CENTRO UNIVERSITÁRIO FEEVALE

CLÁUDIO AURÉLIO DA SILVA

DESCOBERTA DE CONHECIMENTO NA BASE DE DADOS DE UMA ACADEMIA DE MUSCULAÇÃO

Novo Hamburgo, novembro de 2008.

CLÁUDIO AURÉLIO DA SILVA

DESCOBERTA DE CONHECIMENTO NA BASE DE DADOS DE UMA ACADEMIA DE MUSCULAÇÃO

Centro Universitário Feevale
Instituto de Ciências Exatas e Tecnológicas
Curso de Ciência da Computação
Trabalho de Conclusão de Curso

Professor Orientador: Juliano Varella de Carvalho

Novo Hamburgo, novembro de 2008.

RESUMO

As técnicas de mineração de dados, de maneira concisa, são utilizadas para auxiliar a extração de conhecimento. Algumas bases de dados não são aproveitadas da melhor maneira possível, pois informações desconhecidas, que podem se tornar importantes para as organizações, deixam de ser exploradas pelos *decision makers*. O software Weka, através de algoritmos de *data mining*, possibilita um aproveitamento dos dados com maior eficácia. A boa compreensão desse software, das técnicas e dos algoritmos, permite que o conhecimento sobre o ambiente em questão seja realizado de maneira mais completa e eficiente. Este trabalho visa extrair conhecimento da base de dados de uma academia de musculação, através do uso de técnicas de mineração de dados, a fim de obter dados estatísticos sobre os clientes e as atividades em geral dentro desta organização.

Palavras-chave: Mineração de Dados. Descoberta de Conhecimento. Análise de Dados. Classificação. Weka.

ABSTRACT

The techniques of data-mining, in a concise way, are used to assist the extraction of knowledge. Some databases are not exploited in the best possible way, because unknown information that may become important for organizations are no longer explored by decision makers. The Weka software through data mining algorithms, allows a more efficient use of data. The good understanding of this software, the techniques and algorithms, allows that the knowledge about the environment in question is held in a more complete and efficient way. This paper aims to extract knowledge from the database of an academy of weight training, through the use of data-mining techniques in order to obtain statistical data about customers and activities in general within this organization.

Key words: Data mining, Knowledge Discovery, Data analysis, Classification Method, Weka.

LISTA DE FIGURAS

Figura 1.1 – Etapas do Processo de KDD 14
Figura 1.2 – Interdisciplinaridade da Mineração de Dados 18
Figura 2.1 – Tipos de Nodos de uma Árvore de Decisão 23
Figura 2.2 – Estrutura de uma Rede Neural Simples 27
Figura 3.1 – Interface Principal do *Sipina* 32
Figura 3.2 – Interface do *QwikNet* 33
Figura 3.3 – Regras Geradas pelo *Sipina* 34
Figura 3.4 – Localização do município de Capixaba, Acre 36
Figura 3.5 – Prevalência da Esquistossomose em 197 Municípios de Minas Gerais 40
Figura 3.6 – Árvore de Decisão Gerada pelo Algoritmo J4.8 42

LISTA DE TABELAS

Tabela 3.1 – Matriz de Erros Obtida com o Algoritmo de Árvore de Decisão 38
Tabela 3.2 – Matriz de Erros Obtida com o Algoritmo J4.8 41

LISTA DE GRÁFICOS

Gráfico 1.1 – Percentagem de esforço para cada etapa do processo de KDD.....20

LISTA DE ABREVIATURAS E SIGLAS

DDD	Discagem Direta a Distância
DDI	Discagem Direta Internacional
DM	Data Mining
EUA	Estados Unidos da América
IBGE	Instituto Brasileiro de Geografia e Estatística
ID3	Idemized Dichotomizer 3
KDD	Knowledge Discovery Database
LMT	Logistic Model Tree
MLP	Perceptron Multi-Camadas
RNA	Rede Neural Artificial
SGBD	Sistema Gerenciador de Banco de Dados
SIG	Sistemas de Informação Geográficos
SR	Sensoriamento Remoto
TM	Thematic Mapper
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

INTRODUÇÃO 10

1 KNOWLEDGE DISCOVERY DATABASE 13
1.1 Pré-Processamento 15
1.1.1 Limpeza dos Dados 15
1.1.2 Seleção dos Dados 15
1.1.3 Transformação dos Dados 16
1.2 Mineração de Dados 16
1.3 Pós-Processamento 19

2 CLASSIFICADORES 21
2.1 Árvores de Decisão 21
2.1.1 Histórico 22
2.1.2 Conceitos 22
2.1.3 Vantagens e Desvantagens 24
2.1.4 Algoritmo C4.5 25
2.1.5 Algoritmo LMT 26
2.2 Redes Neurais 26
2.2.1 *Perceptron* Multi-Camadas 29

3 APLICAÇÕES DE MINERAÇÃO DE DADOS 31
3.1 Análise do perfil do usuário de serviços de telefonia utilizando técnicas de mineração de dados (JUNIOR; PEREZ, 2006) 31
3.2 Avaliação da exatidão do mapeamento da cobertura da terra em Capixaba, Acre, utilizando classificação por árvore de decisão (CARVALHO; FIGUEIREDO, 2006) 35
3.3 Uso de árvore de decisão para predição da prevalência de esquistossomose no Estado de Minas Gerais, Brasil (MARTINS et al, 2007) 39

CONCLUSÃO 44

REFERÊNCIAS BIBLIOGRÁFICAS 45

INTRODUÇÃO

Em 1989 foi formalizado um termo para denominar o abrangente conceito de buscar conhecimento a partir de bases de dados: o KDD – *Knowledge Discovery Database*. Historicamente a Descoberta de Conhecimento em Bases de Dados foi consolidada a partir de várias disciplinas, sendo que entre elas pode-se citar a Estatística, a Inteligência Artificial, o Reconhecimento de Padrões e Banco de Dados (GOLDSCHMIDT, 2005).

KDD é um processo, de várias etapas, para descoberta de informações que sejam úteis e estejam implícitas nas grandes bases de dados. O exponencial aumento no volume de dados que não podem ser restaurados de uma maneira adequada pelos limites e capacidades de consultas dos SGBD's atuais, faz com que a utilização do KDD tenha uma grande importância na atualidade (BORGES, 2006).

O uso de técnicas para exploração de grandes quantidades de dados, a fim de descobrir padrões e relações, que demandariam grande trabalho por parte do ser humano, é definido como *Data Mining* (CARVALHO, 2005). A Mineração de Dados é a principal etapa do processo de KDD, onde efetivamente é feita a busca por conhecimentos que possam se tornar úteis no conjunto da aplicação da Descoberta de Conhecimento em Bases de Dados (GOLDSCHMIDT, 2005).

Para a solução de diversos casos que relatam os problemas de mau aproveitamento nas bases de dados, a aplicação de técnicas de mineração auxilia na abstração de conhecimento. Carvalho (2005) cita um exemplo de caso de segmentação de mercados, onde a venda foi otimizada através da análise e mineração na base de dados que contém os registros de venda, a fim de indicar quais as tendências de aquisição de determinados clientes, considerando vários aspectos. Também exemplifica a situação de uma empresa que utilizou as técnicas de redes neuronais artificiais para previsão de mercados financeiros e se tornou uma

das mais importantes no ramo, nos EUA, durante sete anos consecutivos, tendo crescimento de sua carteira, neste período, de 25% a 100% ao ano.

Existem diversas técnicas para Descoberta de Conhecimento em Bases de Dados, sendo que neste trabalho será utilizada a de Classificação. A técnica tem como objetivo classificar casos em classes distintas, levando em conta os atributos comuns a um determinado conjunto de objetos, pertencentes a uma base de dados. Esse modelo gerado possibilita a descoberta das classes de novos objetos a serem adicionados na base (SOUSA, 1998).

O processo de classificação é definido por dois passos, onde no primeiro as características dos dados formam um modelo de classificação, e no segundo, esse modelo criado é empregado com o objetivo de classificar novos objetos. Segundo Passini e Toledo (2002), a construção do modelo pode ser dividida em três fases: o treinamento, onde são estabelecidos parâmetros para se treinar o modelo, a fase de teste, onde a precisão do mesmo é testada com a aplicação de dados diferentes, e a aplicação, que é responsável pela execução da técnica.

Dessa forma, este trabalho visa extrair conhecimento da base de dados de uma academia de musculação, através do uso de técnicas de classificação, a fim de obter dados estatísticos sobre os clientes, além de encontrar alguma relação das atividades dos alunos com a sua freqüência na academia, evolução do peso e medidas dos alunos de acordo com faixa etária, sexo entre outras informações. Com isso, é possível criar perfis de alunos e focar determinados tipos de serviços de acordo com as características destes perfis.

A base disponibilizada está no Microsoft Office Access, contendo 42 tabelas no total. Possui desde dados sobre vendas de produtos até informações sobre os clientes. São no total 6.993 membros desde a data da implementação do sistema, no ano de 1997.

É realizada uma avaliação periódica nos alunos, medindo seu peso, altura, cintura, e várias outras características dos mesmos. Visto que cada acesso dos membros é gravado na tabela Entradas, é possível obter um controle da assiduidade de cada aluno para levantamento de dados, como a evolução de cada pessoa de acordo com seu desempenho.

Auxiliar a tomada de decisão na academia, promovendo ações que resultem em um conhecimento mais profundo de seus clientes, e por conseqüência permitir uma análise de tendências, é uma motivação adicional para a realização deste trabalho.

Sendo assim, no capítulo 1 será abordado todo o processo de KDD, desde a seleção dos dados até o pós-processamento, onde o conhecimento obtido é analisado e fica pronto para aplicação, se o resultado for considerado satisfatório. No capítulo 2 serão abordados dois tipos de classificadores: árvores de decisão e redes neurais. Também será abordado neste capítulo o funcionamento de alguns algoritmos utilizados por estas técnicas. No capítulo 3 serão apresentados três estudos de casos a fim de demonstrar que as técnicas de mineração de dados estão sendo cada vez mais utilizadas em diversas áreas e que seus resultados são altamente satisfatórios.

1 KNOWLEDGE DISCOVERY DATABASE

Atualmente, em vários segmentos do mercado, o volume de dados cresce exponencialmente, gerando grandes problemas para as organizações. Os responsáveis por armazenar essas informações perdem o controle sobre o conteúdo armazenado, precisando encontrar formas para resolver este problema. Quanto maior a base de dados, mais difícil fica a extração de algo que possa ser útil à empresa, caso os mesmos não estejam armazenados de maneira correta.

Todas as informações armazenadas nas bases de dados podem ser vistas de maneira superficial pelos proprietários de empresas. Contudo, alguns dados não são aproveitados da melhor maneira possível, pois informações desconhecidas, que podem se tornar importantes para as organizações, deixam de ser exploradas pelas pessoas responsáveis. Sendo assim, mostra-se necessária a criação e utilização de tecnologias para o processo de recuperação de informações.

