos supervisionados.

O objetivo dos algoritmos de árvores de decisão é a criação de uma árvore na qual cada nó indica o teste de um atributo. De acordo com Frank & Witten (2005, p.

62) normalmente o teste em um nó compara o valor de um atributo com uma constante. Entretanto, algumas árvores comparam dois atributos um com o outro, ou ainda utilizam uma função que correlaciona um ou mais atributos.

Geralmente, a escolha de atributos é feita com base nos menores valores de entropia encontrados. Carvalho (2005, p. 158) define entropia neste contexto como "a quantidade de informação adicional necessária para se entender um fenômeno ou sistema". Para calcular a entropia de um determinado atributo (coluna) da base de dados é necessário, inicialmente, calcular a entropia dos valores possíveis desse atributo a partir da fórmula proposta por Shannon (1948, p. 13), exposta na figura 2. Na qual $A = v_j$ significa que o atributo A tem o valor v_j, n é o número de classes diferentes e p(i) é a probabilidade de um registro pertencer à classe c_n.

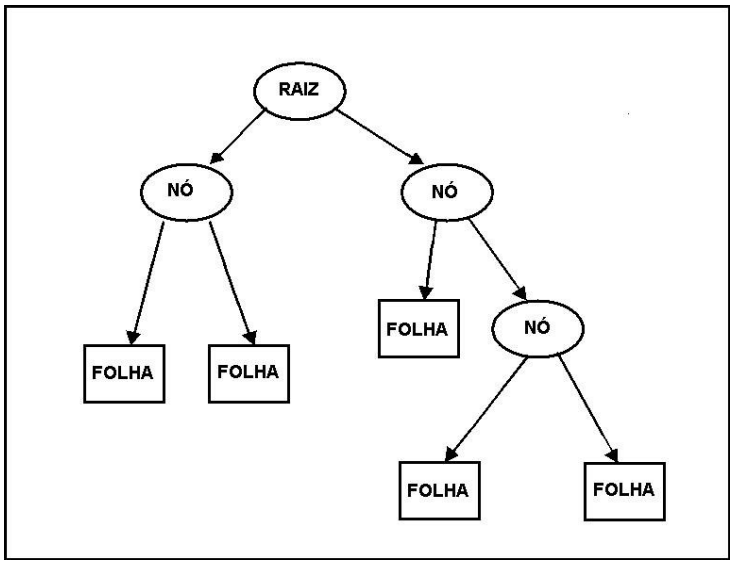
FIGURA 2 – FÓRMULA DO CÁLCULO DA ENTROPIA

$$E(A = v_j) = -\sum_{i=1}^{n} p(i) \times \log_2(p(i))$$

FONTE : O AUTOR

Quanto menor o valor da entropia encontrado, maior será o ganho da informação, isto é, maior será a qualidade de classificação do atributo. Em outras palavras, o atributo será mais previsível, com comportamento mais óbvio e conjunto de dados mais homogêneo, quanto menor for o valor de sua entropia e maior o ganho da informação. Desta forma, pode-se afirmar que o atributo que melhor classificar os dados deve ser escolhido como um nó da árvore, conforme detalhado na figura 3.

FIGURA 3 – ÁRVORE DE DECISÃO



FONTE: O AUTOR

Além de classificar e predizer amostras desconhecidas, analisando-se a árvore gerada é possível estabelecer regras de classificação, com o objetivo de melhor sintetizar e representar o conhecimento extraído na mineração. Dentre os algoritmos mais comuns aplicados à representação de árvores de decisão, tem-se o ID3 (QUINLAN, 1986) e o C4.5 (QUINLAN, 1993).

3.3.2 Regras de classificação

A tarefa de classificação necessita que um determinado método seja aplicado de acordo com o conhecimento que se deseja obter da base de dados. Como já mencionado, o conhecimento representado através do método de árvores de decisão pode ser extraído e representado na forma de regras de classificação. Na verdade, de acordo com Frank & Witten (2005, p. 67) as regras de classificação representam uma síntese do que pode também ser obtido através da leitura da árvore por si só, uma vez que cada regra é um caminho na árvore, da raiz até uma das folhas. Em muitas situações, as regras são muito mais compactas que as árvores, principalmente se for possível obter uma regra “padrão” que cubra casos não especificados pelas outras regras.

Vasconcelos (2002, p. 26) afirma que os dois modelos mais conhecidos de regras de classificação são:

a) regras de classificação *stricto sensu*: esse modelo obedece o seguinte formato: se “a” (condição), então “b” (classificação), cuja interpretação mostra que, se os valores assumidos pelos atributos de um registro satisfazem as condições do antecedente da regra, então, o registro recebe a classe indicada pelo valor do atributo de classificação;
b) regras de classificação indiretas: obtidas sob a forma de árvores de decisão, numa seqüência hierárquica de testes construídos ao longo de uma estrutura arbórea, com os nós folhas da árvore representando diferentes classes. Percebe-se, assim, que cada árvore pode exprimir diferentes regras de classificação.

Seja qual for o modelo de regras de classificação adotado, o desempenho do algoritmo e a qualidade das regras obtidas irá depender diretamente da forma que o conjunto de dados foi selecionado, pré-processado e transformado nas etapas que antecedem a aplicação da mineração de dados propriamente dita. Dentre os inúmeros algoritmos que geram regras de classificação, há os algoritmos genéticos, que posteriormente serão melhor abordados neste trabalho.

3.3.3 Regras de associação

As regras de associação são normalmente utilizadas para encontrar relações de associação e correlação de itens em grandes conjuntos de dados. Segundo Frank & Witten (2005, p. 69), as regras de associação são somente diferentes das regras de classificação porque podem predizer qualquer atributo, não apenas a classe, e isto dá a liberdade para predizerem também combinações de atributos. De acordo com Agrawal & Seikant (citado por Domingues, 2004, p. 17), as regras de associação caracterizam o quanto a presença de um item implica a presença de (pelo menos) outro item no mesmo conjunto de dados. Desta forma, a

associação pode ser classificada como uma atividade de mineração de dados descritiva.

Relacionados às regras de associação, normalmente são encontrados também os termos de suporte e confiança. Em artigo publicado pela UTL (2005) estes termos são definidos como:

a) suporte**:** o suporte de uma regra é uma medida de quão freqüente os itens envolvidos, nessa regra, ocorrem juntos, usando-se noção probabilística.
b) confiança**:** a confiança de uma regra de associação (X–>Y), quantifica com que freqüência X e Y ocorrem juntos como fração do número de vezes que X ocorre. Por exemplo, se a confiança for 50%, X e Y ocorrem juntos em 50% das vezes em que X ocorre. Sabendo que X ocorre, a probabilidade de Y também ocorrer nesse mesmo registro é de 50%.

Uma regra de associação que poderia ser encontrada num banco de dados corporativo, por exemplo, seria o indicativo de que 90% dos clientes que compram o produto X, também adquirem o produto Y. Neste exemplo, 90% corresponderiam à confiança da regra.

Normalmente um algoritmo de regras de associação pode encontrar inúmeras regras semelhantes a esse modelo. Cabe ao usuário selecioná-las de acordo com um valor mínimo de suporte e confiança que garanta a consistência da relação e atenda de forma eficiente aos objetivos da mineração de dados. Dentre os algoritmos que geram regras de associação, o mais utilizado é o Apriori.

3.3.4 Regras de exceção

Retornando às regras de classificação, uma extensão natural é permitir que elas possuam exceções. Então, segundo Frank & Witten (2005, p. 70) é preferível que modificações incrementais sejam feitas em um conjunto de regras ao invés de se remodelar todo o conjunto.

Um dos principais motivos que estimulam a geração de regras de exceção é o grande volume de padrões que contêm redundâncias ou modelos irrelevantes, frequentemente encontrados na mineração de dados, o que muitas vezes dificulta a

interpretação dos resultados e, por conseqüência, inviabiliza o seu uso no apoio à tomada de decisão.

É interessante enfatizar que uma regra de exceção deve ser vista como o incremento de uma regra comum e, portanto, contradiz o padrão previsto por esta última. Este método assume que regras comuns estabelecem padrões conhecidos pelo usuário, levando-se em consideração que elas são aplicadas a grandes volumes de dados, ao contrário das regras de exceção, que em geral são desconhecidas e se aplicam a uma pequena porção minoritária de registros.

A partir de um conjunto de regras gerais do tipo se “a”, então “b”, por exemplo, o formato da regra de exceção poderia ser: se “a” e “c”, então (não) “b”. Desta forma, as regras de exceção tendem a ser surpreendentes, por representarem uma contradição em relação às regras de classificação comuns. Algoritmos baseados em lógica *fuzzy* representam um bom exemplo que obedece ao princípio das regras de exceção.

3.3.5 Análise de agrupamento

Análise de Agrupamento, de acordo com Azambuja (2005, p. 32), corresponde a um conjunto de métodos cujo objetivo é identificar padrões e formar grupos homogêneos (dados semelhantes pertencem a um mesmo grupo) a partir de n observações ou elementos existentes em um banco de dados. Pelo fato deste método não se basear na existência de classes previamente definidas no conjunto de dados, os algoritmos de agrupamento são considerados não-supervisionados.

Esta forma de análise é uma maneira clássica de se efetuar pesquisas exploratórias da MD em bancos de dados cujo conjunto de registros é pouco conhecido e há grande número de objetos, o que consequentemente dificulta a exploração dos dados por meio de análises meramente humanas. Desta forma, a análise de agrupamento organiza os dados em estruturas que facilitam sua interpretação, gerando grupos baseados em semelhanças e distinções identificadas num conjunto de dados.

