Instituto Unificado de Ensino Superior Objetivo
Ciência da Computação

Mineração de Dados

Trabalho de Conclusão de Curso

Goiânia-GO
2006

Carla Moema Moreira Duarte

MINERAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado ao Curso de Ciência da Computação do Instituto Unificado de Ensino Superior – SOES como pré-requisito para obtenção do grau de bacharel em Ciência da Computação

Orientador:
Cláudio Pinho

Goiânia-GO
2006

Carla Moema Moreira Duarte

MINERAÇÃO DE DADOS

Trabalho de Conclusão de Curso apresentado ao Curso de Ciência da Computação do Instituto Unificado de Ensino Superior – SOES como pré-requisito para obtenção do grau de bacharel em Ciência da Computação

Orientador:
Cláudio Pinho

Banca Examinadora

_____

Professor Edmar Gonçalves Vieira

_____

Professor André Luiz Barreto Esperidião

_____

Professor Cristiano Penido

Goiânia-GO
2006

Dedico este trabalho aos meus pais e minhas irmãs por todo amor, apoio e paciência.

AGRADECIMENTOS

A Deus por estar sempre zelando por mim.

A minha família pelo apoio constante.

Aos professores Cláudio Pinho Resende, Davy Cestari Vinaud e Edmar Gonçalves Vieira pela orientação e atenção que tiveram comigo no desenvolvimento deste trabalho.

“A educação faz um povo fácil de ser liderado, mas difícil de ser dirigido; fácil de ser governado, mas impossível de ser escravizado.”

Henry Peter Brougham

RESUMO

A Mineração de Dados é um processo de natureza interativa responsável por encontrar padrões em grandes conjuntos de dados, com o intuito de extrair conhecimento válido, útil e inovador a partir destes. Surge com o objetivo de auxiliar na tomada de decisões em ambientes corporativos. Considerando a grande quantidade de informações armazenadas em um banco de dados, grandes corporações têm investido cada vez mais na formação e na busca por profissionais dessa área que é vista como parte de um processo maior, chamado KDD - *Knowledge Discovery in Databases* (Descoberta do Conhecimento em Bancos de Dados) sendo de fundamental importância a participação humana no processo. Esta pesquisa tem como objetivo fornecer uma introdução sobre o assunto, apresentando métodos, suas pré-condições e efeitos, técnicas e tarefas aplicáveis a cada situação a depender da base de dados que se quer extrair o conhecimento. Algumas técnicas são demonstradas com a utilização da ferramenta Weka. São apresentadas também, ao longo deste trabalho, algumas aplicações de Data Mining para ilustrar o potencial desta tecnologia.

SUMÁRIO

1. INTRODUÇÃO 11

1.1 MOTIVAÇÃO E OBJETIVOS 11
1.2 ORGANIZAÇÃO DO TRABALHO 11

2. DESCOBERTA DO CONHECIMENTO EM BASES DE DADOS 13

3. MINERAÇÃO DE DADOS 17

3.1. MÉTODOS DE MINERAÇÃO DE DADOS 20
3.1.2 - Métodos Baseados em Instâncias (Emmanuel Passos & Ronaldo Goldschmidt , 2005) 21
3.1.3 – Métodos Estatísticos 22
3.2 TAREFAS DE MINERAÇÃO DE DADOS 26
3.2.1 - Classificação 27
3.2.2 - Clusterização ou Agrupamento 29
3.2.3 - Associação 31
3.2.4 - Estimativa 33

4. POSSÍVEIS ÁREAS DE APLICAÇÃO DE DATA MINING 35

5. UTILIZANDO A FERRAMENTA WEKA 37

5.1 - EXEMPLO DE APLICAÇÃO DE ÁRVORES DE CLASSIFICAÇÃO 37
5.2 - EXEMPLO DE APLICAÇÃO DE REGRAS DE ASSOCIAÇÃO 40

6. CONCLUSÃO 43

REFERÊNCIAS 44

LISTA DE FIGURAS

Figura 1 - Etapas Operacionais do processo de KDD.....15
Figura 3 - Conjunto contendo dados sobre clientes que receberam crédito 22
Figura 4 - Resultado do K-NN.....22
Figura 5 - Fluxograma K-Means 25
Figura 6 – Árvore de Classificação.....28
Figura 7 - Aplicativo Weka em execução.....37
Figura 8 – Weather 38
Figura 9 – Atributos.....39
Figura 10 - Árvore de Decisão.....39
Figura 11 - *weather.nominal.arff*.....40
Figura 12 - Parâmetros do *Apriori*.....41
Figura 13 - Melhores regras geradas.....41

LISTA DE TABELAS

Tabela 1 - Back-Propagation e C4.5s 20
Tabela 2 – JOGA-TÊNIS.....23
Tabela 3 – Informações sobre alguns indivíduos.....28
Tabela 4 – Informações sobre consumidores.....31
Tabela 5 – Clusters criados com a tarefa 31
Tabela 6 – Dados de jogadores de basquete 33
Tabela 7 – Dados do novo jogador e altura estimada 34

1. INTRODUÇÃO

1.1 Motivação e Objetivos

Surge a tríade “dado, informação e conhecimento”. O computador, em essência, serve para transformar dados em informações; o conhecimento é o uso inteligente da informação é a informação utilizada na prática (REINALDO VIANA – 2005).

Os bancos relacionais nos permitem a extração de diversas informações usando consultas SQL. Basta que as questões sejam definidas, sendo assim, as informações extraídas são na verdade respostas a problemas existentes. A Mineração de Dados (Data Mining) consiste em descobrir informações totalmente novas tendo em mãos grandes quantidades de dados. Podendo estar na forma de Banco de Dados Relacional, arquivos texto, Data Mart, entre outras. Possuindo uma vasta aplicação nos mais diferentes segmentos, tanto acadêmicos como corporativos, além de uma série de desafios relevantes que podem motivar trabalhos científicos. O avanço tecnológico e a oferta de ferramentas não dispensam o especialista do domínio minerado. A experiência profissional, a convivência com os processos e a leitura dos padrões descobertos são atributos que propiciam ao minerador amplas chances de sucesso no processo.

O presente Trabalho de Conclusão de Curso tem como intuito explorar as técnicas de Mineração de Dados, cobrindo os aspectos práticos e fornecendo a base teórica para um melhor entendimento do assunto. Os fundamentos serão apresentados, bem como diferentes áreas de pesquisa e aplicação desta tecnologia. Visando um enfoque prático e aplicado, atividades de mineração serão descritas aqui e auxiliarão na fixação dos conceitos apresentados, bem como numa melhor concepção do potencial desta desafiadora e recente área de pesquisa. O texto mescla uma abordagem conceitual e formal com linguagem acessível, recomendada a todos os tipos de leitores.

1.2 Organização do Trabalho

O Capítulo 2 complementa a visão geral da área de KDD com conceitos básicos necessários aos capítulos posteriores.

O Capítulo 3 apresenta diversos métodos de Mineração de Dados assim como algumas das principais tarefas que são também exemplificadas.

No Capítulo 4 são apresentadas possíveis áreas de aplicação das técnicas de Mineração de Dados.

O Capítulo 5 traz exemplos práticos de utilização da ferramenta Weka, usando Árvores de Classificação e Regras de Associação.

O Capítulo 6 faz um resumo dos requisitos para se obter sucesso no processo de Mineração de Dados.

2. DESCOBERTA DO CONHECIMENTO EM BASES DE DADOS

Um Banco de Dados – BD representa uma coleção de dados que possui algum significado e objetiva atender a um conjunto de usuários. Por exemplo, um catálogo telefônico pode ser considerado um BD. Sendo assim, um BD não necessariamente está informatizado. Quando resolvemos informatizar um BD, utilizamos um programa especial para realizar essa tarefa. Tal programa é denominado SGBD1 – Sistema Gerenciador de Banco de Dados. Podemos citar como exemplos de SGBDs: SQL Server, Oracle, Firebird, MySQL, Interbase, entre outros. Estes programas em geral são chamados SGBDs relacionais. Em um SGBD relacional, enxergamos os dados armazenados em uma estrutura chamada tabela. Neste modelo, as tabelas de um BD são relacionadas, permitindo assim que possamos recuperar informações envolvendo várias delas.

Os constantes avanços na área da Tecnologia da Informação têm viabilizado o armazenamento de grandes e múltiplas bases de dados. Tecnologias como a internet, sistemas gerenciadores de banco de dados, leitores de códigos de barras, dispositivos de memória secundária de maior capacidade de armazenamento e de menor custo e sistemas de informação em geral são exemplos de recursos que tem facilitado a criação de inúmeras bases de dados de diversas naturezas.

Em contra partida, a análise de grandes quantidades de dados pelo homem é inviável sem o auxílio de ferramentas computacionais adequadas. Com isso, tornou-se indispensável o desenvolvimento de ferramentas que auxiliam, de forma automatizada e inteligente, a tarefa de analisar, interpretar e relacionar esses dados para que se possa desenvolver e selecionar estratégias de ação em cada conjunto de aplicação.

A maioria dos sistemas de informação opera sobre bancos de dados chamados transacionais2. Esses bancos de dados contêm informações detalhadas que permitem às empresas acompanhar e controlar processos operacionais. Gerentes e executivos necessitam de recursos computacionais que forneçam subsídios para apoio ao processo decisório, sobretudo nos níveis tático e estratégico das empresas. Nas bases de dados tradicionais, os dados encontram-se voltados para a representação de detalhes operacionais corporativos. Em

[1] Software que possui recursos capazes de manipular as informações do banco de dados e interagir com o usuário. Um BD relacional é composto por um único tipo de construção: Tabela, composta por linhas (tuplas) e colunas (atributos); a ligação entre linhas de diferentes tabelas é feita através do uso de valores de atributos.