O termo em inglês Knowledge Discovery Database (KDD), utilizado para referenciar a descoberta de conhecimento em bases de dados, segundo Fayyad, Piatetsky-Shaphiro e Smyth (1996, p. 6), "é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados". De um modo geral, a tarefa mais difícil do processo de KDD é perceber e interpretar corretamente vários fatos observáveis durante o processo, e a dificuldade em conjugar dinamicamente essas interpretações de forma a tomar a decisão de quais ações devem ser realizadas em cada caso (GOLDSCHMIDT, 2005).

Uma característica muito relevante que deve ser levada em consideração é que a descoberta do conhecimento não se dá exclusivamente por métodos e algoritmos, é necessário que exista a interferência humana a fim de delimitar quais são os níveis e interpretar se as

respostas geradas serão úteis dentro do contexto (GONCHOROSKI, 2007). O KDD é caracterizado como um processo formado por várias etapas operacionais, conforme representado na figura 1.1 (GOLDSCHMIDT, 2005).

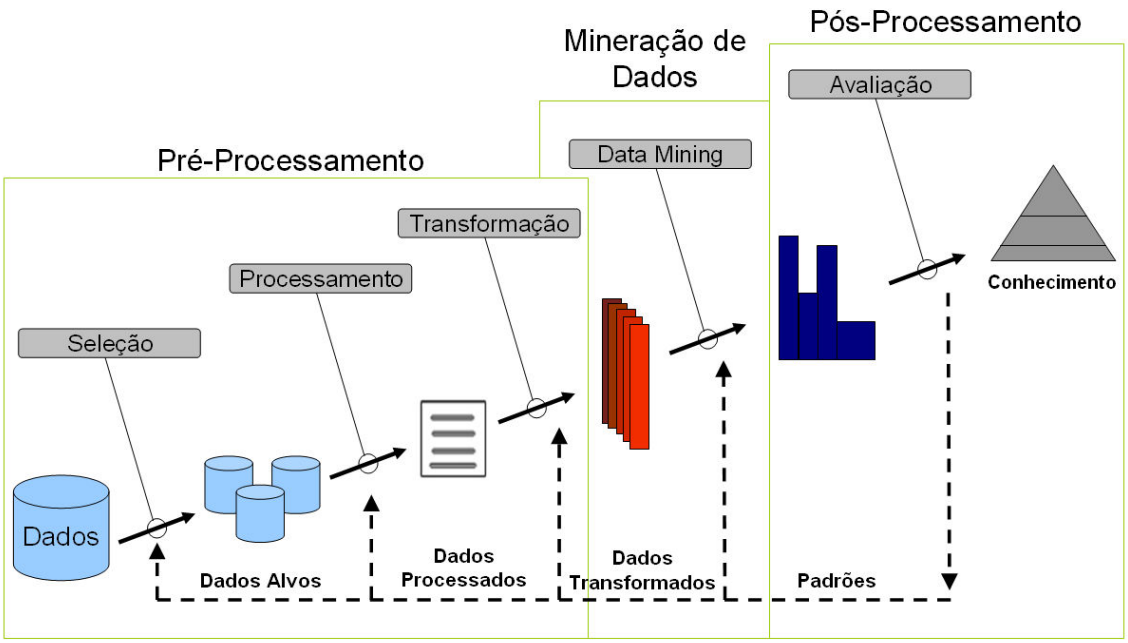

Figura 1.1 – Etapas do Processo de KDD
Fonte: Adaptado de GONCHOROSKI, 2007, p. 30

De acordo com Borges (2006), o processo de KDD pode ser dividido em três grandes fases: o Pré-Processamento, que abrange todas as funções relacionadas à captação, organização e o tratamento dos dados; a Mineração de dados onde é feita a descoberta de padrões; e o Pós-Processamento que abrange os padrões e conhecimento descoberto. Em primeiro lugar, antes de executar as etapas do processo de KDD, é preciso que seja feita a análise e definição das metas a serem alcançadas com a extração do conhecimento no contexto da aplicação. É neste momento que são definidos itens importantes como a união entre o escopo de aplicação e a tecnologia KDD, visando ter a relação custo-benefício ao aplicar esta tecnologia (GONCHOROSKI, 2007).

1.1 Pré-Processamento

A etapa de pré-processamento possui um papel essencial no processo de descoberta de conhecimento. Nela é realizada desde a correção de dados obsoletos até a adequação da formatação dos dados para os algoritmos de Mineração de Dados a serem empregados (GOLDSCHMIDT, 2005). Outro fator importante nesta etapa é a verificação de predominância de classes, sendo que uma vez constatada, é necessário excluir alguns dos registros da classe predominante ou adicionar registros de outras classes. Este processo objetiva balancear a base de dados de tal forma que, no processo do aprendizado, determinada classe não seja beneficiada, impedindo que o sistema fique tendencioso (ALMEIDA, 2003).

1.1.1 Limpeza dos Dados

Presente na etapa de pré-processamento, a limpeza dos dados pode ser realizada utilizando o conhecimento do domínio. Pode-se localizar registros com valores nulos em algum atributo, granularidade incorreta ou exemplos errôneos, por exemplo. A limpeza pode também ser feita independente de domínio, como decisão da estratégia para tratamento de atributos incompletos, remoção de ruídos, entre outros (REZENDE, 2005). Os campos devem ser tratados por um analista que pode fazer interpolações, entrar códigos especiais nestes campos, ou simplesmente eliminar os registros com estas informações. Esta medida deve considerar o tipo de dados e seu impacto no processo de descoberta de conhecimento (BORGES, 2006).

1.1.2 Seleção dos Dados

Na etapa de seleção identificam-se as bases de dados e quais variáveis e tipos de dados serão extraídos na fase de Mineração de Dados. Por exemplo, alguns dados, como telefone, endereço e e-mail, poderiam ser descartados por não terem utilidade no contexto de associações entre compras de clientes. (CARVALHO, 2000). Esta etapa pode ter dois enfoques diferentes: a escolha de atributos ou a escolha de registros que devem ser levados em conta no processo de KDD (GOLDSCHMIDT, 2005.)

1.1.3 Transformação dos Dados

De acordo com Rezende (2005) algumas transformações comuns podem ser aplicadas aos dados, entre elas: resumo, onde dados sobre vendas, por exemplo, podem ser agrupados para formar relatórios diários; transformação de tipo: quando algum atributo é transformado em outro tipo de dados para ser aproveitado da melhor maneira possível por algum algoritmo de extração de padrões. Portanto, a principal finalidade desta etapa é transformar os dados pré-processados, a fim de ajustá-los de acordo com a entrada de algum dos diversos algoritmos de Mineração de Dados (CARVALHO, 2000).

As próximas etapas do processo serão compostas pela própria Mineração dos Dados, onde será feita a escolha de quais algoritmos serão utilizados e, por fim, o pós-processamento, onde é realizada a análise dos resultados obtidos de modo a verificar se o conhecimento adquirido será útil ao contexto proposto. É no momento da análise que é feita a constatação se será necessário que o processo seja reiniciado, caso o resultado não esteja dentro dos objetivos definidos inicialmente (GONCHOROSKI, 2007).

A principal etapa do processo de KDD, chamada de Mineração de Dados, é composta por diversas técnicas conhecidas como seus algoritmos, com uma grande complexidade. Sendo assim o processo e algumas dessas técnicas serão abordadas a seguir.

1.2 Mineração de Dados

Na atualidade, sistemas que usufruem do potencial de Mineração de Dados têm sido utilizados em vários ramos do mercado, com intuito de extrair conhecimento de grandes bases de dados. Qualquer algoritmo que gere um padrão a partir de dados é um algoritmo de Mineração de Dados. (CARVALHO, 2000)

O processo de Mineração de Dados baseia-se na interação entre várias classes de usuários, e grande parte do seu sucesso depende dessa interação. Existem três classes diferentes nas quais podem ser divididos os usuários deste processo: especialista do domínio, que deve oferecer apoio para a execução do processo e possuir grande conhecimento do domínio da aplicação; analista, que deve conhecer profundamente todas as etapas que fazem parte do processo e é o usuário especialista no processo de extração de conhecimento; e o usuário final, que utiliza o conhecimento obtido no processo para a tomada de decisão (REZENDE, 2005).

De acordo com Borges (2006) a Mineração de Dados ou *Data Mining* (DM) é a principal etapa do processo KDD, que tem como finalidade extrair padrões dos dados. Esta fase é considerada o centro de todo o processo onde a maior preocupação é ajustar modelos ou determinar padrões de acordo com os dados observados. Pode ser vista também como uma maneira de selecionar, explorar e modelar grandes quantidades de dados a fim de detectar padrões de comportamento. Já de acordo com Fayyad, Piatetsky-Shaphiro e Smyth (1996), a mineração de dados é o processo de extrair informações desconhecidas, válidas e acionáveis de grandes conjuntos de dados para então aplicar o conhecimento obtido em decisões cruciais no mundo dos negócios.

Na prática, os objetivos de DM são a predição ou a descrição. Compreende-se por predição a utilização de alguns atributos da base de dados para predizer valores desconhecidos ou futuros de outras variáveis de interesse. Já a descrição procura por padrões que descrevem os dados interpretáveis pelos seres humanos (MARTINHAGO, 2005).

Mineração de Dados é uma área interdisciplinar que integra principalmente estatística, inteligência artificial e banco de dados. Pode-se afirmar isto, pois ao realizar várias medidas estatísticas, os algoritmos de *data mining* conseguem, por exemplo, classificar ou relacionar itens de uma base de dados. Os algoritmos podem também ser aplicados em um grande conjunto de dados armazenados aproveitando-se de métodos de indução com base na Inteligência Artificial. A Figura 1.2 ilustra a interdisciplinaridade da tecnologia de Mineração de Dados (CARVALHO, 2000).

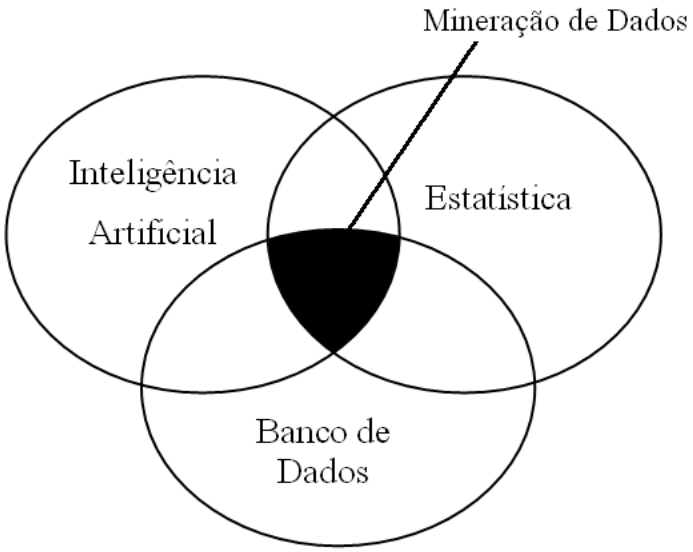

Figura 1.2 – Interdisciplinaridade da Mineração de Dados
Fonte: Adaptado de CARVALHO, 2000, p. 25

Os algoritmos de Mineração de Dados são essenciais nesta etapa do processo de KDD, pois alguns deles têm capacidade de aprender a partir de exemplos. Tais algoritmos assimilam relacionamentos eventuais existentes entre os dados, utilizando o resultado deste aprendizado nos modelos de conhecimento gerados (GOLDSCHMIDT, 2005).