Segundo Frank & Witten (2005, p. 81), alguns algoritmos de agrupamento permitem que uma instância pertença a mais de uma classe, desta forma, as

instâncias são agrupadas em duas (ou mais) dimensões, representadas através de um diagrama.

O resultado de uma análise de agrupamento apresenta um arranjo dos objetos numa escala de distância, cuja finalidade é identificar a afinidade entre os grupos. Portanto, a subdivisão de um conjunto de dados em grupos homogêneos é o objetivo principal da análise de agrupamento. Independente de qual o método de agrupamento utilizado, a qualidade de seu resultado vai depender muito da correta interpretação dos grupos formados. Para tanto, muitas vezes é necessário que haja um bom conhecimento prévio do conjunto de dados por parte do analista, para que os grupos sejam entendidos com clareza.

De qualquer forma, a análise de agrupamento é a representação clássica do conhecimento obtido através de métodos não-supervisionados na mineração de dados. Um dos algoritmos mais utilizados dentro deste contexto é o KNN (*K-Nearest Neighbor*), também conhecido como “método do vizinho mais próximo”. Na seção de algoritmos o KNN é abordado com mais detalhes.

3.4 PRINCIPAIS ALGORITMOS

Nesta seção faz-se uma breve explicação do funcionamento dos principais algoritmos utilizados na mineração de dados.

3.4.1 ID3

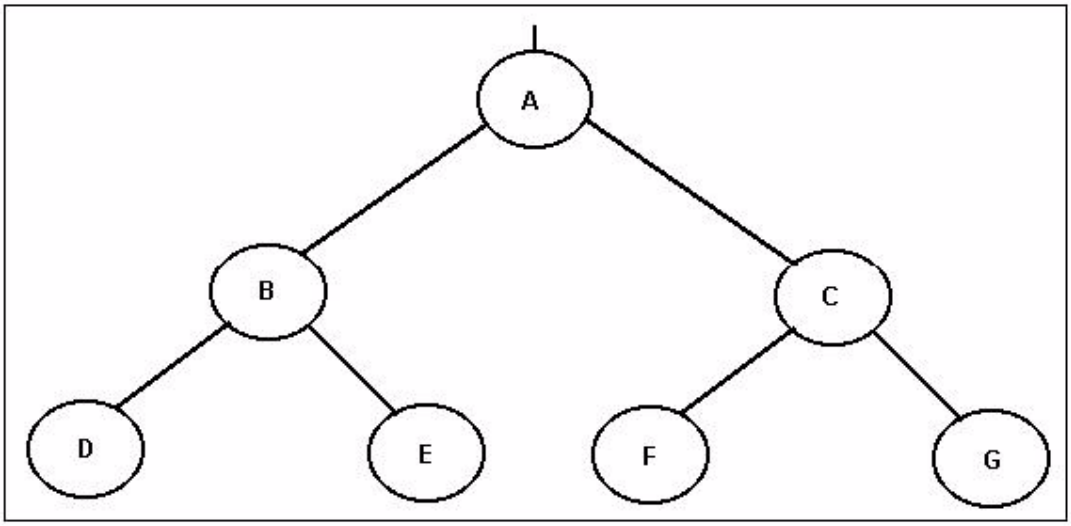
O ID3, algoritmo proposto por Ross Quinlan (1986), é um dos mais utilizados para a geração de árvores de decisão. Seu nome é uma sigla, cujo significado em inglês é *Induction of Decision Trees* ou, em português, indução de árvores de decisão. Sua estrutura é baseada no critério da entropia para selecionar os atributos geradores da árvore de decisão.

Segundo Vasconcelos (2002, p. 40) o ID3 analisa o conjunto de dados e constrói a árvore a partir de sua raiz. Primeiramente, escolhe-se o atributo “a” com a melhor função de avaliação (baseado no menor valor de entropia calculado) para

particionar estes dados. Para cada valor “i” do atributo “a”, um ramo “r” é criado junto com o correspondente subconjunto de dados que possuem o valor de “a = i”.

Deste modo, para cada ramo “r” é criado um nó “n” na árvore que poderá ser categorizado como uma classe ou um outro atributo. Caso os exemplos analisados representem a mesma classe no padrão “a = i”, o nó será atribuído a esta classe. Porém, se os exemplos revelarem a existência de outra classe no padrão “a = u”, por exemplo, o nó será considerado um outro atributo. Portanto, com base nesta lógica é gerada a árvore de decisão, conforme figura 4.

FIGURA 4 – ESTRUTURA DE UMA ÁRVORE



FONTE: O AUTOR

Embora o ID3 seja normalmente um algoritmo com bom desempenho e um dos mais populares na indução de árvores de decisão, outros algoritmos baseados em sua estrutura foram desenvolvidos posteriormente com recursos incrementais.

3.4.2 C4.5

Proposto também por Quinlan (1993), o algoritmo C4.5 foi desenvolvido a partir do ID3 e é também aplicado à indução de árvores de decisão. Em comparação com este último, o C4.5 apresenta algumas vantagens, pois permite minerar atributos numéricos, lidar com registros vazios e dados com ruído, além de gerar poda (síntese da árvore original) e regras de classificação a partir das árvores de decisão.

De acordo com Vasconcelos (2002, p. 43), o algoritmo C4.5 normalmente gera regras de classificação por um método chamado poda postergada. Primeiramente, é gerada uma árvore de decisão. Esta árvore então é convertida em regras de classificação (uma regra para cada ramo da árvore). Estas regras são generalizadas com a remoção dos termos redundantes e inconsistentes para ser gerada uma árvore podada, resultando numa espécie de síntese da árvore original.

Por último, as árvores podadas e não podadas são comparadas com o conjunto de dados inicial para confirmar a consistência das regras geradas. Para cada árvore podada é gerado um conjunto de regras, no qual cada regra normalmente deve conter atributos, valores, a classificação e a porcentagem que indica a exatidão da regra, conforme exemplo a seguir:

Se “Sexo” = “Feminino” e “Idade” > 40, então “Assiste Novelas” (74%)

Neste caso, a porcentagem indica que em setenta e quatro porcento dos testes realizados no conjunto de dados, mulheres acima de quarenta anos assistem novelas.

3.4.3 Apriori

O algoritmo *Apriori*, desenvolvido por Agrawal & Seikant (1994), percorre o banco de dados para identificar quais registros são freqüentes e, desta forma, é utilizado para encontrar regras de associação em um conjunto de dados.

Segundo Domingues (2004, p. 25), inicialmente, o algoritmo faz uma varredura no banco de dados, contando a ocorrência de conjuntos de itens freqüentes (*itemsets*). Em seguida, este conjunto de itens freqüentes é utilizado, através da função *apriori-gen*, descrita por Agrawal & Seikant (1994, p. 6), para gerar o conjunto de itens candidatos "C_k".

O banco de dados é então percorrido para se determinar o valor de suporte dos itens candidatos. Este suporte (porcentagem de ocorrência) tem a finalidade de eliminar padrões fracos e irrelevantes, para selecionar somente as regras consistentes. Os itens candidatos que tenham uma freqüência superior a um suporte mínimo estipulado "L_k" são selecionados para em seguida se identificar os elementos que dentre eles são mais freqüentes.

Por último, é feita a união dos conjuntos "L_k" de *itemsets* mais freqüentes. Considera-se somente os conjuntos cujo grau de confiança atende uma porcentagem mínima pré-estabelecida e então descartam-se os demais. A figura 5 demonstra, passo a passo, os procedimentos executados pelo algoritmo *Apriori*.

FIGURA 5 – AS ETAPAS DO ALGORITMO *APRIORI*

```
1: L_1 := {1-itemsets freqüente};
2: for (k := 2; L_{k-1} ≠ ∅; k++) do
3:    C_k := apriori-gen(L_{k-1}); //Gera novos conjuntos candidatos
4:    for all (transações t ∈ T) do
5:       C_t := subset(C_k,t); //Conjuntos candidatos contidos em t
6:       for all candidatos c ∈ C_t do
7:          c.count++;
8:       end for
9:    end for
10:   L_k := {c ∈ C_k | c.count ≥ sup-min};
11: end for
12: Resposta := ⋃_k L_k;
```

FONTE: ADAPTADO DE AGRAWAL & SEIKANT (1994, p. 5)

3.4.4 Algoritmos genéticos

A teoria de evolução das espécies de Darwin (1859), resumidamente, coloca que a evolução genética ocorre através da seleção natural. As espécies se adaptam

às alterações do meio-ambiente por meio de mutações genéticas para garantirem a sua sobrevivência. Os indivíduos que melhor se adaptam a estas mudanças do ambiente são justamente aqueles que sobrevivem e evoluem através das transformações de seus genes.

Os algoritmos genéticos são baseados em princípios de mutação genética e conceitos de evolução biológica, tais como genes, cromossomos, cruzamentos e seleção natural. De acordo com Goldberg (1989, p. 1), estes algoritmos combinam a seleção natural em conjuntos de dados com trocas de informações estruturadas e aleatórias, para formar algoritmos de busca que contêm o instinto de investigação humana. São gerados a partir de diversas regras de classificação que competem entre si, através de diferentes tipos de operadores, efetuando predições de dados. As regras que não apresentam desempenho satisfatório são descartadas, ao passo que as mais eficientes proliferam, produzindo assim variações de si mesmas. De acordo com Santos (2001, p. 127), estas regras são representadas através de um conjunto finito de caracteres, normalmente restritos ao código binário (0,1).