[2] As bases de dados transacionais contêm todas as informações operacionais da empresa, tais como: cadastros de clientes, produtos, fornecedores. São todas as informações históricas armazenadas ao longo do tempo de vida da empresa, geralmente imensos volumes de dados e muitas vezes espalhados em várias fontes de dados.

Data Warehouses, os dados encontram-se consolidados de forma a viabilizar consultas, descoberta de tendências e análises estratégicas a partir dos dados.

Para atender esse novo contexto, surge uma nova área denominada Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases - KDD). A Mineração de Dados é apenas uma fase do processo de KDD. Em geral, o processo de KDD pressupõe que os dados sejam organizados em uma única estrutura tabular bidimensional contendo casos e características dos problemas a ser analisado. É importante destacar que o processo não requer que os dados a serem analisados pertençam a Data Warehouses. No entanto, o tratamento e a consolidação dos dados necessários à estruturação e cara neste tipo de ambiente é extremamente útil e desejável ao processo de KDD.

Um Data Warehouse é um conjunto de dados baseado em assuntos, integrado, não-volátil, variável em relação ao tempo, e destinado a auxiliar em decisões de negócios. A orientação a assunto, aliada ao aspecto de integração, permite reunir dados corporativos em um mesmo ambiente de forma a consolidar e apresentar informações sobre um determinado tema.

Para compreender o processo de KDD é preciso inicialmente, destacar as diferenças e a hierarquia entre dado, informação e conhecimento. Os dados podem ser interpretados como itens elementares, captados e armazenados por recursos da Tecnologia da Informação. As informações representam os dados processados, com significados e contextos bem definidos. Diversos recursos da Tecnologia da Informação são utilizados para facilmente processar dados e obter informações. O conhecimento é o padrão ou conjunto de padrões cuja formulação pode envolver e relacionar dados e informações.

De acordo com Ronaldo Goldschmidt & Emmanuel Passos (2005) o termo KDD foi formalizado em 1989 em referencia ao amplo conceito de procurar conhecimento a partir de bases de dados. Fayyad et al. (apud GOLDSCHIMIT, 2005): “KDD 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”.

O processo de KDD é caracterizado como um processo composto por várias etapas operacionais. Segue abaixo, na Figura 1 uma representação resumida das etapas executadas. A etapa pré-processamento compreende as funções relacionadas à coleta, organização e ao tratamento dos dados. Esta mesma etapa tem como objetivo a preparação dos dados para os algoritmos da etapa seguinte, a Mineração de Dados. Durante a Mineração de Dados é realizada a busca real por conhecimentos úteis no conjunto da aplicação KDD. A etapa pós-

processamento abrange o tratamento do conhecimento obtido e tem como objetivo, viabilizar a avaliação da utilidade do conhecimento descoberto (Fayyad et al., 1996).

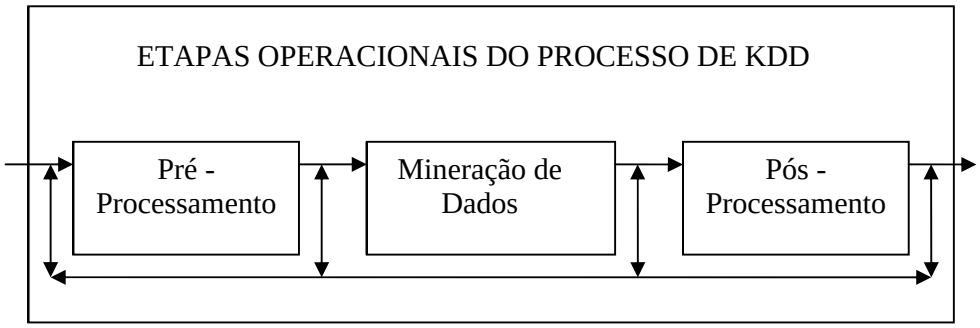

Figura 1 - Etapas Operacionais do processo de KDD

A complexidade do processo de KDD está na dificuldade em perceber e interpretar adequadamente inúmeros fatos observáveis durante o processo e na dificuldade em conjugar dinamicamente tais interpretações de forma a decidir quais ações devem ser realizadas em cada caso (Goldschmidt, 2003). Cabe ao analista humano a tarefa de orientar a execução do processo de KDD.

O uso do termo interativo na descrição do processo KDD se deve ao fato de que o homem é responsável pela atuação direta no controle do processo. Utilizando recursos computacionais disponíveis em função da análise e da interpretação dos fatos observados e resultados obtidos ao longo do processo. Enquanto o termo iterativo sugere a possibilidade de tarefas serem repetidas parcial ou completamente durante o processo na busca de resultados suficientes por meio de refinamentos sucessivos.

A expressão “padrão válido” indica que o conhecimento deve ser verdadeiro e adequado ao contexto da aplicação de KDD. Enquanto que a expressão “padrão novo” deve acrescentar novos conhecimentos aos conhecimentos existentes.

Um conhecimento útil é aquele que pode ser utilizado em benefício da aplicação em questão.

A Descoberta de Conhecimento em Bases de Dados é uma matéria multidisciplinar e, historicamente, origina-se de diversas áreas, dentre as quais podem ser destacadas (Goldschmidt, 2006):

* Estatística
* Inteligência Computacional e Aprendizado de Máquina
* Reconhecimento de Padrões
* Banco de Dados

A ação do usuário no processo de KDD é de grande importância devendo ser especializada essa participação humana. Cabe ao analista humano a tarefa de orientar a execução do processo de KDD. Diante de cada cenário, o homem utiliza sua experiência anterior, seus conhecimentos para interpretar e combinar subjetivamente os fatos de forma a decidir qual a estratégia a ser adotada (Fayyad et al., 1996b; Wirth et al., 1997)

No entanto, a formação de especialistas em KDD constitui-se em uma tarefa complexa, longa e exaustiva, pois requer não somente uma fundamentação teórica sobre a área, mas também a participação destes em inúmeras experiências práticas reais (Goldschmidt, 2005).

3. MINERAÇÃO DE DADOS

O foco da Mineração de Dados é como transformar dados armazenados, em conhecimento, expresso em termos de formalismos de representação, tal como regras e relações entre dados. Existe conhecimento que pode ser extraído diretamente de dados sem o uso de qualquer técnica, entretanto, existe também muito conhecimento que está de certa forma "embutido" na Base de Dados, na forma de relações existentes entre itens de dados que, para ser extraído, é necessário o desenvolvimento de técnicas especiais.

É a principal etapa do processo de KDD, na qual ocorre a busca efetiva por conhecimentos novos e úteis a partir dos dados (Goldschmidt, 2005). É comum encontrarmos autores que se referem à Mineração de Dados e ao Processo de KDD como se fossem sinônimos, já que é nesta fase em que são extraídos os padrões e regras dos dados, mas o processo é maior envolvem outras etapas importantes. A execução desta etapa compreende a aplicação de algoritmos sobre os dados procurando abstrair conhecimento. Estes algoritmos são baseados em técnicas que procuram, de acordo com certos paradigmas, explorar os dados de forma a produzir modelos de conhecimento. A forma de representação do conhecimento em um modelo depende diretamente do algoritmo de Mineração de Dados utilizado.

Todo *conjunto de dados* (visto no Capítulo 2) corresponde a uma *base de fatos* ocorridos que devem ser interpretados como um *conjunto de pontos* em um hiperespaço de dimensão K. A dimensão da base de fatos é determinada pelo número de atributos do conjunto de dados em análise. A Figura 2 mostra um exemplo no contexto da análise de crédito na qual, três informações estão representadas em um plano cartesiano. Os eixos correspondem aos atributos *Renda e Despesa*. Cada ponto representa um caso. O símbolo associado a cada caso ("círculo" ou "xis") fornece a terceira informação, que corresponde ao comportamento do cliente quanto ao pagamento do credito concedido.

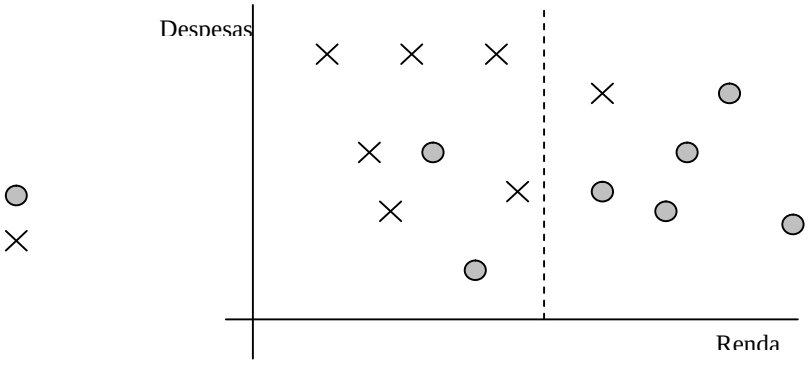

Figura 2 - Exemplo de uma aplicação com "hiperespaço" de dimensão 2

Todo processo de KDD deve ser norteado por objetivos. Estes objetivos compreendem a definição da *tarefa de* KDD a ser executada e da expectativa que os conhecedores do domínio da aplicação tenham com relação ao modelo de conhecimento a ser gerado. A partir dessas definições, o especialista tem condições de esquematizar que *tipos de padrões* devem ser extraídos a partir dos dados.

Imaginemos que a financeira responsável pelos dados deseja obter um modelo de conhecimento que preveja o comportamento de futuros clientes quanto ao pagamento de suas dívidas com uma taxa máxima tolerável de erro de 5%. Esta intenção aliada à base de dados disponível nos conduz à *tarefa de classificação* dos clientes. Esta tarefa consiste em gerar um modelo de conhecimento, a partir do histórico de casos disponível, que consiga, a partir dos dados de novos clientes, prever em qual classe de comportamento o cliente deverá se enquadrar.