Abordando técnicas estatísticas em DM, é possível influenciar significativamente todas as áreas de uma organização. As técnicas estatísticas ajudam a assimilação e reação às mudanças de mercado, fazendo com que a organização se torne mais produtiva e competitiva, além de tomar decisões baseadas em fatos (BORGES, 2006).

A tecnologia de Mineração de Dados tem grande potencial para auxiliar as organizações a extrair importantes informações oriundas das suas bases de dados, formulando padrões e comportamentos futuros, ajudando a responder questões que demandariam muito tempo para serem resolvidas, possibilitando melhores decisões de negócio apoiadas no conhecimento extraído. Assim, é possível afirmar que a Mineração de Dados é um recurso em grande ascensão e se tornará obrigatório aos mercados mais competitivos (MARTINHAGO, 2005).

1.3 Pós-Processamento

A última etapa do processo de KDD é de grande relevância, mesmo sendo a mais simples, uma vez que não exige grande esforço computacional e tempo do usuário em relação às outras etapas. É nesta etapa que o usuário que acompanhou todo o processo define se o conhecimento gerado será útil e aplicável. É muito comum que ao final do processo, os algoritmos utilizados apresentem ao usuário algumas informações não relevantes, sendo que cabe a este analisá-las e decidir aplicá-las ou não, dependendo do contexto da aplicação (GONCHOROSKI, 2007).

Segundo Santos (2008), esta etapa deve ser executada pelo analista de dados, responsável pelas etapas anteriores, e pelo analista de negócios, responsável por analisar se o conhecimento obtido será útil. Pode-se também contatar o executivo responsável para que o mesmo forneça esclarecimentos sobre o conhecimento descoberto, relacionando-os aos objetivos do negócio a fim de validá-los. Já de acordo com Martinhago (2005), a etapa de pós-processamento consiste na validação do conhecimento extraído da base de dados, identificação de padrões e interpretação dos mesmos, transformando-os em conhecimentos apoiadores na tomada de decisão. O objetivo de interpretar os resultados é filtrar as informações que serão apresentadas aos tomadores de decisão.

Geralmente, a meta principal desta etapa é fazer com que o conhecimento descoberto seja compreendido da melhor maneira possível, validando-os através de medidas da qualidade da solução e da percepção de um analista de dados. O conhecimento gerado será consolidado em forma de relatórios, sendo feita a documentação e explicação das informações obtidas em cada etapa do processo de KDD (BORGES, 2006).

Visto que o interesse para um determinado padrão gerado no processo de extração de conhecimento varia de acordo com cada usuário e ramo de mercado, medidas subjetivas são necessárias. Quando um conjunto de regras interessantes é selecionado, estas medidas consideram que fatores específicos devem ser tratados, como o conhecimento do domínio e o interesse do usuário (REZENDE, 2005).

Durante as etapas do processo de KDD esforços diferentes são consumidos. A fase que mais consome esforço computacional e tempo do usuário é a de preparação dos dados, conforme mostra o gráfico 1.1 (CARVALHO, 2000).

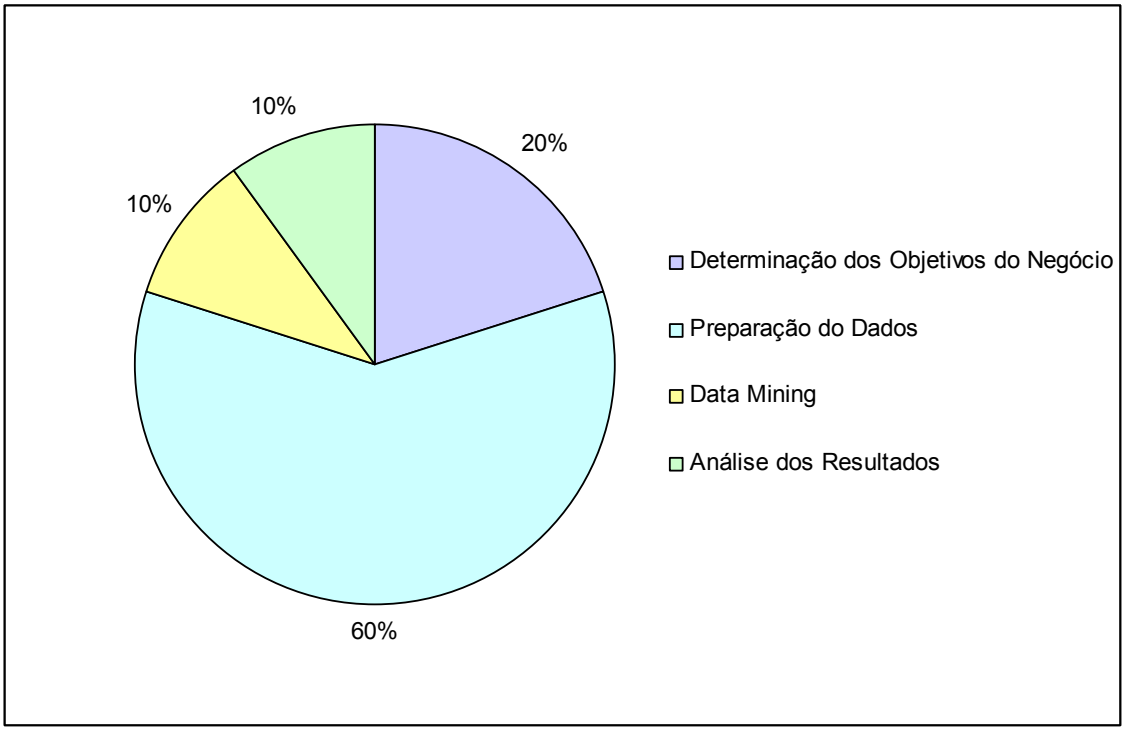

Gráfico 1.1 – Percentagem de esforço para cada etapa do processo de KDD
Fonte: Adaptado de CARVALHO, 2000, p. 24

A etapa de pós-processamento encerra o ciclo da descoberta do conhecimento e é nela que se coloca em ação todo o conhecimento adquirido durante as etapas anteriores. Após interpretar e avaliar o resultado obtido, o usuário vai identificar a necessidade ou não de reiniciar o processo e gerar outro tipo de regra ou informação. Se os resultados não forem satisfatórios, faz-se necessário repetir a etapa de Mineração de Dados ou retomar qualquer um dos estágios anteriores. O conhecimento é encontrado somente após a avaliação e validação dos resultados (MARTINHAGO, 2005).

Como dito anteriormente, algumas técnicas e algoritmos são utilizados no processo de Descoberta de Conhecimento. Na etapa de Mineração de Dados um dos métodos mais conhecidos e utilizados é a Classificação, que dispõe de algumas técnicas para extrair conhecimento. Entre essas pode-se citar as árvores de decisão e redes neurais como sendo as mais estudadas. Visando uma abordagem mais completa e detalhada sobre estas técnicas, torna-se necessário um estudo sobre o histórico, vantagens e desvantagens entre outras características das mesmas.

2 CLASSIFICADORES

Na tarefa de classificação, existem algumas técnicas que são utilizadas para a extração de conhecimento de bases de dados sendo que a seguir serão abordadas somente duas delas: árvores de decisão e redes neurais. Estas duas técnicas baseiam-se no aprendizado supervisionado, na qual os resultados obtidos necessitam de análise de um especialista que fará a avaliação da relevância dos mesmos. Estas técnicas geram modelos a partir de exemplos de uma base de dados, denominados conjunto de treinamento, representando uma amostra dos registros que serão analisados (GONCHOROSKI, 2007).

Objetiva-se com este estudo a comparação das duas técnicas, observando suas vantagens e desvantagens, de acordo com os resultados obtidos, permitindo a escolha da técnica que revela os resultados mais adequados dentro do contexto da aplicação.

2.1 Árvores de Decisão

As árvores de decisão possuem este nome devido a sua estrutura, muito compreensível e assimilativa, se assemelhar a uma árvore. Suas técnicas dividem os dados em subgrupos, baseadas nos valores das variáveis, sendo que o resultado disto é uma hierarquia de declarações do tipo "Se...então..." que são principalmente aplicadas quando o grande objetivo da mineração de dados é a classificação de dados ou a predição de saídas. (MARTINHAGO, 2005).

De acordo com Goldschmidt (2005, p. 109), uma árvore de decisão pode ser definida como "um modelo de conhecimento em que cada nó interno da árvore representa uma decisão sobre um atributo que determina como os dados estão particionados pelos seus nós filhos". Já de acordo com Sousa (1998), métodos de árvore de decisão representam um tipo de algoritmo

de aprendizado de máquina, que fazem uso de uma abordagem dividir-para-conquistar para classificar casos, representando-os em forma de árvores.

2.1.1 Histórico

Muitas pessoas na área de Mineração de Dados consideram Ross Quinlan, da Universidade de Sydney, Austrália, o criador das árvores de decisão. Isto se deve, em grande parte, pela criação de um novo algoritmo chamado de ID3 (*Itemized Dichotomizer* 3), desenvolvido em 1983. O algoritmo ID3 e versões posteriores como o ID4 e o C 4.5, por exemplo, são estruturados de tal forma que se adaptam muito bem ao serem utilizados em conjuntos com árvores de decisão, visto que eles produzem regras ordenadas por importância. Essas regras são utilizadas na produção de um modelo de árvore de decisão dos fatos que afetam os itens de saída (JERONIMO, 2001).

Pode-se dizer que as árvores de decisão são uma evolução das técnicas que apareceram durante o desenvolvimento das disciplinas de *machine learning*. A partir da aproximação conhecida como Detecção de Interação Automática, desenvolvida na Universidade de Michigan, as árvores de decisão foram ganhando maior importância no meio científico (KRANZ, 2004).

2.1.2 Conceitos

Considerada uma ferramenta completa e bastante conhecida para classificação dos dados e apresentação dos resultados na forma de regras, as árvores de decisão são utilizadas freqüentemente no processo de Descoberta de Conhecimento. A classificação é executada, na maioria das vezes, em duas fases no uso das árvores de decisão: construção da árvore e poda (OLIVEIRA, 2001). Nesta técnica, o usuário escolhe o atributo que quer avaliar para que o algoritmo procure as variáveis mais correlacionadas, gerando uma árvore de decisão com inúmeras ramificações. A árvore criada será utilizada na classificação de novas instâncias, de acordo com os valores dos atributos da nova instância (ARAÚJO, 2006).

Pode-se considerar as árvores de decisão como um algoritmo supervisionado, pois há a necessidade de ser informadas com antecedência as classes dos registros usadas no conjunto de treinamento. Uma árvore de decisão é formada por um conjunto de nós que são conectados através de ramificações, estes nós se dividem em três tipos conforme mostra a figura 2.1, onde

o nodo raiz é o início da árvore, os nodos comuns dividem um atributo e geram novas ramificações e o nodo folha possui as informações de classificação do registro (SANTOS, 2008).

