



Universidade Federal de Pernambuco
Graduação em Ciência da Computação
Centro de Informática

2006.2

Conceitos, técnicas, ferramentas e aplicações de
Mineração de Dados para gerar conhecimento a
partir de bases de dados

Aluno: Thiago Amorim (tmas@cin.ufpe.br)

Orientador: Fernando da Fonseca de Souza (fdfd@cin.ufpe.br)

“A verdadeira perfeição precisa ser imperfeita” – Noel Gallagher (Oasis)

RESUMO

Este trabalho apresenta a análise de técnicas, ferramentas e metodologias para praticar-se uma importante área da administração de dados em uma organização, denominada de Mineração de Dados, visando à melhoria da qualidade dos mesmos. As etapas do processo de mineração de dados são detalhadas. Algumas técnicas utilizadas durante o processo são abordadas. É analisada a utilização da mineração de dados atualmente em diversas áreas importantes de nossa sociedade. São também analisadas ferramentas que auxiliam este processo.

Palavras-chave: Mineração de dados, Qualidade de dados, Administração de Dados, Arquitetura de Dados, Modelagem de Dados.

ABSTRACT

This work presents an analysis of techniques, tools and methodologies for the practice of an important area of Data Administration in an organization, known as Data Mining. It is aimed at improving data quality. The stages of the data mining process are detailed too. Some techniques used during the process are focussed. The use of data mining on different important areas of our society is analyzed. Tools that assist this process are also analyzed.

Keywords: Data Mining, Data Quality, Data Administration, Data Architecture, Data Modeling

Dedico este trabalho a minha família e amigos

AGRADECIMENTOS

Aos meus pais, por sempre terem colocado a minha educação em primeiro lugar, e por sempre terem se sacrificado para que eu pudesse realizar os meus objetivos.

Aos meus irmãos e amigos, por serem fonte de minha inspiração, e da minha felicidade.

Ao meu orientador, Fernando, pelos ensinamentos, apoio e ajuda no desenvolvimento deste trabalho.

SUMÁRIO

1	<i>Introdução</i>	10
1.1	Problemática	10
1.2	Objetivos da Monografia	12
1.2.1	Objetivo Principal	12
1.2.2	Objetivos Específicos	12
1.3	Relevância	12
1.4	Metodologia de Trabalho	14
1.5	Organização da Monografia	15
2	<i>Mineração de Dados</i>	16
2.1	Fases da Mineração de Dados	18
2.1.1	Entendimento do Negócio (Business Understanding)	19
2.1.2	Seleção dos Dados (Data Understanding)	20
2.1.3	Limpeza dos Dados (Data Preparation)	20
2.1.4	Modelagem dos Dados (Modeling)	21
2.1.5	Avaliação do processo (Evaluation)	21
2.1.6	Execução (Deployment)	22
2.2	Técnicas	22
2.2.1	Classificação	23
2.2.2	Estimativa	24
2.2.3	Previsão	25
2.2.4	Análise de Afinidades	25
2.2.5	Análise de agrupamentos	26
2.3	Tipos de Algoritmos	27
2.3.1	Árvores de Decisão	27
2.3.2	Redes Neurais	28
3	<i>Aplicações Práticas</i>	30
3.1	Segmentação de Mercados	30
3.2	Varejo	33
3.3	Mercado Financeiro	34

4	<i>Análise de Ferramentas</i>	36
4.1	SAS Enterprise Miner	36
4.2	IBM Intelligent Miner	39
4.3	Oracle Darwin Data Mining Software	43
5	<i>Conclusões</i>	46
5.1	Trabalhos Futuros.....	46
5.2	Considerações Finais.....	47
6	<i>Referências</i>	48

LISTA DE FIGURAS

Figura 1: Evolução do valor estratégico de bases de dados (Baseado em [Navega, 2002]).	13
Figura 2: Fases do CRISP-DM Process Model (Baseado em [The CRISP-DM Consortium, 2000]).	19
Figura 3: Fluxo da mineração de dados em análise de empréstimo [SAS, 2007].	37
Figura 4: Análise de Agrupamentos [SAS, 2007].	38
Figura 5: Comparação de técnicas [SAS, 2007].	39
Figura 6: Análise de agrupamentos de clientes bancários [IBM, 2007].	40
Figura 7: Gráfico de setores circulares INT_CREDITCARD [IBM, 2007].	41
Figura 8: Tabela de clientes, com os agrupamentos informados [IBM, 2007].	41
Figura 9: Análise de afinidades [IBM, 2007].	42
Figura 10: Wizards para criação de modelos de mineração de dados [Oracle, 2007].	44
Figura 11: Visualização em forma de árvore para segmentação de clientes [Oracle, 2007].	44
Figura 12: Gráficos para visualização de resultados em formato MS Excel [Oracle, 2007].	45

1 Introdução

Cada vez mais, imensos volumes de informação têm sido sistematicamente coletados e armazenados por grandes organizações. A quantidade de informação armazenada ultrapassa a habilidade técnica e a capacidade humana na sua interpretação. Bancos de dados, por exemplo, são medidos hoje em gigabytes e terabytes.