O tipo de padrão desejado é uma função que separe a classe dos negligentes da classe dos não-negligentes. São diversas as possibilidades de *hipóteses*, com o objetivo de separar os dois conjuntos, as que interessam à financeira são aquelas que conduzem a uma taxa de erro de no máximo 5% dos casos.

O conceito de Medida de Interesse é importante ao processo de KDD por dois motivos:

* Podem ser usadas após a etapa de Mineração de Dados (etapa de pós-processamento) a fim de ordenar ou filtrar os padrões descobertos de acordo com o grau de interesse associado a estes padrões.

* Podem ser usadas para guiar ou restringir o espaço de busca, melhorando a eficiência da busca ao eliminar conjuntos de padrões que não satisfaçam a condições predeterminadas.

Há dois tipos de Medidas de Interesse que podem ser associadas aos modelos de conhecimento em Mineração de Dados: *objetivas e subjetivas.*

As *Objetivas* são baseadas na estrutura dos padrões descobertos e nas estatísticas a eles relacionados. No caso do exemplo anterior, a taxa de erro é uma ilustração de *medida de interesse objetiva*.

As *Subjetivas* são em crenças que os especialistas no domínio da aplicação tem com relação aos dados e aos modelos de conhecimento gerados: Surpresas, contradições, ou ainda alternativas de ações estratégicas. A avaliação envolvendo este tipo de medida depende,

muitas vezes, da visualização e da interpretação dos resultados obtidos, normalmente realizadas na etapa de pós-processamento (Emmanuel Passos & Goldschmidt, 2005).

Outro fator que influencia na escolha dos algoritmos de Mineração de Dados a serem utilizados em cada problema diz respeito aos tipos de variáveis envolvidas. Alguns têm restrições quanto aos tipos de variáveis existentes no conjunto de dados.

Um conceito comum e bastante usado em Mineração de Dados é a noção de *similaridade*. O conjunto de dados pode ser interpretado como um conjunto de pontos em um espaço k-dimensional, a similaridade entre dois pontos é representada pela *distância* entre estes pontos. Quanto menor a distância maior a similaridade. Entre os algoritmos de Mineração de Dados que utilizam o conceito de distância entre os registros do Banco de Dados, destaca-se o K-NN (*K Nearest Neighbors – K* vizinhos mais próximos).

Outro conceito importante envolvido no processo de KDD e mais especificamente na etapa da Mineração de Dados refere-se à capacidade que determinados algoritmos têm de aprender a partir de exemplos. Esses algoritmos aprendem os relacionamentos eventualmente existentes entre os dados, retratando o resultado deste aprendizado nos modelos de conhecimento gerados. As abordagens são: *aprendizado supervisionado* e *não-supervisionado.*

O *aprendizado supervisionado*[3] compreende a abstração de um modelo de conhecimento a partir dos dados apresentados na forma de pares ordenados (entrada, saída desejada). O *Back-Propagation* e C4.5 são exemplos de algoritmos que utilizam esta abordagem. Algoritmos desse tipo necessitam de um conjunto de dados de treinamento e um de testes. O modelo de conhecimento é extraído a partir do conjunto de treinamento e avaliado a partir do conjunto de testes.

No *aprendizado não-supervisionado* [4]não existe informação da saída desejada. Os algoritmos partem dos dados, procurando estabelecer relacionamentos entre eles. O K-Means e o Apriori são exemplos de algoritmos que usam esta abordagem.

Há ainda diversos outros tipos de algoritmos de Mineração de Dados, alguns necessitam de conhecimento prévio sobre tecnologias como Redes Neurais, Lógica Nebulosa e Algoritmos Genéticos.

---

[3] Neste tipo, a rede neural recebe um conjunto de entradas padronizado e seus correspondentes padrões de saída, onde ocorrem ajustes nos pesos sinápticos até que o erro entre os padrões de saída gerados pela rede tenha um valor desejado;

[4] A rede neural trabalha os dados de forma a determinar algumas propriedades do conjunto de dados. A partir destas propriedades é que o aprendizado é constituído;

3.1. Métodos de Mineração de Dados

Cada método requer diferentes necessidades de pré-processamento Morik (apud GOLDSCHIMIT, 2005). Estas variam em função do aspecto extensional da base de dados em que o método será utilizado. Os métodos de KDD, sendo os métodos de Mineração de Dados um caso particular, podem ser considerados *operadores* definidos a partir de *precondições e efeitos. U*ma *precondição* é um *predicado* que estabelece um requisito que deve ser cumprido antes da execução do método. Um *efeito* também é um *predicado* que descreve uma situação gerada após a aplicação do método. Um plano de ações de KDD válido é toda seqüência de métodos onde as *precondições* para execução de cada um dos métodos da seqüência sejam devidamente atendidas.

	Descritores	**Métodos**	
		Back-Propagation	**C4.5**
Precondições	Atributos Categóricos	Não	Não importa
	Atributos Quantitativos	Sim	Não importa
	Dados Não Normalizados	Não	Não importa
	Valores Ausentes	Não	Não
Efeitos	Transparência	Não	Sim
	Representação do Conhecimento	Matriz de Pesos	Regras de Produção

Fonte: Emmanuel Passos & Goldschmidt (2005)

Tabela 1 - Back-Propagation e C4.5

3.1.1 - Métodos Baseados em Redes Neurais

Classificação, Regressão[5], Previsão de Séries Temporais[6] e Clusterização são exemplos de tarefas de Mineração de Dados que podem ser implementadas por métodos de Redes Neurais. Alguns modelos de Redes Neurais podem ser aplicados em mais de um tipo de tarefa de Mineração. A topologia da rede neural varia em função do problema e da representação adotada para os dados. Em geral, a camada de entrada do modelo neural recebe

[5] A tarefa de Regressão compreende a busca por funções, lineares ou não, que mapeiem os registros de um banco de dados em valores reais. Está restrita a atributos numéricos e também é conhecida como Estimativa

[6] É um conjunto de observações de um fenômeno ordenadas no tempo. Exemplos: o consumo mensal de energia elétrica de uma casa, registrado durante um ano ou as vendas diárias de um produto no decorrer de um mês. É o processo de identificação das características, dos padrões e das propriedades importantes da série, utilizados para descrever em termos gerais o seu fenômeno gerador.

os dados pré-processados de cada registro de um banco de dados. A rede processa esses dados produzindo uma saída cuja natureza também varia em função da aplicação.

Em Redes Neurais com aprendizado supervisionado, a entrada corresponde aos atributos preditivos enquanto a saída do modelo corresponde ao atributo objetivo do problema. O algoritmo de aprendizado pode estimar o erro, ou distancia, entre a saída produzida pela rede e a saída desejada. Em função do erro calculado, o algoritmo ajusta os pesos das conexões da rede a fim de tornar a saída real tão próxima quanto seja possível da saída desejada. Modelos neurais deste tipo são muito úteis em geral para reconhecimento de padrões e, em particular, para tarefas de Mineração de Dados que envolvam predição.

3.1.2 - Métodos Baseados em Instâncias (Emmanuel Passos & Ronaldo Goldschmidt , 2005)

A expressão “método baseado em instância” indica que o método, ao processar um novo registro, leva em consideração as instancias ou os registros existentes na base de dados.

O método K-NN (K vizinhos mais próximos) é um dos principais métodos baseados em instâncias, sendo muito utilizado em aplicações envolvendo a tarefa de classificação.

Considerando uma base de dados (base de referencia) de um problema envolvendo a tarefa de classificação (que contém um atributo cujos valores são rótulos de classes predefinidas) e cada novo registro a ser classificado (registro da base de teste), os seguintes passos são executados:

* Cálculo da distância do novo registro a cada um dos registros existentes na base de referência.

* Identificação dos k registros da base de referencia que apresentaram menor distância em relação ao novo registro (mais similares).

* Apuração da classe mais freqüente entre os k registros identificados no passo anterior.

* Comparação da classe apurada com a classe real, computando erro ou acerto do algoritmo. Este último passo só deve ser utilizado quando as classes dos novos registros são conhecidas e deseja-se avaliar o desempenho do método K-NN na base de dados em questão.

Considere o exemplo no contexto da análise de crédito, cuja base de dados de referência encontra-se representada na Figura 3. Este conjunto está dividido em duas classes: os negligentes representados com um “x” e os não negligentes, representados por um “●”. Deseja-se avaliar a possibilidade de concessão de crédito a novas solicitações.

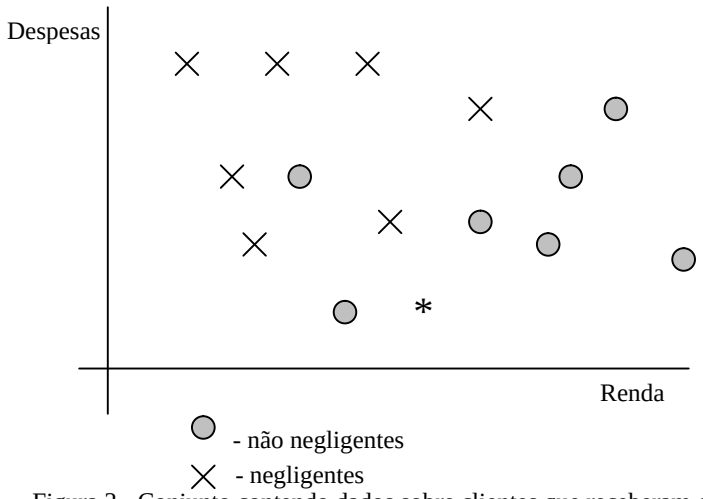

Figura 3 - Conjunto contendo dados sobre clientes que receberam crédito

Apresentando-se um novo registro, representado por “*”, calcula-se a distancia entre o novo registro e todos os registros existentes na base de dados de referencia. Assumindo que o numero de k de vizinhos mais próximos seja três, somente os três registros com menor distancia ao novo registro são considerados.