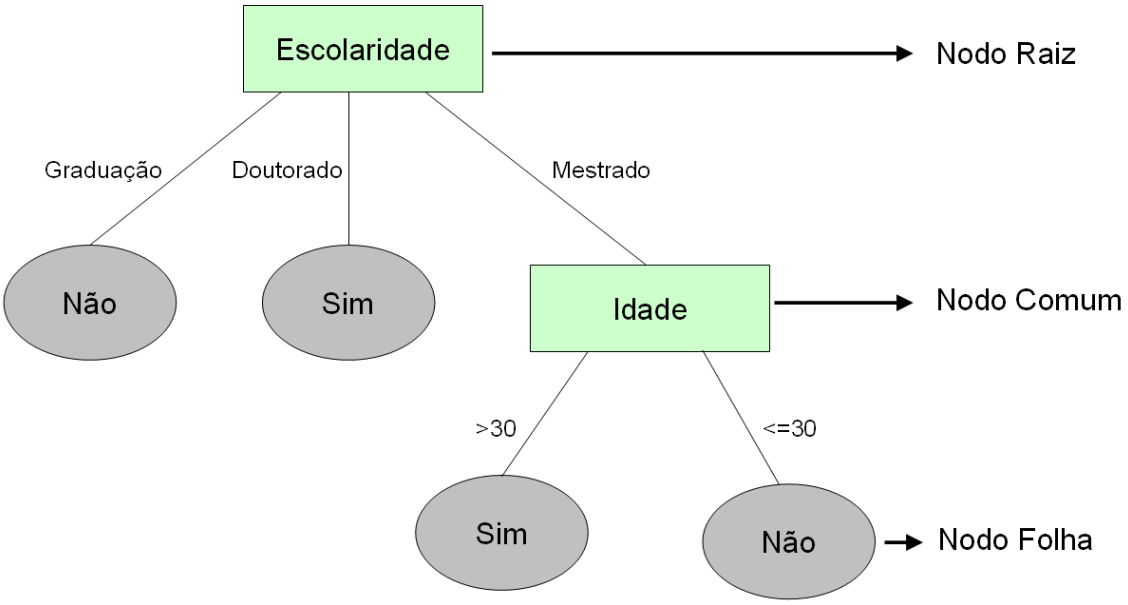

Figura 2.1 – Tipos de Nodos de uma Árvore de Decisão
Fonte: SANTOS, 2008, p. 21

Na fase de construção da árvore, é realizada ramificações na árvore através de sucessivas divisões dos dados com base nos valores dos atributos. Sendo assim, o processo é repetido recursivamente até que todos os registros pertençam a uma classe (OLIVEIRA, 2001).

Um registro entra na árvore pelo nó raiz. A partir deste nodo todos os outros nodos são percorridos até ser alcançado o nodo folha. Cada um dos nodos testa o valor de um único atributo e oferece arestas distintas a serem percorridas na árvore a partir deste nodo, para cada uma de suas valorações. Assim é determinado o próximo nodo no qual o registro irá se posicionar. Podem ser utilizados diferentes algoritmos na escolha do teste inicial, porém, todos têm o mesmo objetivo: escolher aquele que melhor descreve a classe alvo. Quando o algoritmo chega ao nodo folha, todos os registros que terminam na mesma folha são classificados da mesma forma. É importante salientar que existe somente um caminho da raiz

até cada folha, que significa a expressão utilizada para classificar os registros (BORGES, 2006).

Após a fase de crescimento da árvore, pode-se encontrar uma estrutura especializada que está super ajustada aos dados, sendo que desta maneira é oferecida mais estrutura que o necessário. Então, a poda passa a ter um papel crucial, fazendo com que sejam consideradas árvores menores e potencialmente de melhor precisão (SOUSA, 1998).

Na fase de poda, as ramificações que não tem valor significativo são removidas, a fim de criar um modelo de classificação, fazendo a seleção da sub-árvore que contém a menor taxa de erro estimada (OLIVEIRA, 2001).

Após a fase de poda, a árvore gerada pode representar uma estrutura complexa e de difícil compreensão. Nestes casos pode-se utilizar a extração de regras como uma fase final, visando extrair regras menores e menos complexas, porém, com precisão similar (SOUSA, 1998).

2.1.3 Vantagens e Desvantagens

O uso de árvores de decisão possui algumas vantagens em relação às outras técnicas, dentre as quais pode-se citar: facilidade de compreender o modelo obtido, uma vez que tem a forma de regras explícitas, possibilitando a avaliação dos resultados e a identificação dos seus atributos chaves no processo; facilidade de expressar as regras como instruções lógicas sendo aplicadas diretamente aos novos registros; árvores de decisão são relativamente mais rápidas em comparação às redes neuronais, por exemplo, e na maioria das vezes se obtém mais precisão nos resultados quando comparadas à outras técnicas de classificação (OLIVEIRA, 2001).

Segundo Sousa (1998), as principais desvantagens no uso de árvores de decisão estão na necessidade de uma considerável quantidade de dados para desvendar estruturas complexas e na possibilidade de haver erros na classificação, no caso de existirem muitas classes, bem como o tratamento de dados contínuos.

Para melhor compreensão dos algoritmos de árvore de decisão, serão apresentados a seguir dois deles: o C4.5 e o *Logistic Model Tree* (LMT).

2.1.4 Algoritmo C4.5

Considerado um dos mais tradicionais algoritmos na tarefa de Classificação, o C4.5 foi inspirado no algoritmo ID3, sendo que seu método visa abstrair árvores de decisão seguindo uma abordagem recursiva de particionamento das bases de dados (GOLDSCHMIDT, 2005). Também desenvolvido pelo pesquisador australiano Ross Quinlan, em 1993, este algoritmo encontra-se disponível em diversos softwares de mineração. O C4.5 transforma a árvore de decisão em um conjunto de regras ordenadas pela sua importância, possibilitando ao usuário a identificação dos fatores mais relevantes em seus negócios (OLIVEIRA, 2001)

A principal vantagem do algoritmo C4.5 em relação ao ID3, é que ele tem o poder de lidar com a poda (*prunning*) da árvore, evitando o sobre-ajustamento, com a valoração numérica de atributos e com a presença de ruído nos dados (BORGES, 2006). Na maioria das vezes uma árvore originada do algoritmo C4.5 precisa ser podada pela necessidade de redução do excesso de ajuste (*overfitting*) aos dados de treinamento (MARTINHAGO, 2005).

Enquanto o algoritmo ID3 manipula apenas dados nominais, o C4.5 pode manipular também dados numéricos. Entretanto, trabalhar com dados numéricos não é tão simples, pois enquanto os atributos nominais são testados apenas uma vez em qualquer caminho da raiz às folhas, os atributos numéricos podem ser testados diversas vezes no mesmo percurso. Esta característica pode ser considerada uma possível desvantagem do C4.5, pois em alguns casos, a árvore gerada pode ser de difícil entendimento do usuário (BORGES, 2006).

Neste algoritmo é utilizada a abordagem "dividir para conquistar", em que o problema original é dividido em partes semelhantes ao original, porém menores, fazendo com que os problemas sejam resolvidos e suas soluções formem uma combinação para o problema inicial. Ele ainda possui a capacidade de aprimorar a estimativa do erro utilizando uma técnica conhecida como *v-fold*, onde é realizada a validação cruzada com dois ou mais grupos (GONCHOROSKI, 2007).

De acordo com Oliveira (2001), o algoritmo cria uma árvore com uma quantidade aleatória de folhas por nodo e assume os valores das categorias como um divisor, diferentemente do que realiza algoritmos que produzem árvores binárias, por exemplo. Então o *prunning* é realizado de acordo com a taxa de erro de cada nodo e seus descendentes, sendo que a soma dessas taxas compõem a taxa de erro da árvore. Para a identificação do nodo raiz e

de seus descendentes são realizados os cálculos da entropia e do ganho de informação (OLIVEIRA, 2001).

Após a criação de um conjunto de regras, o algoritmo realiza o agrupamento das regras obtidas para cada classe e a eliminação das regras que não possuem relevância na precisão do conhecimento a ser extraído. Como resultado final, se obtém um pequeno conjunto de regras que podem ser facilmente entendidas, criadas pela combinação das regras que induzem à mesma classificação (MARTINHAGO, 2005).

2.1.5 Algoritmo LMT

O algoritmo *logistic model tree* (LMT) aplica os princípios das árvores em problemas de classificação, utilizando para a construção da árvore a regressão logística, que tem como objetivo saber quais variáveis independentes influenciam no resultado, utilizando uma equação para prever um resultado baseado nestas variáveis. Um processo de adaptação por etapas é empregado na construção dos modelos de regressão nos nodos folhas, realizando uma redefinição incremental àqueles construídos em níveis superiores da árvore (ARAÚJO, 2006).

Este algoritmo normalmente é utilizado para a predição numérica, sendo que os nodos folhas gerados armazenam um modelo de regressão logística para geração do resultado. Após a construção da árvore, é aplicada uma regressão para cada nodo interior, utilizando os dados associados a este nodo e todos os atributos que participam nos testes na sub-árvore. Em seguida, os modelos de regressão logística são simplificados, utilizando a poda. Porém, a poda só acontecerá se o erro estimado para o modelo na raiz de uma sub-árvore for menor ou igual ao erro esperado para a sub-árvore. Após a poda é realizado um processo que forma o modelo final, colocando-o no nodo folha (LANDWEHR; HALL; FRANK, 2003).

2.2 Redes Neurais

Uma Rede Neural Artificial (RNA) é uma técnica computacional que cria um modelo matemático, emulado por computador, simulando um sistema neural biológico simplificado, que tem como principal característica a capacidade de aprendizado, generalização, associação e abstração (ARAÚJO, 2006). São consideradas as técnicas mais comuns utilizadas pelos processos de Mineração de Dados e possuem uma característica que

as diferenciam das outras técnicas: podem gerar saídas iguais às entradas, que não existiam durante a fase de treinamento (SANTOS, 2008).

Duas características fazem com que as redes neurais sejam semelhantes ao cérebro: o conhecimento é adquirido pela rede através de seu ambiente utilizando um processo de aprendizagem; e as forças de conexão entre os neurônios, mais conhecidas como pesos sinápticos, são usadas para o armazenamento do conhecimento obtido (HAYKIN, 2001). De acordo com Sousa (1998), os métodos baseados em RNA proporcionam métodos mais práticos para funções de aprendizado, que são representadas por atributos contínuos, discretos ou vetores.

A estrutura de uma rede neural consiste em uma quantidade de neurônios interconectados que são organizados em camadas. O conhecimento através destas camadas se dá através da modificação das conexões, que são responsáveis pela comunicação entre as camadas (ARAÚJO, 2006). A figura 2.2 apresenta a arquitetura de uma rede neural simples em que os círculos representam os neurônios, e as linhas representam os pesos das conexões.