Um algoritmo genético simples, capaz de alcançar bons resultados em inúmeros problemas práticos é composto de pelo menos três operadores: reprodução, cruzamento e mutação. A reprodução é o processo no qual as regras são copiadas de acordo com sua função de avaliação (*fitness*), isto é, sua capacidade de solucionar o problema motivador da aplicação do algoritmo. Quanto melhor for a função de avaliação, maior será a probabilidade desta regra ser selecionada para a próxima geração.

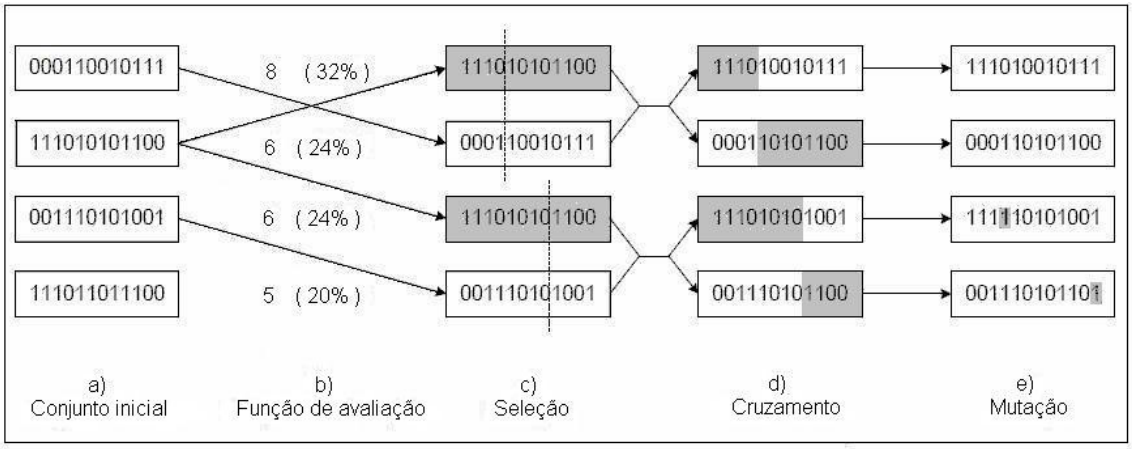
O cruzamento consiste na formação de pares de regras aleatórios que são gerados a partir da mistura das regras anteriormente reproduzidas. Por exemplo, uma regra denominada “c” é gerada a partir do cruzamento da regra “a” com a regra “b”. Este cruzamento é determinado através da união de parte da regra “a” com parte da “b”, para formar uma regra “c” única, resultante do cruzamento das duas primeiras. As novas regras geradas na etapa de cruzamento passarão por todos os processos já mencionados, fazendo com que a cada nova geração sejam encontradas regras mais eficientes para a solução do problema.

Segundo Tsunoda (2004, p. 36) são três os tipos mais comuns de cruzamento. O primeiro é o cruzamento de um ponto, no qual é escolhido alheatoriamente um ponto de corte e então os descendentes recebem parte da carga genética de cada um dos pais. O segundo tipo é o cruzamento de dois pontos,

no qual um dos descendentes fica com a parte central de um dos pais e com as partes externas do outro antecedente, e vice-versa. O terceiro tipo é o cruzamento uniforme no qual cada descendente é preenchido a partir de uma probabilidade que independe da posição que determina o valor de preenchimento de cada um dos antecedentes.

A última etapa do processo é a mutação. Segundo Goldberg (1989, p. 14), este operador atua com pequena probabilidade, cerca de uma mutação para cada mil transferências de bits, e consiste na alteração aleatória do valor binário do caractere de uma regra de 0 para 1 ou de 1 para 0. A figura 6 ilustra de forma resumida as etapas do algoritmo genético.

FIGURA 6 – ETAPAS DO ALGORITMO GENÉTICO



FONTE: (SANTOS, 2001, p. 127)

Desta forma, percebe-se que o algoritmo genético representa uma das técnicas de inteligência artificial mais avançadas, pois simula a evolução natural das espécies de seres vivos, obtendo a cada nova etapa do processo melhores regras e, consequentemente, melhores soluções para o problema.

Os algoritmos genéticos são normalmente utilizados em problemas de otimização complexos, que envolvem muitas variáveis e também muitas possíveis soluções. Têm sido cada vez mais aplicados às áreas de ciências e engenharias.

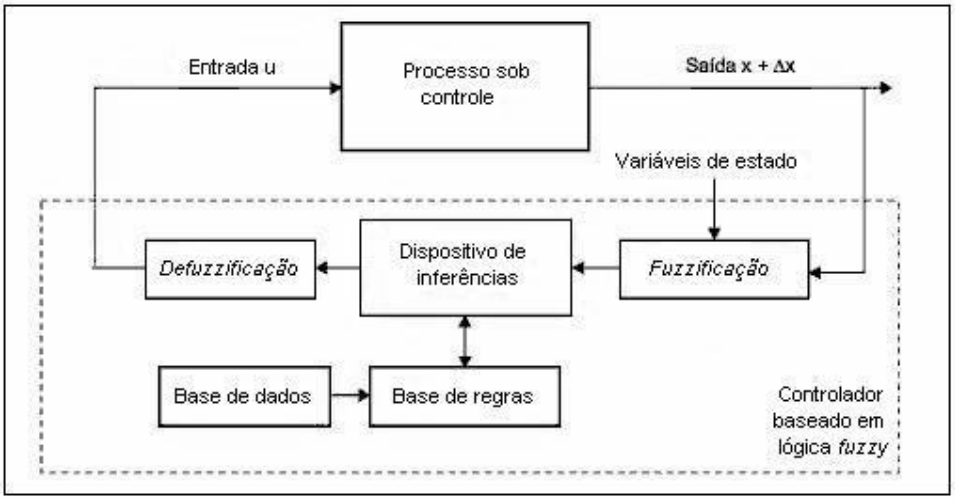
3.4.5 Lógica *Fuzzy*

Os conceitos fundamentais da lógica *fuzzy* foram introduzidos por Lotfi Asker Zadeh em 1965. Desde então ela tem sido muito aplicada nas áreas de controles industriais e sistemas computacionais. De acordo com Ribacionka (1999, p. 6), o termo lógica *fuzzy* costuma ser utilizado de duas formas:

a) com o objetivo de generalizar a lógica clássica aristotélica, na qual as proposições são valoradas apenas como falsas ou verdadeiras, com o intuito de quantificar, formalizar e raciocinar por meio de conceitos imprecisos e subjetivos;
b) como generalização de tudo que envolve o conjunto *fuzzy*.

Junges (2006, p. 4) descreve que, tradicionalmente, uma proposição lógica tem dois extremos: ou ela é completamente verdadeira, ou então completamente falsa. A lógica *fuzzy* obedece estes princípios, contudo, suas premissas variam em grau de verdade de zero (0) a um (1). Ou seja, incluem-se valores adicionais entre os extremos da lógica tradicional e os valores verdadeiro e falso. O controle realizado pela lógica *fuzzy* assemelha-se a um comportamento baseado em regras, ao invés de um controle explicitamente restrito a modelos matemáticos. O seu objetivo é gerar saídas lógicas a partir de um conjunto de entradas imprecisas, com ruídos e até mesmo faltantes. A figura 7 mostra um esquema de como funciona um controlador *fuzzy*:

FIGURA 7 – CONTROLADOR BASEADO EM LÓGICA *FUZZY*



FONTE: SELLITO (2002)

O raciocínio *fuzzy* pode ser dividido em três etapas principais:

a) *fuzzificação*: definição e transformação das variáveis do problema em valores de entrada limitados entre “0” e “1”. Para cada valor de entrada deve ser definida também uma função de pertinência que permite mensurar o grau de verdade das regras (proposições);
b) inferência: nesta etapa definem-se as regras (condicionais e não-condicionais) e analisa-se cada uma delas paralelamente, identificando sua importância para a resolução do problema e sua influência nas variáveis de saída;
c) *defuzzificação*: conversão das variáveis *fuzzy* em valores numéricos aceitos pelo sistema. Neste estágio várias técnicas distintas podem ser utilizadas, tais como: centróide, critério *maxima*, *middle-of-maxima* e *first-of-maxima*.

Portanto, percebe-se que o algoritmo baseado na lógica *fuzzy*, dentre as possíveis aplicações, é adequado para o controle de sistemas continuamente variáveis e para problemas de natureza industrial, biológica e química, que compreendem situações ambíguas.