Desta forma, avaliando os resultados na Figura 4, observa-se que a classe com maior ocorrência dentro da área delimitada pelo algoritmo K-NN foi “cliente não negligente”. Ou seja, pela aplicação do algoritmo K-NN no exemplo apresentado, o crédito seria concedido ao solicitante.

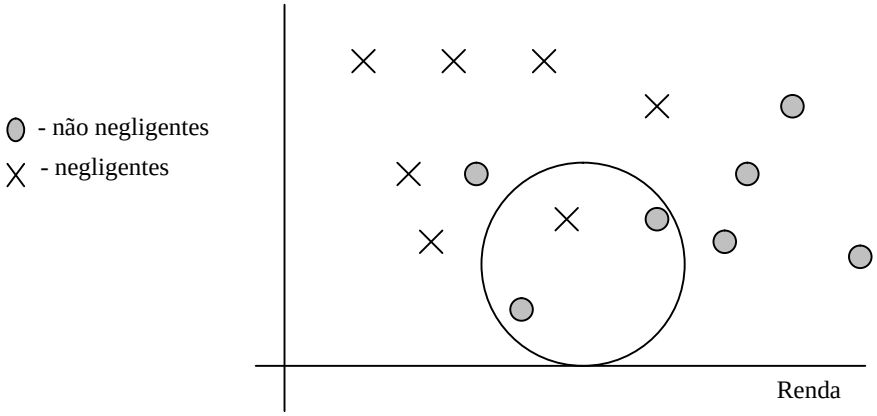

Figura 4 - Resultado do K-NN

3.1.3 – Métodos Estatísticos

Há algoritmos de Mineração de Dados baseados nos princípios da estatística, entre eles o *Classificador Bayesiano Ingênuo* e o *K-Means* (K-meios). Fundamentado na *Teoria de*

Bayes[7], o Classificador Bayesiano Ingênuo está relacionado ao cálculo de probabilidade condicionais. Chamado simples ou ingênuo por considerar que o efeito do valor de um atributo sobre uma determinada classe é independente dos valores dos outros atributos, simplificando sensivelmente as computações envolvidas, possui desempenho e precisão comparável aos classificadores que usam árvores de decisão ou redes neurais.

Exemplo (Custódio Gouvêa – 2005)

Seja o conjunto de treinamento apresentado na Tabela JOGA-TÊNIS e a amostra desconhecida X(TEMPO = Sol, TEMPERATURA = Moderada, UMIDADE = Normal e VENTO = Fraco).

Para classificá-la, deve-se maximizar $P(X \mid Ci)P(Ci)$, para $i = 1, 2$, onde $C1$ corresponde a JOGA-TÊNIS = Sim e $C2$ a JOGA-TÊNIS = Não.

	Atributos				Classe
DIA	TEMPO	TEMPERATURA	UMIDADE	VENTO	JOGA-TÊNIS
1	Sol	Quente	Alta	Fraco	Não
2	Sol	Quente	Alta	Forte	Não
3	Nublado	Quente	Alta	Fraco	Sim
4	Chuva	Moderado	Alta	Fraco	Sim
5	Chuva	Frio	Normal	Fraco	Sim
6	Chuva	Frio	Normal	Forte	Não
7	Nublado	Frio	Normal	Forte	Sim
8	Sol	Moderado	Alta	Fraco	Não
9	Sol	Frio	Normal	Fraco	Sim
10	Chuva	Moderado	Normal	Fraco	Sim
11	Sol	Moderado	Normal	Forte	Sim
12	Nublado	Moderado	Alta	Forte	Sim
13	Nublado	Quente	Normal	Fraco	Sim
14	Chuva	Moderado	Alta	Forte	Não

Tabela 2 – JOGA-TÊNIS

Do conjunto de treinamento, pode-se obter:

• Número total de amostras: $S = 14$

• Número de amostras com classe $C1$: $S1 = 9$

• Número de amostras com classe $C2$: $S2 = 5$

Probabilidades anteriores de cada classe:

• $P(C1)$ = P(JOGA-TÊNIS = Sim) = $S1$/S = 9/14 = 0.643

[7] *A Teoria de Bayes é usada na inferência estatística para atualizar estimativas da probabilidade de que diferentes hipóteses sejam verdadeiras, baseado nas observações e no conhecimento de como essas observações se relacionam com as hipóteses.*

• $P(C2)$ = P(JOGA-TÊNIS = Não) = $S2$/S = 5/14 = 0.357

Para calcular as probabilidades das instâncias de X (Sol, Moderada, Normal e Fraco), relativas a cada atributo (TEMPO, TEMPERATURA, UMIDADE e VENTO), condicionadas a cada classe ($C1$ e $C2$), conta-se o número de amostras de cada classe que possuem a instância considerada.

Probabilidades das instâncias de X relativas a cada atributo, condicionadas em todas as classes:

• $P(x1|C1)$ = P(TEMPO = Sol | JOGA-TÊNIS = Sim) = 2/9 = 0.222

• $P(x1|C2)$ = P (TEMPO = Sol | JOGA-TÊNIS = Não) = 3/5 = 0.600

• $P(x2|C1)$ = P (TEMPER. = Mod. | JOGA-TÊNIS = Sim) = 4/9 = 0.444

• $P(x2|C2)$ = P (TEMPER. = Mod. | JOGA-TÊNIS = Não) = 2/5 = 0.400

• $P(x3|C1)$ = P (UMID. = Normal | JOGA-TÊNIS = Sim) = 6/9 = 0.667

• $P(x3|C2)$ = P (UMID. = Normal | JOGA-TÊNIS = Não) = 1/5 = 0.200

• $P(x4|C1)$ = P (VENTO = Fraco | JOGA-TÊNIS = Sim) = 6/9 = 0.667

• $P(x4|C2)$ = P (VENTO = Fraco | JOGA-TÊNIS = Não) = 2/5 = 0.400

Probabilidade da amostra X condicionada a cada classe:

• P(X êC1) = P(X | JOGA-TÊNIS = Sim) = 0.222 x 0.444 x 0.667 x
x 0.667 = 0.044

• P(X êC2) = P(X | JOGA-TÊNIS = Não) = 0.600 x 0.400 x 0.200 x
x 0.400 = 0.019

• P(X êC1)P(C1) = P(X | JOGA-TÊNIS = Sim) P(JOGA-TÊNIS =
= Sim) = 0.044 x 0.643 = 0.028

• P(X êC2)P(C2) = P(X | JOGA-TÊNIS = Não) P(JOGA-TÊNIS =
= Não) = 0.019 x 0.357 = 0.007

Para usar as regras de classificação, calcula-se:

Logo, como 0.028 > 0.007, isto é, $P(X \mid C1)P(C1) > P(X \mid C2)P(C2)$, o Classificador Bayesiano Simples prediz a classe $C1$ (JOGA-TÊNIS = Sim) para a amostra X.

O algoritmo K-Means usado em tarefas de Clusterização tem como princípio escolher randomicamente os k pontos centróides, tomando k pontos de dados (numéricos) como sendo os centróides dos clusters. Cada ponto (ou registro da base de dados) é atribuído ao cluster cuja distancia deste ponto em relação ao centróide é a menor dentre todas as distancias calculadas. Um novo centróide de cada cluster é computado pela média dos pontos do cluster,

caracterizando a configuração dos clusters para a iteração seguinte. O processo termina quando os centróides dos clusters param de se modificar, ou após um numero limitado de iterações que tenha sido especificado pelo usuário.