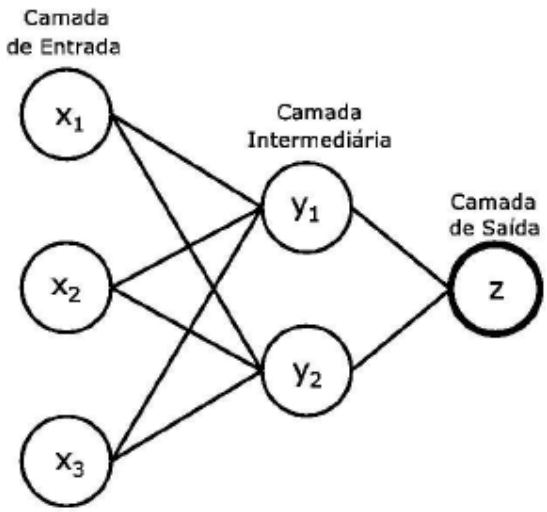

Figura 2.2 – Estrutura de uma Rede Neural Simples
Fonte: ARAÚJO, 2006, p. 30

Todas as camadas de uma rede neural possuem funções específicas. A camada de entrada é a que recebe os dados a serem analisados. A camada intermediária é responsável pelo processamento interno das informações e extraem características, permitindo que a rede

crie sua própria representação. É importante salientar que uma RNA pode conter várias camadas intermediárias, de acordo com a complexidade do problema. A camada de saída recebe os estímulos da camada intermediária, construindo o padrão que será a resposta para o problema em análise (ARAÚJO, 2006).

O processo de aprendizado de uma rede neural pode ser realizado de duas formas:

- Supervisionado: é utilizado um conjunto de pares de dados de entrada e saída desejada. A partir dos conjuntos de entrada, a rede neural cria um conjunto de valores de saída desejado. Na existência de grande diferença entre as saídas, os pesos sinápticos e os níveis de bias[1], são acertados até que a diferença seja diminuída (SANTOS, 2008).

- Não supervisionado: o treinamento da rede se dá apenas através de valores de entrada. Assim, são realizados processos, chamados de competição e cooperação, entre os neurônios para a classificação dos dados, obtendo-se um reconhecimento de padrões (SANTOS, 2008).

Existem redes neurais com apenas um neurônio, chamadas de *perceptron*, que são a unidade mais simples de uma rede neural. Através de um vetor com números reais de entrada, é possível que o *perceptron* calcule uma combinação linear destes atributos para retornar um resultado. Esta rede de apenas uma camada é capaz de solucionar problemas que sejam linearmente separáveis (HAYKIN, 2001).

De acordo com Martinhago (2005), a grande vantagem da utilização de redes neurais é a grande versatilidade que elas possuem, sendo que o resultado é satisfatório até mesmo em áreas complexas, com entradas incompletas ou imprecisas. Além disso, as redes neurais possuem excelente desempenho em problemas de classificação e reconhecimento de padrões (SANTOS, 2008).

As desvantagens da utilização das redes neurais estão ligadas à solução final, que depende das condições finais estabelecidas na rede, uma vez que os resultados dependem dos valores aprendidos. Outra desvantagem das redes neurais é o fato de os resultados obtidos não terem uma comprovação, pois todo o conhecimento adquirido pelos neurônios não podem ser representados. Portanto, não é possível comprovar um resultado adquirido através da utilização de redes neurais (MARTINHAGO, 2005).

[1] Os níveis de bia possibilitam que a saída não seja nula mesmo que as entradas sejam.

Em comparação às árvores de decisão, os algoritmos de redes neurais normalmente necessitam de maior força computacional para serem utilizados. Os tempos de treinamento variam de acordo com o número de casos de treinamento, número de pesos na rede e das configurações dos vários parâmetros do algoritmo de aprendizado (SOUSA, 1998). A seguir será apresentado o *perceptron* multi-camadas e o algoritmo de retropropagação.

2.2.1 *Perceptron* Multi-Camadas

A rede *perceptron* multi-camadas (MLP) é formada por múltiplas camadas de neurônios interconectadas, normalmente em forma de *feedward*, onde cada neurônio de uma camada tem conexões diretas aos neurônios da camada seguinte (HAYKIN, 2001). Estas redes possuem poder computacional elevado em relação as que não possuem camadas intermediárias e podem receber dados que não são linearmente separáveis (BRAGA, 2000).

O processamento realizado por cada neurônio neste tipo de rede é definido pela combinação dos processamentos efetuados pelos neurônios da camada anterior que estão conectados a ele (BRAGA, 2000). Esta rede representa uma generalização do *perceptron* de camada única, tendo o seu funcionamento descrito como uma seqüência de *perceptrons* (HAYKIN, 2001).

Os *perceptrons* de múltiplas camadas são aplicados com bastante sucesso na resolução de problemas difíceis, a partir de seu treinamento de forma supervisionada com um algoritmo bastante conhecido chamado retropropagação de erro. Baseado na regra de aprendizagem por correção de erro, este algoritmo é considerado uma generalização de outro algoritmo bastante conhecido chamado de algoritmo do mínimo quadrado médio (HAYKIN, 2001).

2.2.1.1 Algoritmo de Retropropagação

O surgimento da retropropagação se deu devido ao interesse por parte dos pesquisadores na resolução de alguns problemas existentes dentro do treinamento das redes neurais. Após seu surgimento, este algoritmo acabou tornando-se um dos mais populares para este tipo de treinamento, sendo considerado um dos responsáveis pelo ressurgimento do interesse nesta área (ARAÚJO, 2006).

O algoritmo de retropropagação utiliza pares para ajustar os pesos na rede, através de um mecanismo de correção de erros. Seu aprendizado baseia-se na propagação retrógada do erro para níveis anteriores da rede, de acordo com o nível de participação que cada neurônio teve na camada superior (BRAGA, 2000).

O treinamento através deste algoritmo ocorre em duas fases, chamadas de *forward* e *backward*, sendo que em cada uma delas a rede é percorrida em um sentido diferente. Na fase *forward*, um padrão é apresentado à camada de entrada da rede. A atividade resultante percorre a rede, camada por camada, até que uma resposta seja produzida pela camada de saída. Na fase *backward*, é feita a comparação da saída obtida com a saída desejada para este padrão particular. Se o resultado não estiver correto, o erro é calculado, sendo o erro propagado a partir da camada de saída até a camada de entrada, modificando os pesos das conexões das unidades das camadas internas, de acordo com a retropropagação do erro (BRAGA, 2000).

A partir das técnicas estudadas, busca-se uma relação entre teoria e prática. Sendo assim, a seguir serão apresentados estudos de casos nos quais as soluções para os problemas existentes foram encontradas utilizando técnicas de mineração de dados.

3 APLICAÇÕES DE MINERAÇÃO DE DADOS

De acordo com o estudo feito através de artigos que utilizam Mineração de Dados e mais especificamente árvores de decisão, foram selecionados três com o objetivo de exemplificar alguns casos em que o uso de árvores de decisão auxiliou a extração de conhecimento de bases de dados em áreas distintas. Assim, é possível perceber que o uso de árvores de decisão se torna cada vez mais comum em diversas áreas, auxiliando a tomada de decisão por parte de homens de negócios e até mesmo por parte de organizações governamentais. O primeiro artigo utiliza a Mineração de Dados para criar perfis de usuários inadimplentes de empresas de telefonia, o segundo artigo utiliza árvores de decisão na tentativa de controlar o desmatamento da Amazônia na cidade de Capixaba, Acre. Por fim, o último artigo aborda a utilização de Mineração de Dados para predição da prevalência de esquistossomose no Estado de Minas Gerais, Brasil.

3.1 Análise do perfil do usuário de serviços de telefonia utilizando técnicas de mineração de dados (JUNIOR; PEREZ, 2006)

O crescimento exponencial do prejuízo que as operadoras de telecomunicações absorvem devido à inadimplência e utilização ilícita dos seus recursos, fez com que as mesmas procurassem alternativas para diminuir este problema. Com isso, percebeu-se que técnicas de mineração de dados, quando aplicadas neste setor, se tornam um poderoso recurso para identificar o perfil de usuários inadimplentes. Quanto mais rápido forem identificados esses usuários, menor será o prejuízo que a operadora de telefonia terá e, conseqüentemente, mais recursos poderão ser oferecidos aos usuários.

Por outro lado, conhecer o perfil dos bons pagadores através de suas preferências na utilização dos serviços, auxilia as empresas de telecomunicações a realizarem campanhas de

marketing, por exemplo, visando promoções e privilégios que as ajudam a manter e agregar clientes novos.

O objetivo principal deste artigo é apresentar o uso de duas técnicas de mineração de dados: redes neurais e árvores de decisão, a fim de avaliar qual delas se torna mais eficaz para identificar o perfil de um usuário inadimplente.

Como ferramenta de Mineração de Dados para geração de árvores de decisão os autores optaram por utilizar o algoritmo C4.5 do software Sipina, pois o mesmo possui licença para uso educacional, implementa o método de classificação e utiliza árvore de decisão para representar o conhecimento obtido. A figura 3.1 apresenta a interface principal do Sipina.

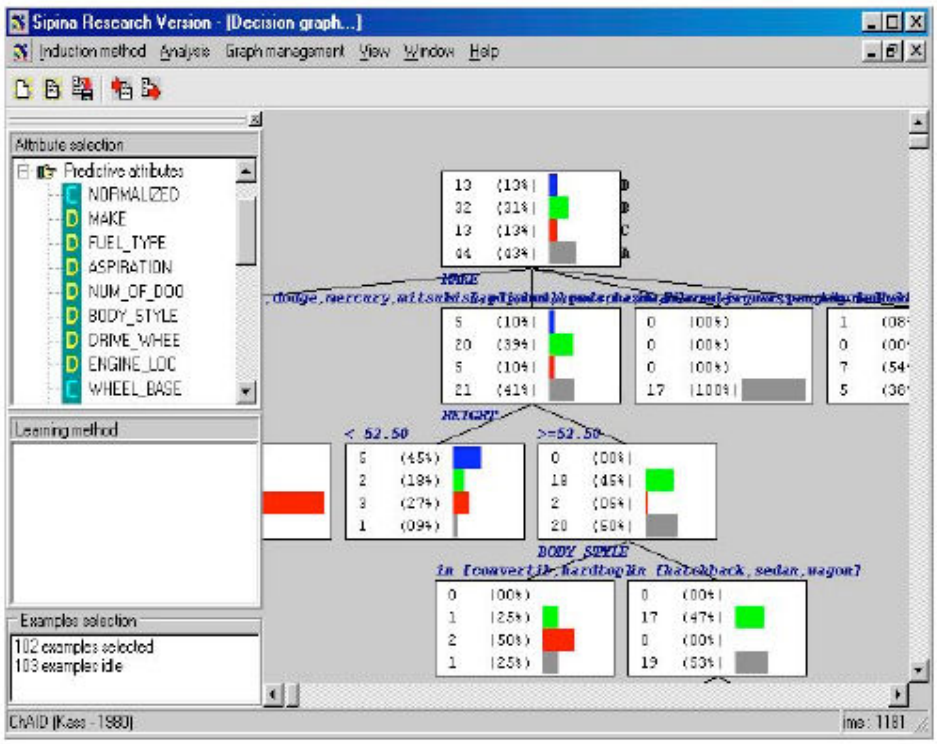

Figura 3.1 – Interface Principal do *Sipina*
Fonte: JUNIOR; PEREZ, 2006, p. 5

Já para a extração do conhecimento utilizando redes neurais foi escolhido o software QwikNet, que simula redes neurais executando vários métodos eficientes para treiná-las e testá-las e tem como característica oferecer uma relação flexível e intuitiva, permitindo

projetar, treinar e testar redes neurais em um ambiente gráfico. A interface do QwikNet é apresentada na figura 3.2.