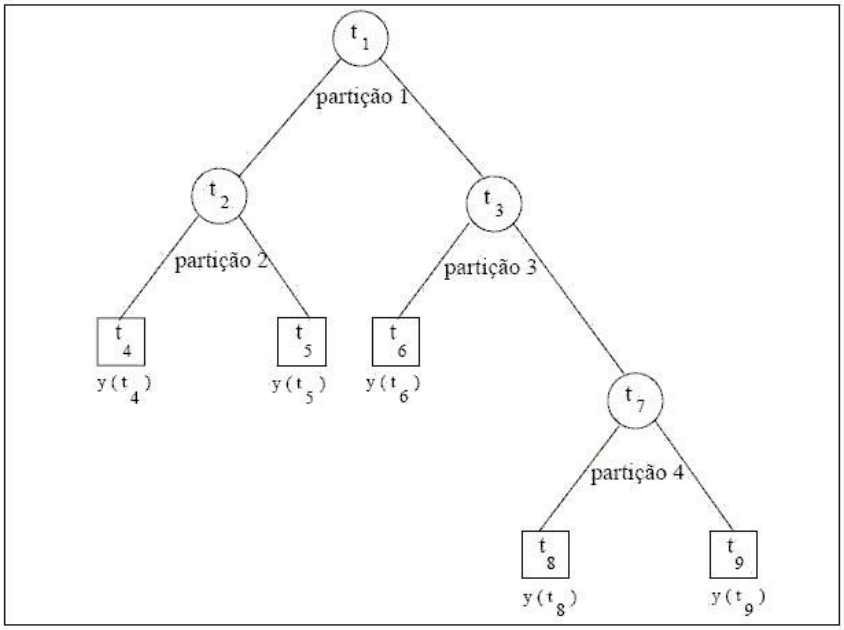
3.4.6 Árvores de regressão

Árvores de decisão aplicadas a problemas de regressão são denominadas árvores de regressão. Cada nó terminal ou folha, nas árvores de regressão, contém uma constante ou uma equação para o valor previsto de um determinado conjunto de dados.

De acordo com Andrade (2003, p. 3), a finalidade dos modelos de regressão é explicar uma ou várias variáveis que se têm como objeto de estudo, em função de outras variáveis que assumem caráter explicativo. Por exemplo, explicar o preço de um imóvel em função de sua localização.

Dentro do escopo da MD, as árvores de regressão correspondem a um método de aprendizagem supervisionado. Este método pode ser representado através de uma árvore binária, com estrutura muito semelhante às árvores de decisão, conforme ilustrado na figura 8.

FIGURA 8 – A ESTRUTURA DAS ÁRVORES DE REGRESSÃO



FONTE: RODRIGUES (2003, P. 29)

3.4.7 KNN

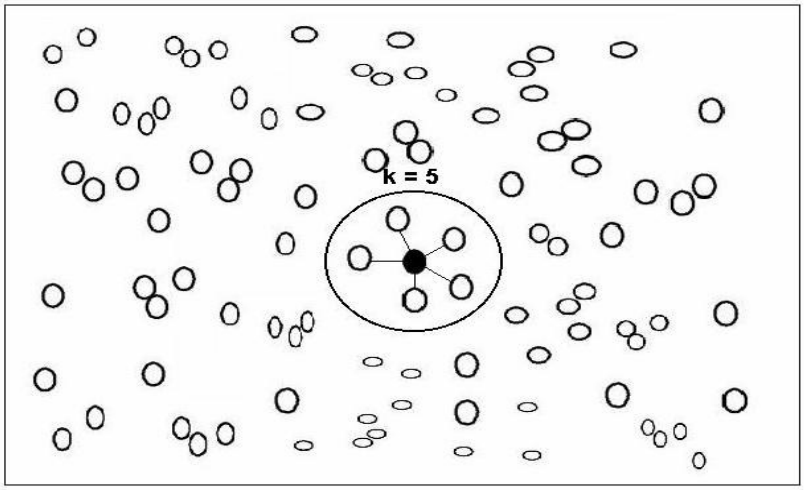
O algoritmo KNN (*k-nearest neighbor*), baseado no vizinho mais próximo, calcula as distâncias entre as amostras do conjunto de dados para gerar grupos de dados. Para isto, calcula uma matriz de distância que compara um determinado elemento com todas as amostras do conjunto de dados, fazendo uma ordenação da variável mais próxima à mais distante e estabelecendo agrupamentos (classes) com base nas distâncias mais próximas encontradas.

Carvalho (2007, p. 2) coloca que o algoritmo KNN classifica um determinado item de dado de acordo com as respectivas classes dos “k” (k ≥ 1) vizinhos mais próximos presentes no conjunto de dados. É medida a distância deste item de dado em relação a cada elemento do conjunto. Então, os dados são ordenados do mais próximo ao mais distante do elemento analisado.

Destes dados ordenados, são selecionados somente os “k” mais próximos, que servem de parâmetro para a regra de classificação. Exemplo: 1-NN é um k-NN sabendo que k é igual a 1, ou seja, é selecionado somente o elemento mais próximo do item de dado que se pretende classificar.

Caso o valor de k seja igual a 5-NN significa que serão selecionados os cinco elementos mais próximos do item analisado. Com base na classe dos cinco elementos analisados é determinada a classe do item de dado que se pretende classificar, conforme a figura 9.

FIGURA 9 – MÉTODO DO VIZINHO MAIS PRÓXIMO (KNN)



FONTE: O AUTOR

3.4.8 Redes neurais

As redes neurais correspondem a modelos simplificados do sistema nervoso central do ser humano. De acordo com Zuben (2003, p. 1) são sistemas de processamento de informação formados pela interconexão de unidades simples de processamento, denominadas neurônios artificiais. Uma grande rede neural pode ter centenas ou milhares de unidades de processamento.

Segundo Alecrim (2004, p. 1) as redes neurais assemelham-se ao cérebro humano em dois aspectos: o conhecimento é armazenado por meio de pesos sinápticos (valores atribuídos nas conexões entre os neurônios) e é adquirido através de etapas de aprendizagem. Esta aprendizagem é feita por meio de um conjunto de regras denominadas de algoritmo de aprendizado. O objetivo destas regras é realizar o processamento de informações tendo como inspiração original a estrutura de neurônios do cérebro humano, capaz de aprender e tomar decisões baseadas em aprendizagem.

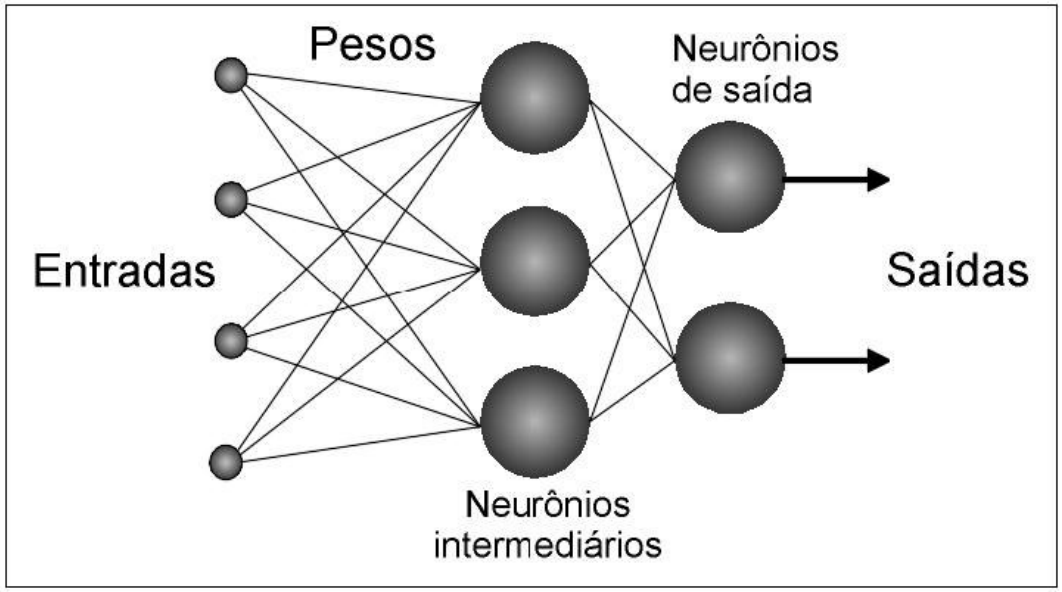
Existem várias maneiras de se projetar uma rede neural. Ela deve ser desenvolvida de acordo com o problema a ser resolvido e em sua arquitetura, dentre os principais aspectos, são determinados o número de camadas de neurônios, a quantidade de neurônios em cada camada e o tipo de sinapse utilizado. As redes neurais são criadas a partir de algoritmos de aprendizado projetados com uma finalidade específica, de acordo com o problema a ser resolvido. Estes algoritmos são desenvolvidos através de modelos matemáticos que simulam o processo de aprendizado do cérebro humano.

Segundo Carvalho (2005, p. 98) a única tarefa que uma rede neural pode efetuar é a associação de padrões. Seu funcionamento ocorre da seguinte forma: é fornecido um padrão de atividade denominado entrada (estímulo) e a rede neural desenvolve então um padrão de saída (resposta). Entre o padrão de entrada e o padrão de saída ocorre propagação de sinal através da rede neural e é realizada a associação de padrões. Esta associação pode ser de dois tipos:

a) auto-associação: quando o padrão de entrada é associado a si mesmo na saída. Importante nas tarefas de reconhecimento de padrões;

b) hetero-associação: quando o padrão de entrada é associado a um padrão de saída diferente. Importante nas aplicações de classificações de padrões, análise de agrupamentos, diagnóstico, previsão, entre outras.

FIGURA 10 – ESTRUTURA TÍPICA DE UMA REDE NEURAL



FONTE: TAFNER (1998, p. 3)

O exemplo mais antigo de redes neurais é a rede *perceptron* com uma camada, proposta por Rosenblatt (1962). Este modelo é formado por uma camada única de neurônios de saída, os quais estão conectados por pesos às entradas. Posteriormente foram desenvolvidas redes *perceptron* multi-camadas, nas quais diversas camadas de unidades computacionais são interconectadas, possibilitando que um neurônio em uma camada específica tenha conexões diretas a neurônios da próxima camada.