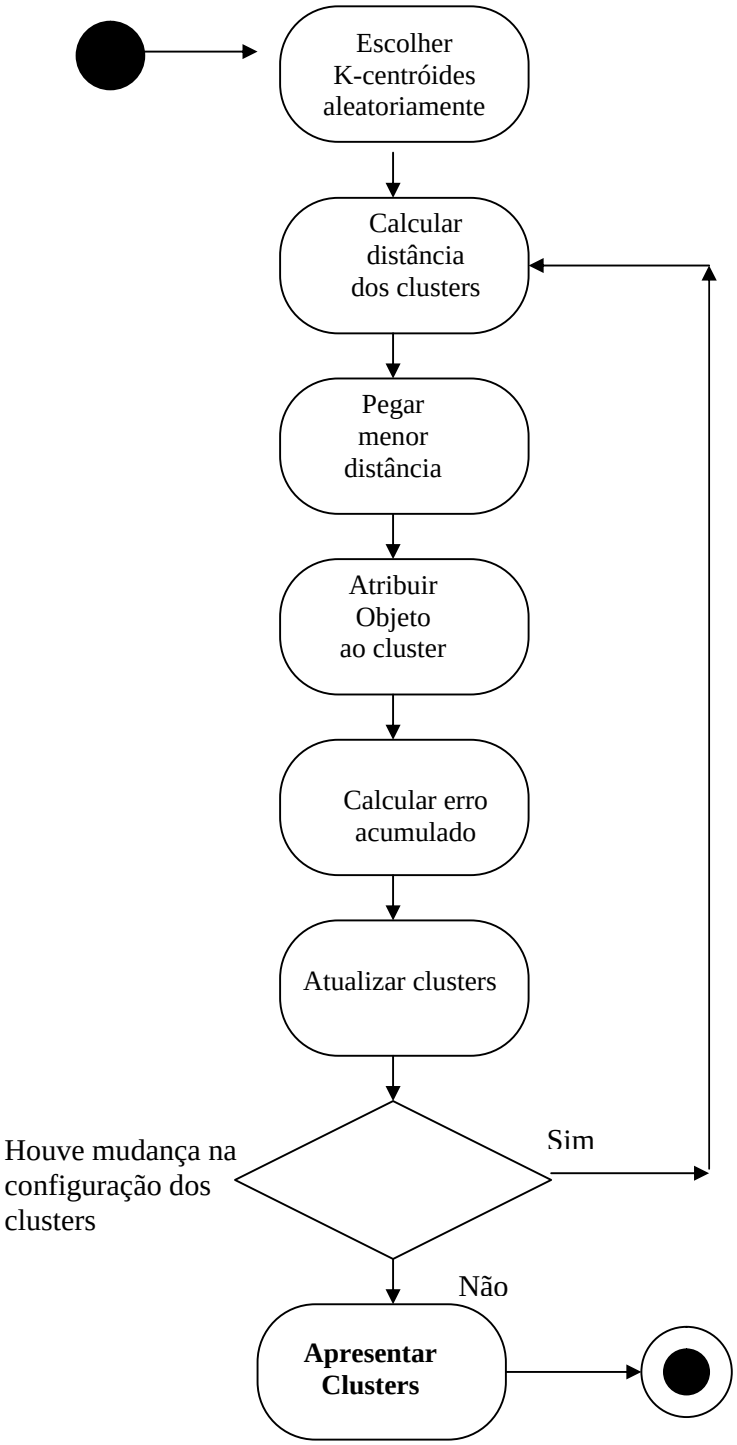

Figura 5 - Fluxograma K-Means

O algoritmo toma um parâmetro de entrada, *k*, dividindo um conjunto de *n* objetos em *k* clusters tal que a similaridade intracluster resultante seja alta, mas a similaridade intercluster seja baixa. A similaridade em um cluster é medida em respeito ao valor médio dos objetos neste cluster (centro de gravidade do cluster). A execução do algoritmo k-means consiste em selecionar aleatoriamente k objetos, que inicialmente representam cada um a média de um cluster. Para cada um dos objetos remanescentes, é feita a atribuição ao cluster ao qual o objeto é mais similar, baseado na distância entre o objeto e a média do cluster. O algoritmo computa as novas médias para cada cluster. O processo vai sendo repetido até que a condição de parada seja atingida.

De acordo com Carlantonio (2001) o método k-means não é adequado para descobrir clusters com formas não convexas ou clusters de tamanhos muito diferentes.

Existem objetos que não seguem o comportamento geral dos dados. Tais objetos, que são diferentes ou inconsistentes em relação ao conjunto de dados formado, são chamados de ruídos (outliers). O método k-means é sensível a ruídos, visto que pequeno numero de dados ruidosos pode influenciar, substancialmente, os valores médios dos clusters. O algoritmo é inicializado com os centros (médias) colocados em posições aleatórias. A busca pelo centro comum se faz de forma iterativa. Após essa inicialização, os objetos restantes são agrupados conforme a distancia em que se encontram das médias.

Há algumas variações do método K-Means são elas:

* O algoritmo K-Modes utilizado para clusterização de dados nominais. No lugar do cálculo da média, calcula-se a moda dos objetos, usando medidas de similaridade para tratar os objetos e métodos baseados em freqüência para atualizar as modas dos clusters.

* O algoritmo K-Prototypes, que é a junção dos dois primeiros, podendo ser aplicado em bases de dados que contenham atributos numéricos e nominais.

* O algoritmos *K-Medoids* que se baseia em encontrar o *medoid* (objeto mais centralmente localizado em um cluster). Os objetos restantes são clusterizados com o medoid ao qual ele é mais similar. Acontece uma troca iterativa, de um medoid por um não medoid, visando a melhoria da clusterização. A qualidade é estimada usando uma função custo que mede a similaridade média entre os objetos e o medoid de seu cluster.

3.2 Tarefas de Mineração de Dados

Harrison (1998) afirma que não há uma técnica que resolva todos os problemas de mineração de dados. Diferentes métodos servem para diferentes propósitos; cada método

oferece suas vantagens e suas desvantagens. A familiaridade com as técnicas é necessária para facilitar a escolha de uma delas de acordo com os problemas apresentados.

As técnicas de mineração de dados podem ser aplicadas a tarefas como classificação, clusterização, associação e estimativa.

3.2.1 - Classificação

A tarefa de Classificação consiste em estudar as características de um objeto e definir uma classe para ele. Esta tarefa tem como objetivo construir modelos que permitam o agrupamento dos dados em classes. É considerada preditiva, já que assim que as classes estejam pré-definidas, quando surgir um novo dado o mesmo já pode ser enquadrado em uma delas.

> Uma população pode ser dividida em categorias para avaliação de concessão de crédito com base em um histórico de transações de créditos anteriores. Em seguida, uma nova pessoa pode ser enquadrada, automaticamente, em uma categoria de crédito específica, de acordo com suas características. (LUIZ HOMERO BASTOS CUNICO- 2005)

Para isso, devem ser considerados dois tipos de atributos que caracterizam o objeto: **atributos preditivos,** cujos valores irão influenciar no processo de determinação da classe; e **atributos objetivos**, que indicam a classe a qual o objeto pertence. Desta forma, a classificação visa descobrir algum tipo de relacionamento entre os atributos preditivos e objetivos. A principal técnica utilizada para a tarefa de classificação é a **Árvore de Classificação (*classification tree*),** se tratando de fato de uma árvore de decisão – com a representação de modelos SE-ENTÃO; lembrando uma pirâmide invertida onde cada caminho possível pode ser definido como *regra de decisão*.

Em um processo de Extração de Dados, a classificação está especificamente voltada à atribuição de uma das classes predefinidas pelo analista a novos fatos ou objetos submetidos ao classificador.

> A partir de uma árvore de decisão é possível derivar regras. As regras são escritas considerando o trajeto do nodo raiz até uma folha da árvore. Estes dois métodos são geralmente utilizados em conjunto. Devido ao fato das árvores de decisão tenderem a crescer muito, de acordo com algumas aplicações, elas são muitas vezes substituídas pelas regras. Isto acontece em virtude das regras poderem ser facilmente modularizadas.[Simone Garcia - Universidade Federal do Rio Grande do Sul- 2000]

Imagine um algoritmo que estuda o banco de dados de uma instituição financeira, visando a aprovação ou não (atributo objetivo) de crédito para empréstimo de seus clientes. Nessa base de dados, existem pessoas adimplentes e inadimplentes sendo cada classe caracterizada por algum tipo de padrão. Neste processo, os clientes do banco de dados cujo

campo resultado venha a ter o valor **não**, representarão os inadimplentes. Para poder preencher esse campo, serão consideradas as características dos clientes (atributos preditivos) existentes no banco. Normalmente, um analista indica quais são os atributos relevantes para a predição.

Já as Redes Neurais com aprendizado não-supervisionado são adequadas para tarefas que envolvam descrição dos dados, como, por exemplo, a tarefa de Clusterização.

Exemplo:

ID	Salário	Idade	Tipo Emprego	Classe
1	3000	30	Autônomo	B
2	4000	35	Indústria	B
3	7000	50	Pesquisa	C
4	6000	45	Autônomo	C
5	7000	30	Pesquisa	B
6	6000	35	Indústria	B
7	6000	35	Autônomo	A
8	7000	30	Autônomo	A
9	4000	45	Indústria	B

Tabela 3 – Informações sobre alguns indivíduos

Neste exemplo temos como objetivo agrupar os indivíduos nas classes A,B,C, de acordo com seus atributos, a árvore de decisão gerada (Figura 6) nos indica que o individuo com salário <= 5000, independente de outros fatores se enquadra na Classe B, no caso de um salário > 5000 e idade > 40 Classe C.

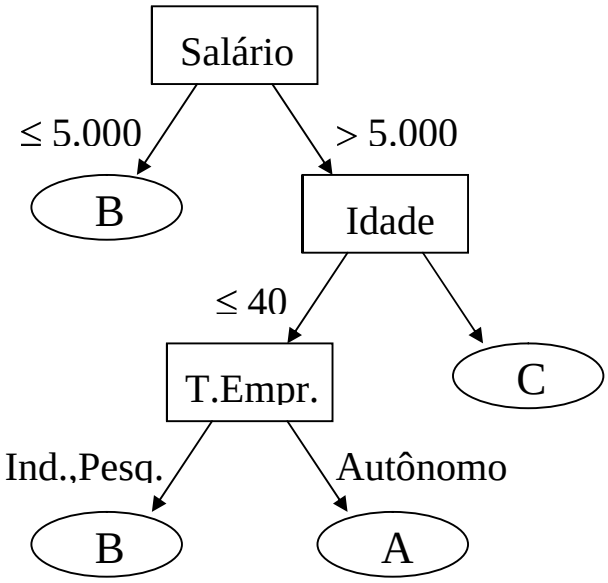

Figura 6 – Árvore de Classificação

Abaixo, temos as Regras de Decisão geradas da mesma forma que a árvore representam cada caminho que pode ser seguido desde a raiz até a folha da arvore, e que nos levam a um resultado igual ao da árvore.

Regras de Decisão
$(Sal \leq 5.000) \Rightarrow Classe = B$

$(Sal > 5000) \wedge (Idade > 40) \Rightarrow Classe = C$

$(Sal > 5000) \wedge (Idade \leq 40) \wedge (TEmpr = Autônomo) \Rightarrow Classe = A$

$(Sal > 5000) \wedge (Idade \leq 40) \wedge ((TEmpr = Indústria) \vee (TEmpr = Pesquisa)) \Rightarrow Classe = B$

3.2.2 - Clusterização ou Agrupamento

Na tarefa de Clusterização os dados são agrupados de acordo com características consideradas similares, encontradas pelo próprio algoritmo. É uma tarefa descritiva, o que diferencia da Classificação, já que as classes não são definidas previamente. O algoritmo de clusterização identifica as classes automaticamente, dividindo os dados em clusters para serem analisados e transformados em conhecimento.