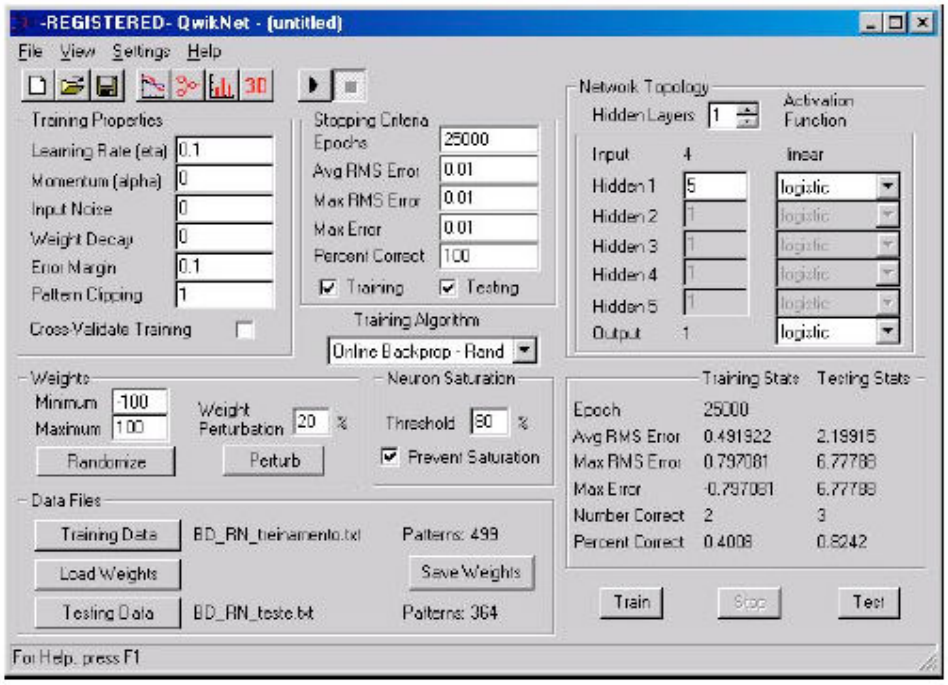

Figura 3.2 – Interface do *QwikNet*
Fonte: JUNIOR; PEREZ, 2006, p. 6

Primeiramente, foi gerado um arquivo texto contendo os dados das chamadas telefônicas presentes no SGBD da companhia. Estes dados foram separados por tabulações e, após a limpeza dos mesmos, sofreram uma codificação, utilizando uma planilha que depois foi salva em formato ".txt", que os enriqueceram e os prepararam para o processo de descoberta de conhecimento. Foram disponibilizados 63.534 registros para testes, com dados de chamadas telefônicas realizadas entre 01 de setembro e 31 de dezembro de 2005, sendo todas estas chamadas realizadas por assinantes inadimplentes de telefones fixos.

Não foi possível a utilização de um número maior de registros devido ao sigilo e também por estes dados serem pontos estratégicos das empresas no combate a inadimplência. Estes registros possuem os seguintes dados: dia da semana que a chamada foi executada, hora inicial da chamada, destino da chamada que identifica o tipo do destino (local, DDD, celular, DDI) e duração das chamadas.

Para a geração da árvore de decisão com o software Sipina, a classe principal foi criada com o atributo “dia da semana” como nodo principal, com o intuito de descobrir os dias da semana em que os usuários fazem mais ligações. Como nodos filhos foram especificados os atributos hora e destino, com a finalidade de descobrir o horário preferido das chamadas e o tempo utilizado nas conversações. Como resultado da aplicação deste algoritmo, pode-se considerar as regras apresentadas na figura 3.3.

Regra 01: Os dias da semana com maior número de chamadas são quarta e quinta feira no horário entre 06:00h – 12:00h:
Quarta feira: 22%
Quinta feira: 31%
Na sexta feira o horário de maior tráfego é entre 18:00h – 24:00h: 21%

Regra 02: Nas segundas feiras o horário entre 12:00h e 18:00h concentra chamadas para serviços especiais: 33%

Regra 03: Nas quartas feiras o horário entre 12:00h e 18:00h concentra chamadas:
para telefone fixo (Local): 23%
para telefone celular (DDD): 22%

Regra 04: Nas quintas feiras o horário entre 12:00h e 18:00h concentram-se chamadas para telefone fixo (DDI): 30%

Regra 05: Nas sextas feiras o horário entre 12:00h e 18:00h concentram-se chamadas:
para telefone celular (Local): 22%
para telefone fixo (DDD): 24%

Figura 3.3 – Regras Geradas pelo *Sipina*
Fonte: JUNIOR; PEREZ, 2006, p. 5

Conforme as regras obtidas, pode-se definir o perfil geral dos usuários, obtendo-se o comportamento generalizado dos inadimplentes. Com essa definição, pode-se estabelecer um parâmetro comparativo, com o qual é realizada a verificação da semelhança do perfil dos usuários individuais com o perfil dos inadimplentes. Analisando os dados de um usuário específico, é feita a comparação dos resultados com as classes pré-determinadas, verificando se determinado usuário se encaixa em um dos perfis já encontrados.

Na realização dos testes utilizando redes neurais, todos os dados foram transformados, conforme realizado no teste com árvores de decisão. A rede neural foi treinada utilizando um arquivo com 499 linhas de dados com informações como dia da semana, horário, destino da chamada e duração, de usuários inadimplentes. Para o teste utilizou-se um

arquivo com 240 linhas com o arquivo de um único usuário para comparar sua semelhança com o perfil de usuário que a rede neural conseguiu aprender.

Foram utilizados quatro neurônios de entrada e um de saída na aplicação do teste, sendo a taxa de aprendizado de 0,1, momentum 0 e critério de parada com 25.000 épocas de treinamento. Dos 364 registros com os quais foram realizados os testes, representando uma pequena amostra do total de registros, três encaixam-se no perfil aprendido pela rede neural. Com isso, considera-se que este indicador é bastante baixo quando aplicado para identificação de perfis de usuários inadimplentes ou fraudulentos.

As redes neurais são utilizadas para aprender com o histórico dos usuários, analisando o comportamento de cada um em diferentes períodos do dia. Realizando uma análise individual, dão condições às empresas de descobrir em tempo real alguma atividade suspeita, interrompendo-a rapidamente colaborando com o aumento da lucratividade da empresa.

Com a técnica de árvores de decisão os dados apresentados representam um padrão de comportamento, contudo foi gerado um número elevado de subdivisões, o que fez com que a leitura do resultado se tornasse um pouco demorada, mas de fácil compreensão. A técnica também permite a geração de regras que definem o padrão, o que facilita a procura de novos padrões quando aplicada em outro volume de dados.

3.2 Avaliação da exatidão do mapeamento da cobertura da terra em Capixaba, Acre, utilizando classificação por árvore de decisão (CARVALHO; FIGUEIREDO, 2006)

Um problema real enfrentado nos dias de hoje é o desmatamento na Amazônia. As instituições brasileiras e internacionais, assim como pesquisadores, organizações não-governamentais e sociedade em geral passaram a ter uma preocupação maior, devido à gravidade deste problema. Diversas experiências e políticas públicas passaram a ser aplicadas na tentativa de reversão deste cenário, entre elas, as que têm como objetivo o monitoramento ambiental da cobertura florestal e que utilizam a aplicação de técnicas de classificação digital e sensoriamento remoto no mapeamento da cobertura da terra.

O objetivo principal deste estudo é realizar uma avaliação, com a maior exatidão possível, do mapeamento da cobertura da terra em Capixaba, utilizando a classificação digital

de imagens de sensoriamento remoto, através de um algoritmo de árvore de decisão. De acordo com dados do IBGE, a área de estudos está localizada no sudeste do Estado do Acre, e corresponde ao município de Capixaba, com a superfície territorial de 1.713 km², equivalendo a 1,1% da área total do estado, conforme representado na figura 3.4.

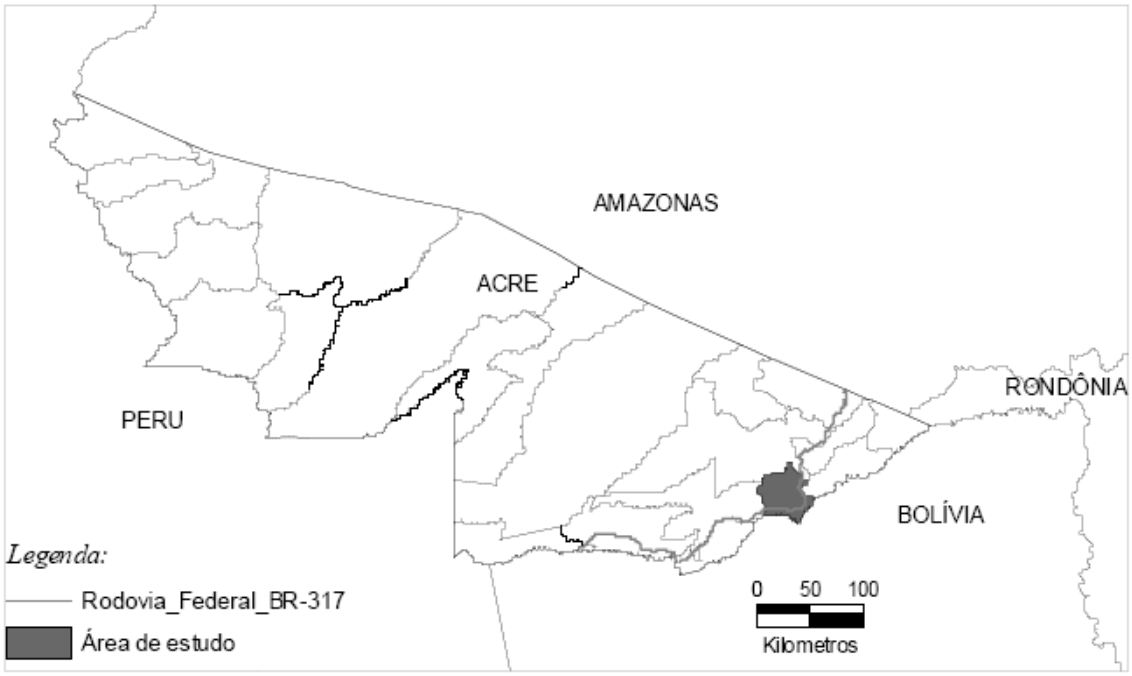

Figura 3.4 – Localização do município de Capixaba, Acre
Fonte: CARVALHO; FIGUEIREDO, 2006, p. 39

Foram utilizadas, para realização deste estudo, imagens multiespectrais, que são imagens de um mesmo objeto, tomadas com diferentes comprimentos de ondas eletromagnéticas, do sensor *Thematic Mapper* (TM) do satélite *Landsat* 5, referentes ao ano de 2003, sendo as imagens derivadas de técnicas de extração de informações. Foram utilizados aplicativos de mineração de dados e de processamento de imagens para a construção da árvore de decisão e geração do mapa temático. Para o mapeamento da cobertura da terra, foi necessária a definição de sete classes temáticas, de acordo com a vegetação e características do relevo das áreas a serem estudas. Estas classes temáticas são as seguintes: Floresta, Capoeira, Pasto alto, Pasto baixo, Solo, Água e Queimada.