O processo de aprendizagem das redes neurais é realizado através das modificações que ocorrem nas sinapses dos neurônios. Estas alterações são efetuadas de acordo com a ativação dos neurônios. Neste contexto, as conexões mais usadas são reforçadas enquanto que as demais são enfraquecidas. Existem basicamente três paradigmas de aprendizado:

a) aprendizado supervisionado: um agente externo indica à rede a resposta desejada para o padrão de entrada;

b) aprendizado não-supervisionado: não existe um agente externo indicando a resposta desejada para os padrões de entrada. A rede processa os dados para determinar propriedades do conjunto de dados;
c) reforço: um crítico externo avalia a resposta fornecida pela rede.

Desta forma, percebe-se que os algoritmos de rede neural, assim como os algoritmos genéticos, podem ser utilizados para a resolução de inúmeros tipos de problema, nas mais diversas áreas. *Scanners* de reconhecimento de caracteres em documentos textuais e programas *anti-spam* são apenas alguns dos exemplos mais simples de suas aplicações. As redes neurais têm sido muito utilizadas atualmente em contextos mais complexos, como em usinas e mercado financeiro.

4 ANÁLISE DE ESTUDOS DE CASO

4.1 BIBLIOTECA CENTRAL DA FURB

A Biblioteca Central da Fundação Universidade Regional de Blumenal (FURB) não apresentava aos seus usuários nenhum Sistema Informatizado de Recuperação (SRI) e Disseminação Seletiva de Informações (DSI).

Porém, segundo Jesus (2004), com o intuito de se conhecer o perfil de cada usuário, delineando suas preferências e interesses, foram aplicadas técnicas de MD na biblioteca para possibilitar a personalização dos processos de SRI e DSI, tornando-os claros e objetivos.

Os principais objetivos do mencionado projeto eram:

a) aplicar técnicas de MD sobre as transações de empréstimos e reservas de obras dos usuários;
b) desenvolver uma personalização dos sistemas de DSI e SRI, de acordo com o perfil de cada usuário.

Neste trabalho, será feita uma descrição deste projeto de MD no que diz respeito à metodologia adotada e resultados encontrados, bem como às conclusões extraídas de tal experimento.

4.1.1 Identificação do problema

Conforme já mencionado, a Biblioteca Central da FURB não possuía sistema automatizado de recuperação e disseminação de informações. Este processo era feito de forma manual por bibliotecários e especialistas da área. Quando os usuários faziam consultas em busca de obras para empréstimo e reserva, a pesquisa retornava grande quantidade de informações, sendo que a grande maioria dos resultados não apresentava relevância nem ordenação, o que gerava uma baixa precisão na recuperação de informações por parte dos usuários.

4.1.2 Metodologia adotada

Primeiramente, foram identificadas as variáveis diretamente relacionadas ao problema: usuários, obras da biblioteca e a classificação decimal de Dewey (codificação universal que classifica obras de acordo com áreas do conhecimento, muito utilizada em bibliotecas).

Após definidas as variáveis, foi feita a coleta, seleção e pré-processamento dos dados. A coleta dos dados foi feita a partir dos sistemas legados da biblioteca e o sistema de identificação única de pessoas com vínculo na instituição. Foram selecionados professores e alunos de pós-graduação da FURB na etapa de seleção. Os atributos que não eram de interesse para a pesquisa como CPF e endereço dos usuários, por exemplo, foram excluídos do conjunto de dados selecionado. Já na etapa de pré-processamento, foi feita a verificação das inconsistências e erros nas variáveis. A data de aquisição continha dados em diferentes padrões e o código decimal de Dewey (CDD) em alguns casos estava fora do padrão de catalogação. Desta forma foi necessário manter os dados dos atributos dentro de um formato homogêneo, a fim de prepará-los para a mineração de dados propriamente dita.

A amostra preparada para a MD ficou da seguinte forma: 17.421 títulos que totalizaram 51.011 volumes, 3.906 usuários, 821 da categoria professores e 3.085 da categoria de pós-graduação. Foram realizadas 68.543 transações, sendo 66.769 de empréstimo e 1.775 de reservas. A média foi de 17,54 transações por usuário.

Finalmente, na etapa de mineração de dados propriamente dita, foi escolhida a tarefa de agrupamento para que, com base nas transações realizadas pelos usuários, as obras pudessem ser agrupadas em áreas do conhecimento a partir da classificação decimal já mencionada. Para tanto, optou-se por um algoritmo de agrupamento hierárquico aplicado às transações dos usuários e, assim, foi possível identificar as áreas de conhecimento de interesse de cada usuário, tornando possível personalizar seu perfil de utilização da biblioteca.

4.1.3 Resultados encontrados

Conforme planejado, a partir da mineração aplicada às transações dos usuários da biblioteca foram identificados vários grupos e subgrupos de interesse

relacionados às obras emprestadas e reservadas pelos usuários da Biblioteca Central da FURB. Como exemplo dos resultados obtidos, o quadro 1 detalha o agrupamento hierárquico da área do Direito.

QUADRO 1 – ÁREAS DE INTERESSE OBTIDAS ATRAVÉS DA MD

Grupo	Descrição do Grupo	CDD	Subgrupo de interesse
1	Direito	341.5	Direito penal
		342.1	Direito civil
		341.2	Direito constitucional
		342.2	Direito comercial
		341.6	Direito do trabalho
		340.1	Filosofia do direito
		340	Direito

FONTE: O AUTOR

No estudo do perfil dos usuários, o grupo 1, por exemplo, representa a grande área de interesse e os subníveis hierárquicos representam as subáreas correspondentes aos interesses dos usuários, com base nas transações analisadas.

A partir destes resultados foi desenvolvido um sistema *web* que oferece os serviços de recuperação e disseminação seletiva de informações personalizadas dinamicamente ao perfil de cada usuário.

Quando o usuário submete a requisição de empréstimo ou reserva, o sistema a processa e retorna páginas html com o conteúdo personalizado ao usuário. O sistema possui ainda um banco de dados que contém informações sobre as obras, usuários, transações e todo o conjunto de informações que serviram de fonte para a aplicação da MD.

4.1.4 Análise dos resultados

Embora não tenha sido feita uma avaliação quantitativa dos resultados encontrados, pode-se concluir que os objetivos do trabalho foram alcançados já que o sistema *web* desenvolvido mostrou-se funcional em seu propósito.

Além disso, pelo caráter inovador do projeto, este pode ser considerado um modelo para a implantação de SRI e DSI em bibliotecas, baseado na aplicação da

MD sobre a classificação CDD das obras (cujo padrão é utilizado em inúmeras bibliotecas), permitindo que os interesses bibliográficos específicos de cada usuário sejam identificados.

Desta forma, percebe-se que a mineração de dados realizada neste experimento gerou grupos com correlações entre livros implícitas, possibilitando que os hábitos e interesses dos usuários passassem a ser melhor aproveitados no aperfeiçoamento dos serviços oferecidos pela biblioteca.

4.2 PROVEDOR DE *INTERNET*

A utilização da *internet* por empresas e instituições se torna a cada dia mais comum no mundo globalizado da era da informação. E, para as empresas conseguirem se manter no mercado de forma competitiva, é fundamental que mantenham estratégias de aproveitamento dos dados disponíveis na *web*.

De acordo com Boscarioli (2003), professores e analistas de sistemas da Universidade Estadual do Oeste do Paraná (UNIOESTE) fizeram um experimento de mineração de dados na *web*. Desta forma, técnicas de MD foram aplicadas com o intuito de descobrir padrões de navegação de usuários freqüentadores do *web site* de uma empresa provedora de acesso à *internet*. O conhecimento do perfil dos usuários do *site* possibilitaria a personalização dos serviços, produtos e atendimento aos clientes.

Com base no experimento acima mencionado, este trabalho irá descrever como a MD pode servir de suporte ao processo decisório através de informações coletadas no ambiente *web*.

4.2.1 Identificação do problema

Com o crescimento de informações disponíveis na *internet* e com o aumento do número de usuários que utilizam este meio para solicitar serviços e produtos, é fundamental que as empresas procurem aproveitar os dados obtidos na *web* no suporte ao processo decisório.

A falta de informações corretamente administradas faz com que oportunidades de negócios sejam perdidas e que o atendimento aos clientes seja impreciso, sendo ofertados serviços e produtos que não satisfazem as reais necessidades e preferências dos clientes.

Portanto, através deste experimento, pretendeu-se obter importantes informações sobre o perfil dos usuários que navegam no *web site* da empresa, tais como serviços mais procurados, páginas mais visitadas e quais os horários de acesso em que os usuários mais freqüentam o *web site* da empresa.

4.2.2 Metodologia adotada

Inicialmente, foi necessário utilizar um mecanismo de *cookie* para gerar um identificador (ID) para cada usuário que acessasse o *site* e assim tornou-se possível identificar cada usuário de maneira única nos *logs* do servidor, pois o valor armazenado no *cookie* ficaria visível também nos arquivos de *log*.

Os dados selecionados para mineração foram coletados no período de uma semana e os atributos escolhidos para análise foram: período (horário de acesso), tipo de internauta (cliente da empresa ou visitante) e a página do *web site* (acessada pelo usuário). Houve também necessidade de configurar o armazenamento de uma variável chamada *referer* que indica de que endereço o usuário veio (caso haja) para acessar o *site* da empresa. Essa informação é útil para indicar a eficiência dos *banners* de propaganda da empresa exibidos em outros *sites*.