> A tarefa básica da clusterização é agrupar um conjunto de objetos em subconjuntos, de acordo com os critérios apropriados. Esses subconjuntos agrupam elementos que têm um alto grau de semelhança ou similaridade, enquanto que, quaisquer elementos pertencentes a grupos distintos tenham pouca semelhança entre si. (VANIA BOGORNY – 2003)

Com a técnica de clusterização é possível identificar grupos diretamente dos dados sem que pra isso, os mesmos sejam previamente conhecidos. Em uma massa de dados geográficos, pode-se, por exemplo, agrupar construções (prédios, casas) de uma área, de acordo com determinadas características similares como categoria, localização geográfica área construída.

> A qualidade dos clusters gerados depende de uma série de definições estabelecidas pelo usuário como, por exemplo, escolha dos atributos, medidas de dissimilaridade, critérios de agrupamento, escolha do algoritmo e definição do número de clusters. A dissimilaridade normalmente é utilizada por essa técnica para avaliar o grau de semelhança entre dois objetos durante o processo de agrupamento. Muitas vezes, essa medida é apresentada como sendo a distância entre dois objetos. (VANIA BOGORNY – 2003)

Os algoritmos de clusterização podem ser divididos em categorias como: hierárquicos, de particionamento, de densidade e de restrições de contiguidade.

Algoritmos Hierárquicos

Um algoritmo bem conhecido utilizado para essa categoria é o HAC. Havendo duas formas de se implementar a técnica sendo:

1- Divisivo: Os dados são agrupados em um só cluster e vão sendo divididos em sub-grupos (menores);

2- Aglomerativo: Os dados são agrupados inicialmente em partes bem pequenas e vão aglomerando em grupos maiores (clusters). (LUIZ HOMERO BASTOS CUNICO-2005)

No Aplicativo Weka o algoritmo que implementa essa categoria é chamado CobWeb.

Algoritmos de Particionamento

Nesta categoria o algoritmo mais utilizado é o *K-means (meios)* que divide o grupo total de dados em subgrupos. Os dados são tratados como um vetor e, cada característica é vista como um componente vetorial (coordenadas: x,y). Os dados qualitativos são transformados em variáveis numéricas, pois o algoritmo trabalha com proximidade entre os pontos. Pontos próximos formam um cluster (grupo). Para plotagem dos dados, são utilizadas (comumente) as funções de distância[8]: Euclidiana[9] e a *Manhattan*[10]. Previamente, o analista define o número de grupos que será criado, o número K. O algoritmo então divide os dados em K grupos e escolhe um ponto chamado centróide, no meio de cada grupo. Em seguida define novamente os centróides de acordo com sua proximidade em relação aos outros pontos do grupo. Com o novo centróide, os grupos são plotados mais uma vez. Os centróides são recalculados repetindo o processo até que os k grupos estejam bem definidos.

Exemplo:

Deseja-se separar os clientes em grupos de forma que aqueles que apresentam o mesmo comportamento de consumo fiquem no mesmo grupo.

Temos na Tabela 4 dados sobre compras de consumidores.

Consumidor	Qtd. Tot.Prods.	Preç..Prods
1	2	1700,00
2	10	1800,00
3	2	100,00

[8] A função de distancia é responsável por calcular quão semelhantes são dois objetos através da análise de seus atributos e, a partir deste cálculo, gerar uma distância no espaço euclidiano entre eles. Os atributos utilizados pela função de distancia podem ser quaisquer dados: numéricos, datas, dados similares ou textuais. É usada para indicar o relacionamento entre os itens de dados, guiando a geração de um gráfico em três dimensões.

[9] .A distância euclidiana é a raiz quadrada da soma dos quadrados das diferenças de valores para cada variável.

[10] A distância *manhattan é definida em um sistema* cartesiano de coordenadas fixo, como a soma dos comprimentos da projeção da linha que une os pontos com os eixos das coordenadas. Depende da rotação do sistema de coordenadas mas não da sua translação ou da sua reflexão em relação a um eixo coordenado.

4	3	2000,00
5	12	2100,00
6	3	200,00
7	4	2300,00
8	11	2040,00
9	3	150,00

Tabela 4 – Informações sobre consumidores

Cada grupo identificado (Tabela 5) é caracterizado por consumidores semelhantes em relação à quantidade total e ao preço médio dos produtos consumidos.

Grupo	Consumidor	Qtd. Total	Preç. Total
1	1	2	1700
	4	3	2000
	7	4	2300
2	2	10	1800
	5	12	2100
	8	11	2040
3	3	2	100
	6	3	200
	9	3	150

Tabela 5 – Clusters criados com a tarefa

Grupo 1 – Gastos com cada produto entre 575 e 850

Grupo 2 – Gastos com cada produto entre 175 e 185

Grupo 3 – Gastos com cada produto entre 50 e 66

3.2.3 - Associação

Uma Regra de Associação caracteriza o quanto a presença de um conjunto de itens nos registros de uma Base de Dados implica na presença de algum outro conjunto distinto de itens nos mesmos registros (Agrawal & Srikant 1994). Procura estabelecer um padrão no relacionamento entre os dados. Considerando as vezes em que há ocorrência de determinado dado em concomitância com outro. É considerada uma tarefa descritiva

O alvo das Regras de Associação é descobrir tendências que possam ser utilizadas para compreender e extrair conhecimento a partir da análise do comportamento dos dados. Por exemplo, analisando as estatísticas de vendas de uma determinada rede de lojas,

percebemos que sete em cada 10 clientes que compram um produto X também compram, naquele momento, o produto Y, nessa regra 70% corresponde à sua confiabilidade.

A Regra de Associação do exemplo envolvendo os produtos X e Y pode ser representada como:

X =>Y, sendo X ≠ Y

*As Regras de Associação estabelecem uma correlação estatística entre atributos de dados e conjuntos de dados. (*Maria Madalena Dias – 1992)

Um algoritmo de extração de Regras de Associação deve criar regras que possuam suporte e confiabilidade de acordo com os parâmetros especificados pelo usuário, podendo conter um ou mais itens; o que varia com o tamanho da base de dados. Diversas regras são geradas, devendo ser avaliadas pelo usuário, para que escolher as mais importantes ou de maior peso na tomada de decisão.

Exemplo:

Id-Transação (TID)	Itens Comprados
1	leite, pão, refrigerante.
2	cerveja, carne
3	cerveja, fralda, leite, refrigerante.
4	cerveja, fralda, leite, pão.
5	fralda, leite, refrigerante.

Neste exemplo, um tipo clássico de utilização da Tarefa de Associação, gerado com o Algoritmo Apriori, onde é possível observar a relação entre os produtos fralda e leite, já que sempre que se comprou fralda o cliente também levou leite. Os resultados dessa análise podem ser úteis na elaboração de catálogos e do layout de prateleiras de forma que produtos que tendem a ser comprados juntos fiquem próximos uns dos outros

{fralda} ⇒ {cerveja}	confiança de 66%	(suporte médio)
{fralda} ⇒ {leite}	**confiança de 100%**	**(suporte alto)**
{leite} ⇒ {fralda}	confiança de 75%	(suporte alto)
{carne} ⇒ {cerveja}	confiança de 100%	(suporte baixo)

3.2.4 - Estimativa

Estima-se um valor para um determinado dado baseado em valores conhecidos de outras variáveis. É uma tarefa preditiva.

A estimação, ao contrário da classificação, está associada às respostas contínuas. Assim, pode-se estar interessado em estimar a renda média de uma família com base em seus bens duráveis informados em um questionário, a expectativa de vida de um novo cliente de uma seguradora com base em seu formulário de admissão, ou a propensão à inadimplência associada a um postulante de empréstimo calculada a partir de suas características pessoais. Os modelos de regressão e as redes neurais são bastante utilizados nestes casos, sendo que especificamente nos casos de regressão para estimação em processos de *Data Mining*, é preciso estar atento para uma diferença teórica existente entre estimação e predição em termos estatísticos.

A previsão, como tarefa típica de Mineração de Dados, está associada à avaliação de um valor futuro de uma variável resposta a partir de seus dados históricos. Assim, pode-se prever o preço de determinada ação, ou o número de clientes que serão perdidos por uma empresa, em um dado horizonte futuro de tempo. As técnicas que podem ser utilizadas aqui são, dentre outras, as redes neurais, os métodos estatísticos de séries temporais, a regressão, as árvores de decisão e o raciocínio baseado em casos.

Exemplo:

Assistencias p/ min	**Tempo em quadra**	**Idade**	**Pontos p/ min**	**Altura**
0,0888	36,02	28	0,5885	2,01
0,1399	39,32	30	0,8291	1,98
0,1107	35,22	25	0,4799	1,93
0,2521	31,73	29	0,5735	1,83
0,1007	28,81	31	0,6318	1,93
0,1067	35,6	23	0,4326	1,96

Tabela 6 – Dados de jogadores de basquete

O objetivo deste exemplo é estimar a altura de um jogador com base nos dados existentes e a relação entre eles. Na Tabela 7, com os dados do novo jogador estimou-se uma altura de 1,88m.

Assistencias p/ min	Tempo em quadra	Idade	Pontos p/ min	Altura
0,19	0,6412	34	25	1,88

Tabela 7 – Dados do novo jogador e altura estimada

4. POSSÍVEIS ÁREAS DE APLICAÇÃO DE DATA MINING

- ***Marketing***. Técnicas de mineração de dados são aplicadas para descobrir preferências do consumidor e padrões de compra, com o objetivo de realizar marketing direto de produtos e ofertas promocionais, de acordo com o perfil do consumidor.

- **Detecção de fraudes**. Muitas fraudes óbvias (tais como, a compensação de cheque por pessoas falecidas) podem ser encontradas sem mineração de dados, mas padrões mais sutis de fraude podem ser difíceis de ser detectados, por exemplo, o desenvolvimento de modelos que predizem quem será um bom cliente ou aquele que poderá se tornar inadimplente em seus pagamentos.