Para a classificação pelo algoritmo de árvore de decisão foi utilizado um aplicativo de mineração de dados, gerando assim um conjunto de regras e a árvore de decisão, sendo

posteriormente utilizado um aplicativo de processamento de imagens para classificação digital.

Nesta técnica de classificação, todos os atributos foram organizados em um único arquivo de imagem, sendo considerados os seguintes atributos: bandas 1, 2, 3, 4, 5 e 7 do *Landsat* 5 com valores em número digital; imagens fração solo, sombra, vegetação e de erro geradas pelo modelo linear de mistura espectral e o índice de vegetação.

Para a construção da árvore de decisão foram utilizadas amostras de treinamento e amostras de teste. Cada *pixel* das amostras e seus respectivos valores nas onze imagens do arquivo de dados precisam ser analisados pelo algoritmo de aprendizado de máquina. Na fase de preparação dos dados utilizados para a implementação da mineração de dados, foram organizados três arquivos, denominados arquivo de nomes, arquivo de dados e arquivo de teste.

No arquivo de nomes está o nome das classes temáticas, dos atributos e valores dos atributos. A classe temática do caso, juntamente com a descrição dos casos ou amostras de treinamento, estão presentes no arquivo de dados. No mesmo formato do arquivo de dados, o arquivo de testes foi utilizado na avaliação do erro da árvore de decisão, criada pela mineração de dados. Foi aplicada também, a técnica convencional de classificação digital por meio do algoritmo de máxima verossimilhança e do algoritmo isodata, com o objetivo de avaliar seu desempenho em relação ao algoritmo de árvore de decisão.

Na construção da matriz de erro foram escolhidas as amostras de validação numa imagem de referência estratificada por classes temáticas. Para a realização da estratificação e produção do mapa temático de referência, foi utilizado um algoritmo de análise de agrupamento e migração.

Com a aplicação do algoritmo de árvores de decisão, foi possível gerar estimativas do percentual de área de cada uma das classes temáticas do município. Foi possível identificar também que a cobertura florestal representa 57,94% da área do município, além do percentual ocupado pela classe capoeira, que totaliza 10,56% da área. Pastagens de propriedades rurais e projetos de assentamento representam 29,05% da área de Capixaba, o que representa 50 mil hectares.

Os melhores desempenhos no emprego do algoritmo de árvores de decisão ficaram por conta das classes de floresta e água, com erros de inclusão ou omissão inferiores a 10%, o que demonstra a eficiência da técnica no mapeamento destas classes temáticas. Já nas áreas de capoeira, os erros de classificação ocorreram devido a semelhança das classes de pasto alto e floresta, o que ocasionou confusão na classificação das mesmas.

Porém, os erros mais significativos obtidos através do emprego do algoritmo de árvore de decisão foram verificados nas classes de pasto alto, pasto baixo e capoeira que obtiveram erros de 18,18%, 17,65% e 16,87%, respectivamente. As características da classe pasto baixo e de solo são semelhantes, gerando alguma confusão na classificação das mesmas, porém a exatidão da classe solo foi altamente satisfatória, com índice de 84,62%.

A tabela 3.1 representa a exatidão global obtida pelo classificador de árvore de decisão com os erros de inclusão e omissão por classe de mapeamento do ano de 2003.

Tabela 3.1 – Matriz de Erros Obtida com o Algoritmo de Árvore de Decisão

	Amostras de validação (pixels)						
Classificação	CA	FL	SO	AG	PA	PR	Total	Inclusão
Capoeira (CA)	69	1	0	0	5	0	75	8,0%
Floresta (FL)	11	194	0	0	0	0	205	5,4%
Solo (SO)	0	0	44	0	0	1	45	2,2%
Água (AG)	0	0	0	77	0	0	77	0,0%
Pasto alto (PA)	3	0	0	0	90	17	110	18,2%
Pasto baixo (PB)	0	0	8	0	0	84	92	8,7%
Total	83	195	52	77	95	102	604	
Omissão	16,9%	0,5%	15,4%	0,0%	5,3%	17,6%		
Exatidão global = 92,38%					Kappa = 0,9044			

Fonte: CARVALHO; FIGUEIREDO, 2006, p. 45

O resultado do mapeamento da cobertura da terra em Capixaba com algoritmo de árvore de decisão foi considerado excelente, sendo superior ao do algoritmo de máxima verossimilhança e ainda maior quando comparado com o método de classificação digital não

supervisionada isodata. Os melhores resultados foram obtidos com as classes de floresta e água, com exatidão superior a 94%. Apesar dos erros de classificação em algumas classes, os resultados demonstram que a técnica é altamente eficiente para aplicação de mapeamento de cobertura de terra.

3.3 Uso de árvore de decisão para predição da prevalência de esquistossomose no Estado de Minas Gerais, Brasil (MARTINS et al, 2007)

No Brasil os hospedeiros da esquistossomose são moluscos límnicos do gênero *Biomphalaria*. Devido a problemas de saneamento domiciliar e ambiental e do baixo nível de educação em saúde da população que vive sob risco, a doença tem caráter social e comportamental. Uma vez que a doença é limitada no espaço e tempo por fatores ambientais, tornam-se extremamente importantes o uso de sistemas de informação geográficos (SIG) e o sensoriamento remoto (SR) na identificação dos fatores ambientais, permitindo que recursos sejam alocados nas áreas com mais risco de contaminação, auxiliando o combate a doença.

O objetivo principal deste trabalho é aplicar uma técnica de mineração de dados, utilizando árvores de decisão, para obter a estimativa da prevalência da esquistossomose no Estado de Minas Gerais. Através dos dados obtidos com sensoriamento remoto, derivadas climáticas e sócio-econômicas, pretende-se empregar ferramentas de SIG para dar auxílio no combate à doença e conscientização da população através dos órgãos competentes.

Como área de estudo, foi utilizado o Estado de Minas Gerais, com área de aproximadamente 590.000 km², contendo 853 municípios, com aproximadamente 18 milhões de habitantes.

Para o estudo, foram utilizadas 197 amostras de um total de 853, que corresponde ao total de municípios do Estado, que possuíam dados de prevalência da doença. Com esta amostra, foi gerada a árvore de decisão, sendo possível, através da árvore selecionada, a extrapolação da estimativa da prevalência da esquistossomose para as demais amostras.

Todos os dados da prevalência foram fornecidos pela Secretaria de Vigilância em Saúde e Secretaria de Estado de Saúde de Minas Gerais. Das quarenta e quatro variáveis utilizadas para amostragem da prevalência, vinte e duas são oriundas de dados de sensoriamento remoto, seis são climáticas e dezesseis são sócio-econômicas.

Para a aplicação da técnica de mineração de dados foi utilizado o software *Weka* (*Waikato Environment for Knowledge Analysis*), por ser um software de código aberto e sua licença gratuita. Para a criação da árvore de decisão foi empregado o algoritmo J4.8, que é a versão na linguagem Java do algoritmo C4.5, mencionado no capítulo anterior.

Como em todo processo de descoberta de conhecimento, foi necessária a realização do pré-processamento dos dados. Como o algoritmo de classificação necessita que a variável a ser explicada seja uma variável nominal, foi necessária a transformação dos dados, utilizando uma regra simples no *Excel*, que classificou os dados em quatro categorias: baixa (0 a 5%); média (>5 a 15%); alta (>15 a 25%); e muito alta (>25%). Esta classificação pode ser observada na figura 3.5.

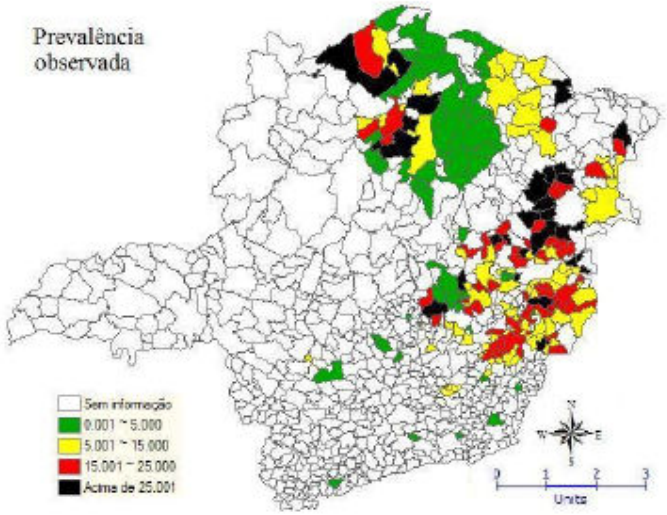

Figura 3.5 – Prevalência da Esquistossomose em 197 Municípios de Minas Gerais
Fonte: MARTINS et al, 2007, p. 2844

Com a utilização de algoritmos de classificação, objetivou-se analisar as diferenças no padrão de comportamento das variáveis em relação à prevalência da doença. Visto que o algoritmo J4.8 gera as regras de decisão e uma matriz de erros, é possível detectar prováveis problemas na classificação e também a comparação entre as classes, conforme mostra a tabela 3.2

Tabela 3.2 – Matriz de Erros Obtida com o Algoritmo J4.8

Classificada / Observada	Baixa	Média	Alta	Muito alta
Baixa (46)	42	3	1	0
Média (73)	17	47	9	0
Alta (51)	5	14	25	7
Muito alta (27)	2	4	8	13

Fonte: MARTINS et al. 2007, p. 2844

É possível observar também que das 197 amostras, 127 são classificadas corretamente, sendo das 70 amostras classificadas incorretamente, 58 classificadas com um erro de classe, 10 com dois erros e somente duas com três erros. Levando em consideração as 46 amostras da classe de prevalência baixa, 91,3% das amostras foram classificadas corretamente. Isto significa que o resultado pode ser considerado satisfatório, visto que os recursos para combate à doença são escassos e com este resultado dificilmente eles serão alocados em áreas com menor prevalência. Com as amostras de prevalência média, 64,4% são classificadas corretamente, já com as amostras de prevalência alta e muito alta, 49% e 48%, respectivamente, são classificadas corretamente.

A árvore de decisão para a prevalência da doença em relação a algumas variáveis preditivas, selecionadas pelo *software Weka* por conterem maior quantidade de informações, é apresentada na figura 3.6.