Foram utilizados os *logs* de acesso semanal para identificar o perfil dos usuários. O intervalo de tempo foi pequeno devido à grande quantidade de arquivos de *log* armazenados. Caso fosse determinado um intervalo de tempo maior, o experimento demandaria muito mais tempo, e, além disso, os recursos de *hardware* não seriam adequados para o grande volume de dados que precisaria ser processado.

Uma longa etapa de preparação dos dados foi necessária para que este experimento se tornasse possível. Primeiramente, os arquivos de *log* foram convertidos do formato de texto para tabelas do *Microsoft SQL Server*. Este procedimento gerou sete tabelas, cada uma relacionada a um dia da semana.

Depois disso, afim de eliminar inconsistências e registros não relacionados aos objetivos do experimento (como arquivos .gif, .jpg, .swf), foi feita a limpeza e tratamento dos dados por meio de *scripts* implementados em ASP.

Em relação ao atributo “página”, devido ao *web site* conter um grande número de páginas, optou-se por escolher os serviços e seções considerados os mais importantes para a empresa. As informações relacionadas aos cliques dos usuários foram obtidas através de tabelas geradas a partir dos *logs*, também por meio de *scripts* em ASP. Para descobrir se o internauta era cliente ou não, foi feita uma análise do seu número de IP registrado nos *logs*.

Os dados foram então convertidos do formato SQL *Server* para um arquivo no formato *Excel* e em seguida para o formato CSV, que separa os atributos por vírgulas. Finalmente, o arquivo CSV foi renomeado para a extensão ARFF (padrão da ferramenta utilizada) para que pudesse ser processado na etapa de MD propriamente dita.

Na realização da mineração de dados, foi utilizada a ferramenta *Weka*, desenvolvida pela Universidade de Waikato, Nova Zelândia. Esta ferramenta está disponível em código aberto (*open source*) e possui uma série de algoritmos, implementados em Java, desenvolvidos para a solução de problemas de mineração de dados. A tarefa escolhida para este experimento foi a de regras de associação e o algoritmo utilizado foi o *Apriori*.

4.2.3 Resultados encontrados

Considerando-se regras com um mínimo de 50% de confiança, foi possível extrair as seguintes informações na etapa de mineração de dados:

a) os clientes acessam o *web site* da empresa à noite, com fator de confiança de 64%;
b) as regras indicam que os usuários costumam acessar a página de busca do site com mais freqüência nos períodos da tarde (58% de confiança) e noite (62% de confiança);

c) quando os clientes utilizam o serviço de *Webmail* estão na maioria das vezes conectados à *internet* pelo próprio provedor de acesso, com fator de confiança de 57%;
d) várias regras de uma forma geral indicam que nos períodos da manhã e tarde, os acessos mais registrados foram de visitantes.

Desta forma, mesmo em um intervalo curto de tempo de análise de dados, foi possível obter informações relevantes acerca dos acessos ao *site* do provedor.

4.2.4 Análise dos resultados

Através deste experimento comprovou-se que uma ferramenta de mineração de dados pode ser eficiente na descoberta de conhecimento e informações relevantes a partir de ambientes *web* corporativos, uma vez que o conjunto de dados minerado foi obtido por meio de arquivos de *logs* de acessos ao servidor *web* que hospeda o *site* de uma empresa provedora de acesso à *internet*.

Para que tendências mais significativas dos usuários pudessem ser descobertas haveria necessidade de se considerar um intervalo de tempo de análise maior. Assim, as chances de se obter informações de relevância estratégica também aumentariam.

De qualquer forma, embora este experimento tenha levado em consideração somente os *logs* de acesso semanal ao *site* do provedor, ainda assim a gerência da empresa pôde ter uma idéia do perfil dos usuários que acessam o *site* em diferentes períodos do dia, bem como quais os serviços mais procurados. Com base nisso surgiu a possibilidade de ofertar serviços e atendimento personalizado aos usuários.

4.3 UNIVERSIDADE TUIUTI

A Universidade Tuiuti do Paraná (UTP), há alguns anos, oferece cursos de graduação, doutorado, mestrado e extensão. Com o intuito de identificar o índice de aproveitamento dos alunos e compara-lo com os objetivos previamente definidos

pelos coordenadores, dois experimentos com mineração de dados foram realizados em parceria com o Instituto Paranaense de Desenvolvimento Econômico e Social (IPARDES): um no ano de 2003 e outro no ano de 2004. Estes experimentos visavam auxiliar o processo decisório de coordenadores e colegiados dos cursos de graduação. A partir dos resultados obtidos em 2003 a instituição tomou algumas decisões que serviram de parâmetro para a mineração realizada em 2004. Segundo Carvalho (2004), foi possível verificar se as decisões tomadas no primeiro experimento surtiram efeito, bem como, obter novos resultados de suporte à tomada de decisão.

O foco deste trabalho é relatar o problema bem como o processo metodológico realizado no segundo experimento, procurando dar ênfase para os resultados obtidos por meio deste.

4.3.1 Identificação do problema

A universidade até então não utilizava nenhum recurso computacional que permitisse não somente coletar informações de um banco de dados mas também trata-las e processa-las com a finalidade de descoberta de conhecimento institucional. Após o experimento com MD realizado em 2003, tornou-se necessário desenvolver um mecanismo computacional que permitisse mensurar o desempenho dos alunos nas disciplinas do curso de Ciência da Computação com a finalidade de se avaliar na prática a relevância das ações tomadas com base no experimento de 2003. Outra necessidade era encontrar possíveis relações entre disciplinas e índices de reprovação.

4.3.2 Metodologia adotada

Os critérios adotados neste experimento foram semelhantes aos utilizados em 2003, com exceção da delimitação do escopo. No primeiro experimento haviam sido considerados todos os cursos de ciência da computação da universidade. No

segundo, porém, foi priorizado o curso de graduação de Ciência da Computação pois, especificamente para este curso, haviam sido tomadas algumas medidas baseadas no primeiro experimento de 2003. Portanto, um dos objetivos do experimento de 2004 é a verificação da eficiência das ações tomadas a partir dos resultados obtidos no primeiro experimento.

O conjunto de dados utilizado em 2004 compreende atributos referentes à performance dos alunos como notas, faltas e aprovações em disciplinas. A tarefa adotada foi de associação e o algoritmo foi o Apriori. As razões por esta escolha dizem respeito aos objetivos definidos para a aplicação da técnica de mineração: identificar disciplinas com reprovações associadas e situações de exceções nestas reprovações.

No atributo relacionado à situação de aprovação nas disciplinas foram necessárias algumas transformações dos dados para adequá-los ao processo de MD. Como não havia o valor “desistência” no atributo foi preciso criar uma regra para analisar os dados: todos os alunos com nota do primeiro bimestre diferente de zero e dos demais bimestres igual a zero foram classificados como desistentes.

Além disso, as disciplinas do curso também foram agrupadas de acordo com as seguintes áreas do conhecimento: básicas, tecnológicas, complementares e humanísticas. Isso permitiu ampliar o experimento, tratando o problema não somente por meio das disciplinas, mas também através das áreas do conhecimento atribuídas a cada uma delas.

4.3.3 Resultados encontrados

De acordo com a mineração de dados descrita foi verificado que no ano de 2003 (após as ações baseadas no primeiro experimento) ocorreu o período com melhor resultado de aprovação desde 1999: 74% dos alunos foram aprovados neste ano. Em 1999, 2000 e 2001, esse número havia sido de 64%, 65% e 60% respectivamente, o que revela um aumento considerável em 2003.

Em relação a desistências, ocorreu considerável redução em comparação com os anos anteriores. Esta redução sofreu uma queda de cerca de 60%, o que acarretou em um maior índice de aprovação. É interessante mencionar que em uma

das disciplinas a redução de desistentes chegou a 69%, resultando em um positivo aumento no número de aprovados.

Em se tratando dos conjuntos de regras encontrados, o número foi extremamente grande (variou de centenas a milhares de regras), o que acabou por dificultar a análise de relevância de cada uma delas. Neste caso, tornou-se necessário identificar situações de exceção para avaliar quais eram as regras mais consistentes como suporte ao processo decisório.

O experimento também identificou duas disciplinas do curso que necessitavam de reformulações, o que resultou em posteriores ações de melhoria por parte da equipe de colegiado.

4.3.4 Análise dos resultados

Através deste experimento com a técnica de MD muitos resultados positivos foram obtidos. Primeiramente, foi verificado que as medidas adotadas em 2003, especificamente para o curso de Ciência da Computação, contribuíram para o aumento do aprendizado dos alunos e também para uma queda significativa no número de desistências.

É importante mencionar que o segundo experimento, embora tenha utilizado a mesma metodologia do primeiro, acabou aperfeiçoando-a com a inclusão de novos atributos gerados a partir da transformação de dados.

Além deste segundo experimento comprovar a eficiência prática de ações anteriormente executadas, com base em um processo de MD, também serviu de apoio para a execução de reformulações metodológicas em algumas disciplinas.

4.4 EMPRESA DISTRIBUIDORA DE ENERGIA ELÉTRICA

Gerir sistemas de distribuição de energia elétrica é algo complexo e caracteriza um constante desafio para as empresas do ramo, pois há fatores elétricos, físicos e humanos envolvidos na manutenção e operação das redes que compõem esses sistemas.

O experimento relatado neste trabalho, Anciutti (2004), descreve a utilização de técnicas de MD aplicadas sobre um conjunto de dados sobre circuitos elétricos de baixa tensão, extraídos de um sistema de *software* corporativo de uma empresa de distribuição de energia elétrica.