- **Medicina**. Caracterizar comportamento de paciente para prever visitas, identificar terapias médicas de sucesso para diferentes doenças, buscar por padrões de novas doenças.

- **Instituições governamentais**. Descoberta de padrões para melhorar as coletas de taxas ou descobrir fraudes.

- **Ciência**. Técnicas de mineração de dados podem ajudar cientistas em suas pesquisas, por exemplo, encontrar padrões em estruturas moleculares, dados genéticos, mudanças globais de clima, oferecendo conclusões valiosas rapidamente.

- **Controle de processos e controle de qualidade**. Auxiliar no planejamento estratégico de linhas de produção e buscar por padrões de condições físicas na embalagem e armazenamento de produtos. (Maria Madalena Dias, 2001).

- **Telefonia**. Com o objetivo de realizar a classificação de clientes de grandes empresas do ramo de telecomunicações de acordo com seu potencial de compra de serviços. Uma vez caracterizado o potencial de compras de todos os clientes, ações de marketing específicas por cliente poderão ser realizadas.

- **Área Financeira**. Gerando um classificador para caracterizar clientes que pagam em dia, clientes que pagam em atraso e clientes que não pagam seus créditos. Assim, podendo prever em novos cadastros prováveis maus pagadores.

Como sugestão comentada por Emmanuel Passos & Ronaldo Goldschmidt (2005) em um projeto importante na área da produção consistiria na aplicação de técnicas de Mineração para geração de modelos que façam a previsão de demanda de energia elétrica para determinados períodos. Nesse caso, a tarefa de previsão de séries temporais deve ser realizada, utilizando-se, para tanto, de registros de consumos de energia elétrica ao longo de períodos anteriores. Aplicação similar pode ser realizada na produção de insumos industrializados, baseado em históricos de volumes de vendas anteriores. Um modelo de

conhecimento que permita previsões deste tipo pode ser incorporado a um sistema de planejamento da produção.

5. UTILIZANDO A FERRAMENTA WEKA

O aplicativo Weka desenvolvido em linguagem Java, pela Universidade de Waikato na Nova Zelândia, é um software de domínio público, tem como grande vantagem sua portabilidade, podendo rodar em diversas plataformas. A ferramenta tem duas formas de utilização: através de linha de comando e de interface gráfica.

Nele é possível executar diversas tarefas de mineração de dados como regras de Classificação, Associação, Clusterização e Estimativa.

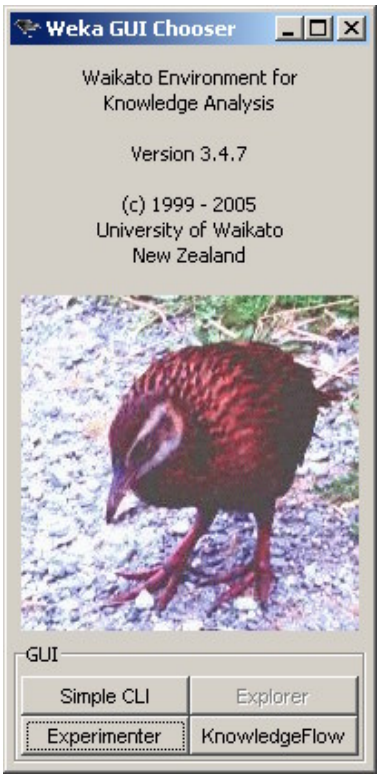

Figura 7 - Aplicativo Weka em execução

5.1 - Exemplo de Aplicação de Árvores de Classificação

Para utilizar a interface gráfica do Weka, basta escolher a opção Explorer ao executar o aplicativo.

O objetivo deste exemplo é gerar uma árvore que auxilie na tomada de decisão entre jogar ou não jogar tênis a depender das condições climáticas. Utilizamos a tabela *weather.arff* (Figura 7), da própria ferramenta. A tabela *weather.arff* possui quatro atributos preditivos e um objetivo (*Play*), são eles:

* *Outlook* (Previsão) que pode ter como valores: *sunny* (iluminado pelo sol), *overcast* (nublado), *rainy* (chuvoso).

* *Temperature* (Temperatura) que é do tipo real, representando a temperatura prevista para o dia.

* *Humidity* (Umidade) que também é do tipo real, representando a umidade do ar prevista para o dia.

* *Windy* (Vento) que tem como possíveis valores *True* (verdadeiro) e *False* (falso), significando a existência ou não de vento.

* *Play* (Jogar) que tem como resultado *Yes* ou *No*, jogar ou não jogar.

```
@relation weather

@attribute outlook {sunny, overcast,
rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
@data

sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Figura 8 – Weather

Na Figura 8 podemos observar como cada atributo **preditivo** influencia no resultado do atributo **objetivo** *Play*. No gráfico a cor azul representa o valor *Yes*, ou seja o dia está bom para jogar tênis; enquanto a cor vermelha representa o valor *No*¸ pois não é um dia bom para o jogo. No ultimo quadro os resultados possíveis para o atributo *Play*, sendo 9 (nove) *Yes* e 5 (cinco) *No.*

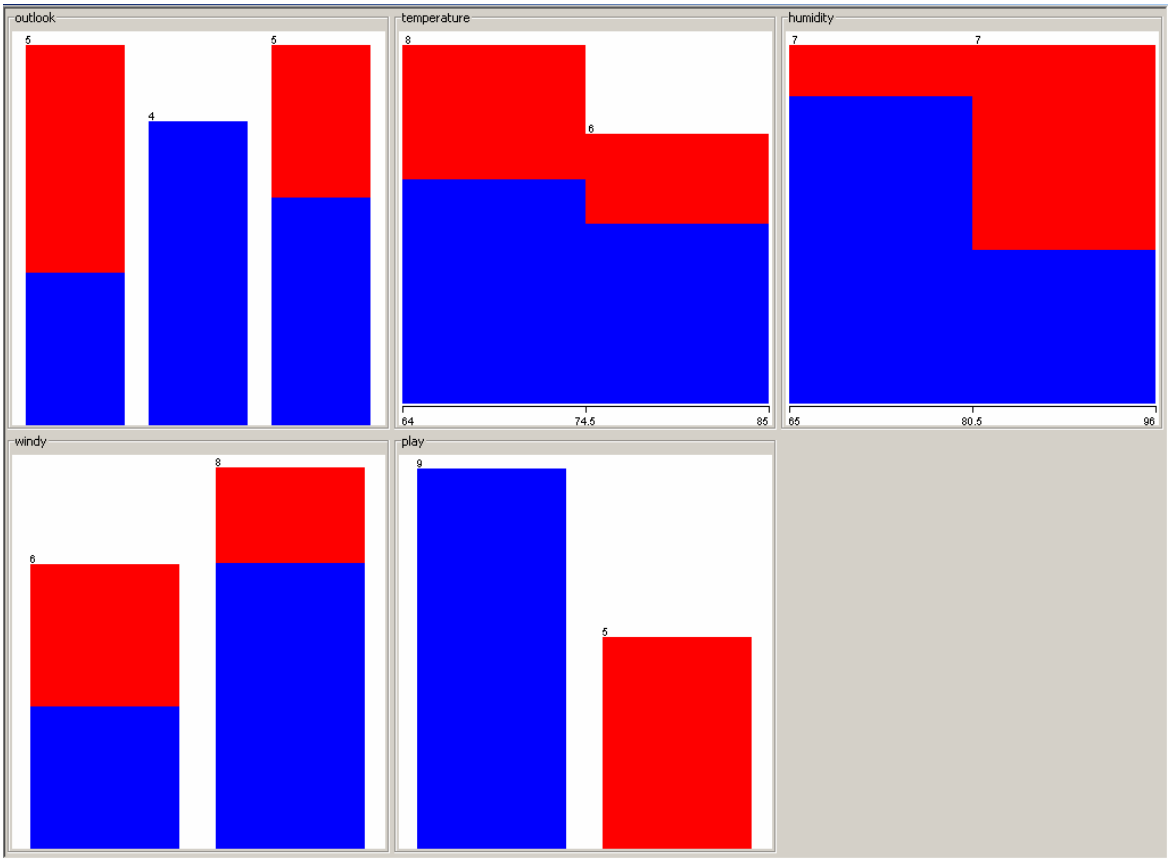

Figura 9 – Atributos

Na Figura 9 está representada a Árvore de decisão da tabela *weather*, com os dados hierarquizados na forma de nós (estágios de decisão) e separados em classes (*Yes/No*). Para gerar a Árvore de decisão utilizamos o algoritmo (do próprio aplicativo) J4.8.

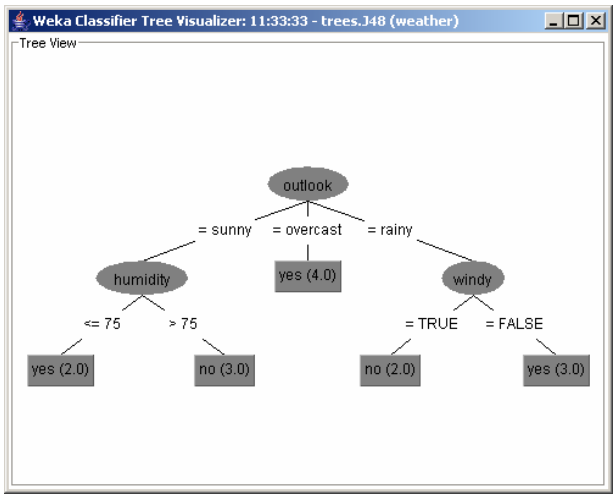

Figura 10 - Árvore de Decisão

5.2 - Exemplo de aplicação de Regras de Associação

Utilizando uma variação da tabela *weather.arff* (do exemplo anterior) onde os atributos preditivos *temperature* e *humidity*, antes numéricos agora são preenchidos com valores nominais, sendo:

* *Temperature* {*hot, mild, cool*}

* *Humidity* {*high, normal*}

Esta conversão é necessária porque o algoritmo utilizado (***Apriori***) só trabalha com *strings*.