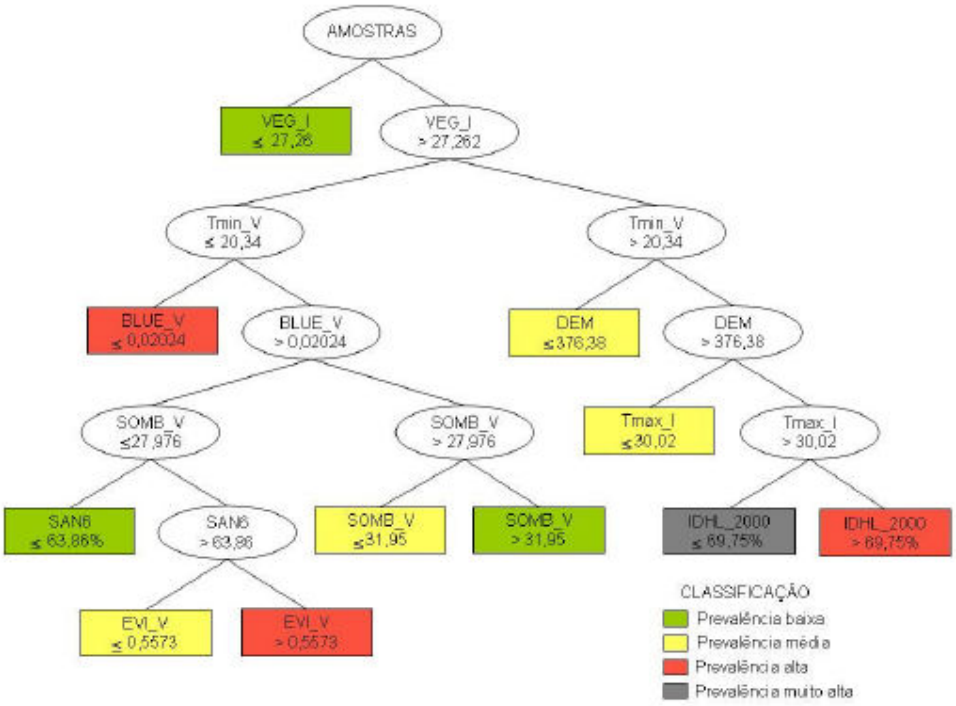

Figura 3.6 – Árvore de Decisão Gerada pelo Algoritmo J4.8
Fonte: MARTINS et al, 2007, p. 2845

Os resultados da utilização de árvore de decisão para verificação da prevalência de esquistossomose podem ser considerados satisfatórios, pois com isso foi possível enxergar os locais com maior índice de contaminação, onde os recursos e a conscientização da população devem ser tratados com maior prioridade. Também é possível verificar que o resultado é coerente, pois o habitat ideal para o caramujo e condições de vida das pessoas são fatores importantes para a existência da doença.

Assim, é possível ter uma melhor percepção da importância da mineração de dados em áreas diferentes, fazendo com que questões sociais e patrimoniais sejam resolvidas de uma forma inteligente e eficaz. O aproveitamento das informações desconhecidas presentes nas bases de dados faz com que conhecimento seja gerado e a utilização das técnicas seja adotada por um número cada vez maior de organizações.

Pretende-se apresentar uma proposta de utilização de técnicas de *data mining* na base de dados de uma academia de musculação, com o intuito de que todo o conhecimento contido neste trabalho seja colocado em prática. Todas as etapas do processo de KDD serão realizadas para que isto seja possível e, através da utilização de algoritmos de classificação, sejam

gerados perfis de alunos a fim de trazer vantagens para a academia através de ações baseadas no conhecimento obtido.

CONCLUSÃO

Pode-se afirmar que muitas vezes os dados presentes nas bases de dados das organizações não são aproveitados da melhor maneira possível. Assim, torna-se fundamental a aplicação de técnicas de mineração de dados nestas bases a fim de transformar informações desconhecidas em conhecimento útil e lucrativo para as empresas. A partir destas técnicas de mineração de dados é possível criar perfis de clientes, fazendo com que ações possam ser tomadas por parte dos homens de negócios na tentativa de aumentar o potencial da empresa.

A aplicação de técnicas de mineração de dados na base da academia de musculação proporcionará um estudo focado no desempenho dos alunos de acordo com sua freqüência e característica da atividade desempenhada na academia. Será possível criar perfis de alunos visando promoções ou ajustes nos serviços prestados, de acordo com as características de cada perfil.

Busca-se também obter dados estatísticos sobre os clientes, além de encontrar alguma relação das atividades dos alunos com a sua freqüência na academia, evolução do peso e medidas, de acordo com faixa etária, sexo, raça, nacionalidade, entre outras informações. As informações adquiridas podem ser aplicadas na organização, podendo ser um diferencial para a academia no que diz respeito ao monitoramento da evolução física dos alunos.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, Leandro Maciel; PADILHA, Thereza Patrícia P.; OLIVEIRA, Fernando Luiz de; PREVIERO, Conceição Aparecida. **Uma Ferramenta para Extração de Padrões**. Revista Eletrônica de Iniciação Científica, v. 3, 2003. Disponível em: <http://www.sbc.org.br/reic/edicoes/2003e4/cientificos/UmaFerramentaParaExtracaoDePadroes.pdf>. Acesso em 20 Ago. 2008.

ARAÚJO, Anderson Viçoso de. **Árvore de Decisão Fuzzy na mineração de imagens do sistema Footscanage.** Curitiba, PR: 2006. Dissertação (Mestrado) – Programa de Pós-Graduação em Informática, Universidade Federal do Paraná, 2006.

BORGES, Helyane Bronoski. **Redução de Dimensionalidade de Atributos em Bases de Dados de Expressão Gênica.** Curitiba, PR: 2006. 123 p. Dissertação (Mestrado) – Programa de Pós Graduação em Informática. Pontifícia Universidade Católica do Paraná, 2006.

BRAGA, Antônio de Pádua; LAUDERMIR, Teresa Bernarda; CARVALHO, André Carlos Ponce de Leon Ferreira. **Redes neurais artificiais: teoria e aplicações.** Livros Técnicos e Científicos Editora S.A., 2000.

CARVALHO, Juliano Varella de. **Reconhecimento de Caracteres Manuscritos Utilizando Regras de Associação.** Campina Grande, PB: 2000. Dissertação (Mestrado) - Centro de Ciências e Tecnologia, Universidade Federal da Paraíba, 2000.

CARVALHO, Luis Alfredo Vidal de. **DataMining : A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração.** Rio de Janeiro: Ciência Moderna, 2005.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery: An overview. In: FAYYAD et al. **Advances in Knowledge Discovery and Data Mining.** G. Cambridge-Mass:AAAI/MIT Press, 1996.

FIGUEIREDO, Symone Maria de Mello; CARVALHO, Luis Marcelo Tavares de. **Avaliação da exatidão do mapeamento da cobertura da terra em Capixaba, Acre utilizando classificação por árvore de decisão.** Cerne, Lavras, v. 12, n. 1, p. 38 – 47, 2006.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Mining: Um Guia Prático: Conceitos, Técnicas, Ferramentas, Orientações e Aplicações.** Rio de Janeiro, RJ: Elsevier, 2005.

GONCHOROSKI, Sidnei Pereira. **Utilização de Técnicas de KDD em um Call Center Ativo.** Novo Hamburgo, RS: 2007. 119 p. Monografia (Bacharelado em Ciência da Computação) – Instituto de Ciências Exatas e Tecnológicas, Centro Universitário Feevale, 2007.

HAYKIN, Simon S. **Redes neurais: princípios e prática.** 2. ed. Porto Alegre, RS: Bookman, 2001. 900 p.

JERONIMO, Paulo Marcelo. **Estudo sobre: Data Mining : Data Warehouse : Cases - Data Warehouse.** Novo Hamburgo, RS: 2001. 73 p. Monografia (Bacharelado em Ciência da Computação) – Instituto de Ciências Exatas e Tecnológicas, Centro Universitário Feevale, 2001.

JUNIOR, Adelir José Schuler; PEREZ, Anderson Luiz Fernandes. **Análise do perfil do usuário de serviços de telefonia utilizando técnicas de mineração de dados.** Revista Eletrônica de Sistemas de Informação, Florianópolis, p. 1 - 8, 01 jun. 2006.

KRANZ, Paulo Henrique. **Business Intelligence: Estudo Aplicado em Cooperativa Médica.** Novo Hamburgo, RS: 2004. 103 p. Monografia (Bacharelado em Ciência da Computação) – Instituto de Ciências Exatas e Tecnológicas, Centro Universitário Feevale, 2004.

LANDWEHR, Niels; HALL, Mark; FRANK, Eibe. **Logistic model trees.** Proceedings of the 14th European Conference on Machine Learning, p. 241-252, 2003.

MARTINHAGO, Sergio. **Descoberta de Conhecimento sobre o processo seletivo da UFPR.** Curitiba, PR: 2005. 114 p. Dissertação (Mestrado) – Departamento de Matemática, Universidade Federal do Paraná, 2005.

MARTINS, Flávia de Toledo; DUTRA, Luciano Vieira; FREITAS, Corina da Costa; FONSECA, Fernanda Rodrigues; GUIMARAES, Ricardo José de Paula Souza e; MOURA, Ana Clara Mourão; SCHOLTE, Ronaldo Guilherme Carvalho; AMARAL, Ronaldo Santos; DRUMMOND, Sandra Costa; FREITAS, Charles R.; CARVALHO, Omar dos Santos. **Uso de árvore de decisão para predição da prevalência de esquistossomose no Estado de Minas Gerais, Brasil.** In: XIII Simpósio Brasileiro de Sensoriamento Remoto, 2007, Florianópolis, SC. XIII SBSR Anais, 2007. p. 2841-2848.

OLIVEIRA, Ivana Corrêa de. **Aplicação de Data Mining na Busca de um Modelo de Prevenção da Mortalidade Infantil.** Florianópolis, SC: 2001. Dissertação (Mestrado) – Engenharia e Sistemas, Universidade Federal de Santa Catarina, 2001.

PASSINI, Sílvia Regina Reginato; TOLEDO, Carlos Miguel Tobar. **Mineração de Dados para Detecção de Fraudes em Ligações de Água.** XI SEMINCO – SEMINÁRIO DE COMPUTAÇÃO, 2002, Blumenau, SC. Anais do XI Seminco. Blumenau, SC: s.n., 2002. p. 229- 242.

REZENDE, Solange Oliveira. **Mineração de Dados.** In: XXV Congresso da Sociedade Brasileira de Computação, 2005. Anais do XXV Congresso da Sociedade Brasileira de Computação. São Leopoldo: SBC, 2005. p. 397-433.

SANTOS, Daiana Pereira dos. **Mineração em Notas Fiscais de entrada de uma empresa calçadista.** Novo Hamburgo, RS: 2008. 93 p. Monografia (Bacharelado em Ciência da

Computação) – Instituto de Ciências Exatas e Tecnológicas, Centro Universitário Feevale, 2008.

SOUSA, Mauro Sérgio Ribeiro de. **Mineração de Dados: Uma Implementação Fortemente Acoplada a um Sistema Gerenciador de Banco de Dados Paralelo.** Rio de Janeiro, RJ: 1998. 75 p. Dissertação (Mestrado) – Programa de Pós Graduação de Engenharia. Universidade Federal do Rio de Janeiro, 1998.