4.4.1 Identificação do problema

As redes de distribuição de energia elétrica normalmente enfrentam situações de imprevisibilidade em virtude de mudanças metereológicas, variações na demanda de potência relacionadas à sazonalidade do consumo, falhas de equipamentos e até mesmo a utilização clandestina de eletricidade.

Estas situações costumam gerar altos custos às companhias elétricas e por isso, elas têm investido em novas tecnologias que possam otimizar as redes de distribuição, melhorando assim a qualidade dos serviços prestados e reduzindo os gastos com multas às agências controladoras e com manutenção e compra de equipamentos.

Portanto, o experimento aqui relatado descreve a utilização de algoritmos genéticos para encontrar regras de associação em um conjunto de dados sobre circuitos de baixa tensão, de uma empresa de distribuição de energia elétrica.

4.4.2 Metodologia adotada

Primeiramente foi necessário definir quais características dos circuitos de baixa tensão eram relevantes para o experimento. Para isso um engenheiro eletricista especialista em redes de distribuição de energia prestou suporte e ajudou a delimitar os dados a serem selecionados.

O conjunto de dados adotado tinha um total de 5210 registros, sendo cada registro correspondente a um circuito elétrico, resultando em 120 atributos. Porém, dentre estes campos somente 30 foram selecionados para análise. No pré-

processamento dos dados apenas 3072 indivíduos (circuitos elétricos) foram considerados adequados a partir da amostra original com 5210 registros.

Alguns atributos não estavam consolidados de forma que pudesse ser feita uma análise direta, então foram elaborados critérios através do cálculo de uma média matemática ponderada, denomidada centróide. De acordo com os valores obtidos para as centróides foram estabelecidas faixas de valores para realizar a classificação dos circuitos.

Além disso, devido ao fato de muitos valores analisados serem contínuos, antes que os dados pudessem ser minerados foi necessário discretizá-los, isto é, foi preciso substituir os valores numéricos por faixas nominais correspondentes aos seus intervalos. Isto se deu por meio de cálculos probabilísticos seguindo a fórmula de Sturges (1995).

Desta forma, foram encontradas 12 faixas, sendo posteriormente necessário separar cada valor de determinado atributo de acordo com a classe a que pertencia. Para atributos já discretizados (formato nominal), em um número de classes menor do que aquele encontrado pela fórmula de Sturges, manteve-se a classificação original.

Para selecionar os atributos que comporiam o cromossomo, isto é, que mais teriam influência em uma regra cujo conseqüente (classificação) é conhecido, foi realizada uma análise estatística descritiva, verificando-se a correlação de todos os atributos (que indicam a condição) com o atributo conseqüente da regra. Deste modo, os campos que atingiram mais de 50% de correlação com as variáveis independentes foram selecionados para compor o cromossomo e os atributos que não eram relevantes para o objetivo da regra foram descartados, reduzindo assim o tamanho do cromossomo.

Conforme mencionado anteriormente, cada indivíduo foi definido como um circuito de baixa tensão e suas características (fator de potência, índice de carregamento, etc) como seus genes. E, embora o alfabeto binário seja o mais indicado para aplicação de algoritmos genéticos e para a representação cromossômica, neste experimento esta estrutura não seria adequada, pois o universo de soluções seria limitado a somente duas classes, gerando resultados não tão significativos quanto aqueles obtidos por meio de mais variáveis.

Quanto ao tamanho da população, cerca de 30% dos registros que compunham o conjunto total de dados foram selecionados para serem processados

pelo algoritmo, o que representou 1000 registros (extraídos dos 3072 anteriormente selecionados).

O algoritmo genético aplicado ao conjunto de dados foi uma adaptação daquele proposto por Goldberg (1989), com modificações específicas para as necessidades deste experimento. Primeiramente o algoritmo busca de forma aleatória no conjunto de dados a população de indivíduos com tamanho pré-definido pelo usuário, 1000 registros no caso deste experimento. Então, é feita a análise da população calculando-se a sua função de avaliação (*fitness*), com base nos parâmetros relacionados ao escopo do problema. Em seguida, é feito o processo repetitivo de otimização da população produzindo-se novas gerações de acordo com o valor do *fitness* alcançado por elas. Por fim ocorre a aplicação do cruzamento (*crossover*) e da mutação. Quanto ao valor das probabilidades, escolheu-se os valores de 60% para cruzamento (*crossover*) e 8% para mutação.

Preferiu-se deixar ao usuário e não ao algoritmo a análise quanto à significância da variação obtida na população ou sobre a eficácia dos operadores genéticos. Depois da evolução de seis gerações, a população passou a não alterar o seu material genético, ou seja, neste caso nenhum indivíduo novo surgia e o mesmo cromossomo passava a ser somente replicado.

4.4.3 Resultados encontrados

Da aplicação do algoritmo genético foram obtidos os resultados detalhados no quadro 2, que ilustra os intervalos de valores para os principais atributos de cada regra de associação detectada pelo algoritmo genético.

QUADRO 2 – RESULTADOS DOS ALGORITMOS GENÉTICOS

SITUAÇÃO	CARACTERÍSTICAS ENCONTRADAS	VALOR
Circuitos com tensão adequada (220 a 213,4V)	Potência perdida no circuito;	<= 0,085
	Índice de queda de tensão entre fase A e o neutro	<= 0,675
Circuitos com potência demandada inferior a 14,15 kVA	Quantidade de KWh consumido para fase elétrica A	<= 30,5
	Índice de carregamento do trafo	<= 42,5
Circuitos com índice de carregamento de potência acima de 86,5%	Demanda de potência de consumidores secundários	>= 75,05

	Energia total (KWh) medida para consumidores residenciais	>= 10,5

FONTE: O AUTOR

Desta forma é possível identificar os principais motivos que podem provocar alterações de valores nos indicadores vitais para os sistemas de distribuição de energia, tais como tensão, potência e carga. Informações que podem ser utilizadas tanto em projetos como também na operação de redes de distribuição de energia elétrica, possibilitando a melhoria da qualidade do serviço prestado pela empresa.

4.4.4 Análise dos resultados

Através deste estudo, foi possível verificar que muitas vezes é necessário um longo tempo despendido na parte de pré-processamento e tratamento dos dados, para que se possa obter um conjunto de dados adequado para a mineração.

No caso deste experimento, foi necessária a assistência de um especialista em redes de distribuição de energia elétrica para fosse efetuada a seleção dos dados. Entretanto, mesmo tendo o suporte de um especialista no escopo do problema, em alguns aspectos metodológicos os autores do experimento não detalharam com clareza os critérios adotados na escolha dos atributos e nem a razão pela qual apenas um terço dos registros escolhidos foram de fato processados pelo algoritmo.

Outro fator que é interessante mencionar, é que na apresentação das regras de associação encontradas não foi informada a função de avaliação (*fitness*) de cada uma delas. De qualquer forma, segundo os autores, o algoritmo foi executado com várias combinações de parâmetros para assegurar que as regras de associação encontradas fizeram uso do potencial máximo do algoritmo. Assim, percebe-se que a utilização de algoritmos genéticos na mineração de dados pode ser muito mais explorada na gestão organizacional, sendo eficiente para minerar dados com inúmeras variáveis para solucionar complexos problemas.

5 CONSIDERAÇÕES FINAIS

Através do levantamento de dados realizado para a compilação deste trabalho foi possível confirmar a pouca exploração das técnicas de mineração de dados no Brasil. A princípio a idéia era encontrar casos de utilização da ferramenta em Curitiba-PR, porém as poucas empresas que já haviam feito experimentos nesta área não se dispuseram a abrir espaço para este estudo investigativo, o que de certa forma é compreensível tendo em vista o caráter estratégico dos dados normalmente utilizados na MD.

Então, optou-se por analisar estudos de caso realizados em todo o âmbito nacional que atendessem aos objetivos desta pesquisa. Assim foi possível identificar algumas características comuns presentes nos experimentos de MD analisados.

Percebeu-se que a grande maioria dos bancos de dados organizacionais não é projetada com o intuito de se fazer um processamento estratégico das informações armazenadas. Isto ocasiona um longo período gasto nas tarefas de seleção, pré-processamento, tratamento e demais etapas que antecedem a MD propriamente dita.

Outro fator que merece ser mencionado é que em muitos casos os processos metodológicos, sob o ponto de vista científico, podem ser bastante aperfeiçoados no que diz respeito aos critérios adotados para definição de aspectos tais como: seleção de atributos, escolha de algoritmos e tarefas mais adequados para a aplicação da tecnologia de MD. Quanto mais cuidadosa for a definição destes critérios, maior será a qualidade da MD e, possivelmente, maior será a relevância dos resultados obtidos com o experimento. De acordo com os casos estudados, a MD tem sido muito útil no apoio à tomada de decisão operacional e tática, isto é, de curto e médio prazo respectivamente.

Porém, maior será a qualidade da MD quanto melhor projetados forem os bancos de dados organizacionais ou, em outras palavras, mais adequados à extração de informações estratégicas e descoberta de conhecimento. Bancos de dados consistentes permitem que um grande volume de dados possa ser processado, garantindo uma relevância estratégica para os resultados da aplicação de MD.

De qualquer forma, por mais que esta área ainda esteja em estágio embrionário no Brasil, já tem mostrado ótimos resultados. Nesta última década, muitos alunos de cursos de computação de universidades públicas e privadas vêm realizando inúmeros experimentos com esta tecnologia, tanto no âmbito acadêmico como também no ambiente real de empresas e instituições.