```
@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

Figura 11 - *weather.nominal.arff*

Os parâmetros escolhidos para definir a execução dos algoritmos ***Apriori*** podem ser vistos na Figura 10, são eles:

* O valor de confiança default é de 0.09 (minMetric 90%).

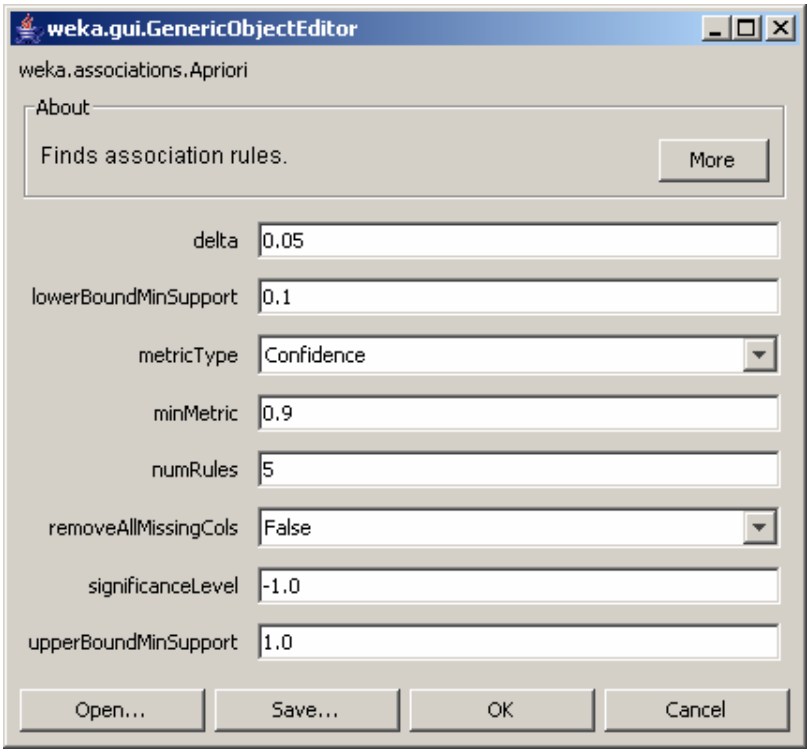

Figura 12 - Parâmetros do *Apriori*

* O valor mínimo de suporte começa com 1 (100%), vai diminuindo 0.05 (delta 5%) até que 5 regras (numRules) seja geradas, ou que o valor de suporte chegue a 0.01 (lowerBoundMinSupport).

O Resultado são as cinco melhores regras de Associação encontradas pelo algoritmo.

```
Melhores regras:

1. humidity=normal windy=FALSE 4 ==> play=yes 4    conf:(1)
2. temperature=cool 4 ==> humidity=normal 4    conf:(1)
3. outlook=overcast 4 ==> play=yes 4    conf:(1)
4. temperature=cool play=yes 3 ==> humidity=normal 3    conf:(1)
5. outlook=rainy windy=FALSE 3 ==> play=yes 3    conf:(1)
```

Figura 13 - Melhores regras geradas

O Símbolo ‘=>’ separa o antecedente do conseqüente.

O numero antes do símbolo ‘=>’ indica o suporte da regra, em quantos registros a regra aparece, nesse caso em 4 de 14 registros.

No final de cada regra é mostrada a confiança de cada uma delas (conf: 1 ou 100%).

Com as regras de associação podem ser gerados novos conhecimentos, como se vê na Regra 1 em que a umidade normal e a ausência de vento são suficientes para o dia ser bom para jogar tênis, independente de outras variáveis.

6. CONCLUSÃO

Considerando os avanços e a relevância que a área da Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases – KDD)* vem ganhando no contexto mundial, esta pesquisa tem como principal objetivo fornecer uma introdução ao assunto. Foram apresentados alguns conceitos básicos, técnicas, orientações e exemplos de aplicações em KDD e Mineração de Dados.

A Mineração de Dados surgiu com o objetivo principal de dar suporte à tomada de decisões na empresa. Assim, a aplicação de técnicas em sistemas de descoberta de conhecimento busca a descoberta de regras e padrões em dados que trarão o conhecimento suficiente e adequado para aquelas pessoas responsáveis pela tomada de decisões na empresa. É necessário que o usuário de um sistema de KDD tenha um grande conhecimento do negócio da empresa assim como da própria base de dados que está lidando, para ser capaz de selecionar corretamente os conjuntos de dados e as classes de padrões relevantes.

Conforme foi exposto no decorrer da pesquisa, a participação humana na condução do processo de KDD é de fundamental importância para o sucesso das aplicações. O que requer um amplo estudo e o domínio sobre a fundamentação teórica (envolvendo conceitos, técnicas e algoritmos) aliada a uma intensa vivência prática de diversas experiências reais. Todo o processo de Descoberta é bastante complexo e trabalhoso, pois envolve a execução de diversas tarefas, configuração de parâmetros e grande interação com o usuário. No entanto, o sucesso do processo pode trazer uma recompensa de grande valia para as organizações.

A atual escassez de profissionais com conhecimento para atuar na área de KDD, combinada com a velocidade com que as bases de dados vêm aumentando de tamanho e se multiplicando, tem contribuído decisivamente para uma intensificação da demanda por aplicações nesta área. Assim, diante deste cenário de oportunidades, a formação de pessoal especializado em KDD vem se tornando uma opção de grande procura e interesse na atualidade.

REFERÊNCIAS

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. ***Mining Association Rules Between Sets of Itens in Large Databases.*** ACM SIGMOD Conference Management of Data, 1993.

ÁLVARES, Reinaldo Viana. Mineração de Dados: Introdução e Aplicações. **SQL Magazine**, São Paulo: Editora DevMedia, p. 30-36, Abril, 2006.

BOGORNY, Vânia. **Algoritmos e Ferramentas de Descoberta de Conhecimento em Bancos de Dados Geográficos.** Programa de Pós-Graduação em Computação, Universidade Federal do Rio Grande do Sul, Instituto de Informática, 2003. Disponível em: <http://www.inf.ufrgs.br/~vbogorny/ti3-final.pdf> Acesso em: 10 de Outubro de 2006.

CARLANTONIO, L. M. **Novas Metodologias para Clusterização de Dados.** Dissertação de Mestrado, Engenharia Civil, COPPE, Universidade Federal do Rio de Janeiro, 2001.

CUNICO, Luiz Homero Bastos. **Técnicas em *data mining* aplicadas na predição de satisfação de Funcionários de uma rede de lojas do comércio varejista**. Dissertação (Mestrado) - Universidade Federal do Paraná, 2005. Disponível em: <http://www.ppgmne.ufpr.br/disser.htm> Acesso em: 09 de Outubro de 2006.

DIAS, Maria Madalena - **Parâmetros na escolha de técnicas e ferramentas de mineração de dados.** Graduação em Formação de Tecnólogo em Processamento de Dados - Universidade Estadual de Maringá, UEM, Paraná, 2001, Brasil. Disponível em: <http://www.ppg.uem.br/docs/ctf/Tecnologia/2002> Acesso em: 12 de Outubro de 2006

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. *From Data Mining to Knowledge Discovery: An Overview.* Knowledge Discovery and Data Mining, Menlo Park: AAAI Press, 1996a.

GARCIA, Simone. **O uso de árvores de decisão na descoberta de conhecimento na área da saúde.** Semana Acadêmica 2000. Universidade Federal do Rio Grande do Sul, 2000, RS,

Brasil. Disponível em: < http://www.inf.ufrgs.br/pos/SemanaAcademica/Semana2000/ SimoneGarcia> Acesso em: 12 de Outubro de 2006

GOLDSCHMIDT, Ronaldo; PASSOS Emmanuel. **Data Mining Um guia prático**. Rio de Janeiro: Editora Campus, 2005. 261p.

GOLDSCHMIDT, Ronaldo; PASSOS Emmanuel; VELLASCO, M.; PACHECO, M. *Task Definition Assistance in KDD Applications.* CLEI'03 - XXIX Conferência Latino Americana de Informática. La Paz, 2003.

HARRISON, T.H. - ***Intranet data warehouse***. São Paulo: Editora Berkeley Brasil, 1998.

MOTTA, Custódio Gouvêa Lopes. **Introdução a Técnicas de *Data Mining* – DM - Classificação de Dados –** Universidade Federal de Juiz de Fora LNCC/MCT – 02 de Fevereiro de 2005. Disponível em: < http://www.lncc.br/verao/verao05/arquivos/ MiniCurso DMLNCC050202C.pdf> Acesso em: 10 de Outubro de 2006.

REZENDE, Solange Oliveira - **Mineração de Dados**. (Livre docência) Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação, Departamento de Ciências de Computação e Estatística, SP, 2004.

SANTOS, Marcelino Pereira. **Mineração de Dados. Conceitos, Aplicações e Experimentos com Weka**, Dissertação (Mestrado) - Universidade do Estado do Rio Grande do Norte, Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, Brasil. Disponível em: <www.sbc.org.br/bibliotecadigital> Acesso em 12 de Outubro de 2006

VICTOR, André O. **Conceitos e Técnicas de Mineração de Dados (Data Mining),** 2005. Disponível em: www.sbc.org.br/bibliotecadigital. Acesso em 12 de Outubro de 2006

WIRTH, R.; SHEARER, C.; GRIMMER, U.; REINARTZ, T.; SCHLOSSER, J.; BREITNER, C.; ENGELS, R.; LINDNER, G. *Towards Process-Oriented Tool Support for Knowledge Discovery in Databases.* Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery. Trondheim, 1997.