Para o Curso de GI da UFPR, a MD é certamente muito promissora. Primeiramente, porque esta tecnologia é capaz de processar dados brutos fornecendo subsídios cruciais para a gestão de informações organizacionais e para o apoio a processos decisórios. E, em segundo lugar, porque ainda há muito o que explorar na área, tanto do ponto de vista computacional, como também gerencial.

Deste modo, assim como os livros de Direito representam o suporte necessário para um advogado exercer sua profissão, técnicas que permitam ao gestor da informação obter conhecimento organizacional a partir de dados brutos são certamente imprescindíveis. Esta é a proposta funcional da MD.

Pode-se concluir que tanto o objetivo geral, como também os objetivos específicos desta pesquisa foram atingidos. Os alunos e profissionais formados em GI, sem dúvida alguma, terão uma grande oportunidade de colocação no mercado de trabalho ao apresentarem as habilidades e competências necessárias para a manipulação desta tecnologia.

REFERÊNCIAS

ADRIAANS, Pieter e ZABTINGE, Dolf. ***Data Mining***. England: Addison Wesley Longmann, 1996.

ALECRIM, Emerson. **Redes neurais artificiais**. Disponível em: <http://www.infowester.com/redesneurais.php> Acesso em: 19 nov 2007.

ALMEIDA, Leandro Maciel e PADILHA, Thereza Patrícia P. **Um modelo do aprendizado de grupos de alunos em ambientes colaborativos**. Disponível em: <http://www.ulbra-to.br/ensino/43020/artigos/anais2003/anais/modeloaprendizado-encoinfo2003.pdf> Acesso em: Acesso em 24 ago. 2007.

ANCIUTTI, Isabela et. al. **Uma aplicação de data mining sobre circuitos elétricos de baixa tensão utilizando algoritmos genéticos**. Disponível em: <http://www.inf.ufsc.br/~frank/papers/WorkCompSul2004-Isabela.pdf> Acesso em: 10 de out de 2007.

ANDRADE, D. F. et al. **Modelos de regressão**. Disponível em: <http://www.inf.ufsc.br/~ogliari/arquivos/projeto_modelos_de_regressao.doc> Acesso em: 23 set 2007.

BERRY , Michael J. e LINOFF, Gordon. **Data mining techniques for marketing, sales and customer support**. New York: John Wiley & Sons, 1997.

BOSCARIOLI, C. et. al. **Análise de *logs* da *web* por meio de técnicas de *data mining***. Disponível em: <http://conged.deinfo.uepg.br/~iconged/Artigos/Artigo_03.pdf> Acesso em: 20 de out de 2007.

CAMPOS, Omar Barbosa. **Data mining**: overview. Disponível em: <http://www.de9.ime.eb.br/~intec/Data%20Mining/Artigos%20de%20Suporte/Overview%20Data%20Mining.pdf>. Acesso em 14 jun. 2007.

CARVALHO, Déborah Ribeiro. **Data mining:** gestão pedagógica de cursos de graduação. Disponível em: <http://www.niee.ufrgs.br/cbcomp/cbcomp2004/html/pdf/Intelig%EAncia_Artificial/t170100237_3.pdf> Acesso em: 03 de set de 2007.

CARVALHO, Luis Alfredo Vidal de. **Datamining**: a mineração de dados no marketing, medicina, economia, engenharia e administração. Rio de Janeiro: Ciência Moderna, 2005.

CARVALHO, Tiago Buarque Assunção de. **Avaliação do comportamento de métricas de distância com ponderação de características.** Recife, 2007**.** Disponível em: <http://www.cin.ufpe.br/~tg/2007-1/tbac-proposta.doc> Acesso em: 24 set 2007.

DANTON, GIAN. **Metodologia científica**. Disponível em: <http://www.centro-filos.org.br/download/ebooks/gian_danton_metodologia_cientifica.pdf> Acesso em 20 jun. 2007.

DOMINGUES, Marcos Aurélio. **Generalização de regras de associação**. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-10082004-154242/> Acesso em: 15 jul 2007.

DOMINGUES, Marcos Aurélio e REZENDE, Solange Oliveira. **Descriçâo de um algoritmo para generalização de regras de associação**. Disponível em: <ftp://ftp.icmc.usp.br/pub/BIBLIOTECA/rel_tec/RT_228.pdf> Acesso em: 15 jul 2007.

FAYYAD, U.M. *et al***. Advances in knowledge discovery and data mining**. Cambridge, (UK): AAAI Press / MIT Press, 1996.

FAYYAD, Usama *et al*. The KDD proccess for extracting useful knowledge from volumes of data. **Communications on the ACM**, New York, v. 39, n. 11, p. 27-34, Nov. 1996.

FRANK, Eibe e WITTEN, Ian H. **Data mining**: pratical machine learning tools and techniques. Morgan Kaufmann, 2005

GOLDBERG, D. E. ***Genetic algorithms in search, optimization and machine learning***. Massachusetts: Addison-Wesley, 1989. 432 p.

HAN, Jiawey e KAMBER, Micheline. **Data mining**: concepts and techniques. San Diego, CA: Morgan Kaufmann, 2001.

HOUAISS, A. Villar. **Dicionário Houaiss da Língua Portuguesa**. Rio de Janeiro: Objetiva, 2001.

JESUS, Alberto Pereira et. al. **Personalização de sistemas *web* utilizando *data mining*:** um estudo de caso aplicado na biblioteca central da FURB. Disponível em: <http://www.inf.furb.br/seminco/2004/artigos/104-vf.pdf> Acesso em: 03 out 2007.

JUNGUES, Luís Carlos Dill. **Introdução à lógica fuzzy**. Disponível em: <http://s2i.das.ufsc.br/tikiwiki/apresentacoes/logica-fuzzy.pdf> Acesso em: 20 set 2007.

OLIVEIRA, Robson Butaca Taborelli de. **O processo de extração de conhecimento de base de dados apoiado por agentes de software**. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-23092001-231242/>. Acesso em 17 jun. 2007.

PICHILIANI, Mauro. **Data mining na prática**: árvores de decisão. Disponível em: <http://www.imasters.com.br/artigo/5130/sql_server/data_mining_na_pratica_arvores_de_decisao/> Acesso em: Acesso em 15 ago. 2007.

QUINLAN, J. R. (1986). Induction of Decision Trees. In: ***Machine learning***, v.1, n.1, p.81-106.

QUINLAN, J. R. (1993). ***C4.5 Programs for Machine Learning***. San Mateo: Morgan Kaufmann.

RIBACIONKA, Francisco. **Sistemas baseados em lógica *fuzzy***. Disponível em: <http://www.ime.unicamp.br/~laeciocb/orientacoes.htm> Acesso em: 14 jul 2007.

RODRIGUES, Aline Visconti. **Árvores de regressão com dados amostrais complexos**. Disponível em: <http://www.ence.ibge.gov.br/pos_graduacao/mestrado/dissertacoes/pdf/2005/aline_visconti_rodrigues.pdf> Acesso em: 20 set 2007.

ROSENBLATT, Frank. ***Principles of Neurodynamics****: perceptrons and the theory of brain mechanisms.* Spartan Books, New York, 1962.

SANTOS, Maribel Yasmina Campos Alves. **Padrão**: um sistema de descoberta de conhecimento em bases de dados geo-referenciadas. Disponível em: <https://repositorium.sdum.uminho.pt/handle/1822/202> Acesso em: 10 set 2007.

SELLITTO, Miguel Afonso. **Inteligência artificial:** uma aplicação em uma indústria de processo contínuo. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-530X2002000300010&lng=in&nrm=iso&tlng=in> Acesso em: 24 set 2007.

SHANNON, C. E. ***A mathematical theory of communication***. Disponível em: <http://plan9.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf> Acesso em: 24 nov 2007.

TAFNER, Malcon Anderson. **As redes neurais artificiais**: aprendizado e plasticidade. Disponível em: <http://www.cerebromente.org.br/n05/tecnologia/plasticidade2.html> Acesso em: 19 nov 2007.

TSUNODA, Denise Fukumi. **Abordagens evolucionárias para a descoberta de padrões e classificação de proteínas.** Disponível em: <http://www2.cpgei.cefetpr.br/diss_teses/Ano_2004/teses/Abordagens_Evolucionarias_para_a_Descoberta_de_Padroes_e_Classificacao_de_Proteinas.pdf> Acesso em: 30 set de 2007.

UNIVERSIDADE TÉCNICA DE LISBOA. Disponível em: <http://mega.ist.utl.pt/~jpcr/tfc/doku.php?id=regrasassociacao> Acesso em: 15 ago de 2007.

VASCONCELOS, Benitz de Souza. **Mineração de regras de classificação com sistemas de banco de dados objeto-relacional.** Disponível em: <http://copin.ufcg.edu.br/twiki-public/pub/COPIN/DissertacoesMestrado/BenitzDeSouzaVasconcelos.pdf> Acesso em: 24 ago. 2007.

VIEIRA, Valter Afonso. **As tipologias, variações e características da pesquisa de marketing**. Disponível em: <http://www.fae.edu/publicacoes/pdf/revista_da_fae/fae_v5_n1/as_tipologias_variacoes_.pdf > Acesso em: 03 nov 2007.

ZUBEN, Fernando J. Von. **Uma caricatura funcional de redes neurais artificiais**. Disponível em: < http://www.sbrn.org.br/fls/volume2/artigo1.pdf> Acesso em: 19 nov 2007.