O custo de uma má qualidade desses dados pode ser decisivo para o sucesso de uma empresa. Com isso, é essencial o estudo da Administração de Dados para garantir a qualidade dos dados que são inseridos e recuperados em uma base de dados. A administração de dados consiste no desenvolvimento e execução de estratégias, práticas e procedimentos para o processo de gerência do ciclo de vida completo dos dados em uma empresa [DAMA International, 2006].

1.1 Problemática

As organizações têm se mostrado extremamente eficientes em capturar, organizar e armazenar grandes quantidades de dados, obtidos de suas operações diárias. Porém, a maioria delas ainda não usa adequadamente essa gigantesca massa de dados para transformá-la em conhecimentos que possam ser utilizados em suas próprias atividades. Com a geração de um volume cada vez maior de informação, é essencial tentar aproveitar o máximo possível desse investimento.

Uma excelente prática de Administração de Dados é o enriquecimento dos dados, gerando ainda mais informação e conhecimento, melhorando assim a sua qualidade. Esta informação preciosa está na verdade implícita escondida sob uma grande massa de dados, e não pode ser descoberta utilizando-se sistemas de gerenciamento

de banco de dados convencionais¹. A solução existe, e chama-se Mineração de Dados [Navega, 2002]. Este processo é uma etapa de outro conhecido como extração de conhecimento em bases de dados ou Knowledge-Discovery in Databases (KDD)² [Fayyad et al., 1996].

Mineração de dados, ou data mining, é o processo de análise de conjuntos de dados que tem por objetivo a descoberta de padrões interessantes e que possam representar informações úteis. O processo de mineração de dados permite que se investiguem esses dados à procura de padrões que tenham valor para a empresa. Este conceito está se tornando cada vez mais popular como uma ferramenta de gerenciamento de informação, que deve revelar estruturas de conhecimento, podendo guiar decisões em condições de certeza limitada.

Com a utilização da mineração de dados, pode-se analisar dados comportamentais, obtendo conhecimento que estava “escondido” na base de dados, gerando aprendizado e dados complementares que podem influenciar no desenvolvimento de estratégias na organização. Esta tecnologia está sendo usada para descrever características do passado, assim como prever tendências para o futuro. Sua utilização permite avanços tecnológicos e descobertas científicas, além de garantir uma vantagem competitiva invejável [Sferra e Corrêa, 2003].

¹ De acordo com o Wikipedia, a enciclopédia livre: Um Sistema Gerenciador de Banco de Dados ou Sistema Gestor de Base de Dados (SGBD) é o conjunto de programas de computador (software) responsáveis pelo gerenciamento de uma base de dados. O principal objetivo é retirar da aplicação cliente a responsabilidade de gerenciar o acesso, manipulação e organização dos dados. O SGBD disponibiliza uma interface para que os seus clientes possam incluir, alterar ou consultar dados. Em bancos de dados relacionais a interface é constituída pelas API ou drivers do SGBD, que executam comandos na linguagem SQL [Wikipedia, 2006].

² Uma definição importante para KDD foi elaborada por Usama Fayyad (Fayyad et al. 1996): "...o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis" [Fayyad, 1996].

1.2 Objetivos da Monografia

Nesta seção serão apresentados o objetivo principal desse trabalho e o detalhamento dos objetivos específicos.

1.2.1 Objetivo Principal

O objetivo principal desta monografia é destacar o tópico da mineração de dados como uma importante área da Administração de Dados, sendo um processo para melhorar a qualidade dos dados de uma organização.

1.2.2 Objetivos Específicos

Ainda serão abordados assuntos específicos, com a finalidade de:

- Esclarecer a relevância do tema como meio de agregar valor a diferentes áreas da nossa sociedade; e
- Analisar a utilização de ferramentas para auxiliar no processo de mineração de dados.

1.3 Relevância

Quando aplicada em uma empresa, a mineração de dados melhora a interação entre empresa e cliente, aumenta vendas e dirige as estratégias de marketing. A mineração de dados, porém, pode ser aplicada a qualquer massa de dados, sejam eles oriundos da Medicina, Economia, Astronomia, Geologia, entre outras áreas de estudo. A relevância deste trabalho fundamenta-se na importância da adoção de técnicas de mineração de dados para melhorar a qualidade de dados em um SGBD, como parte do trabalho de Administração de Dados.

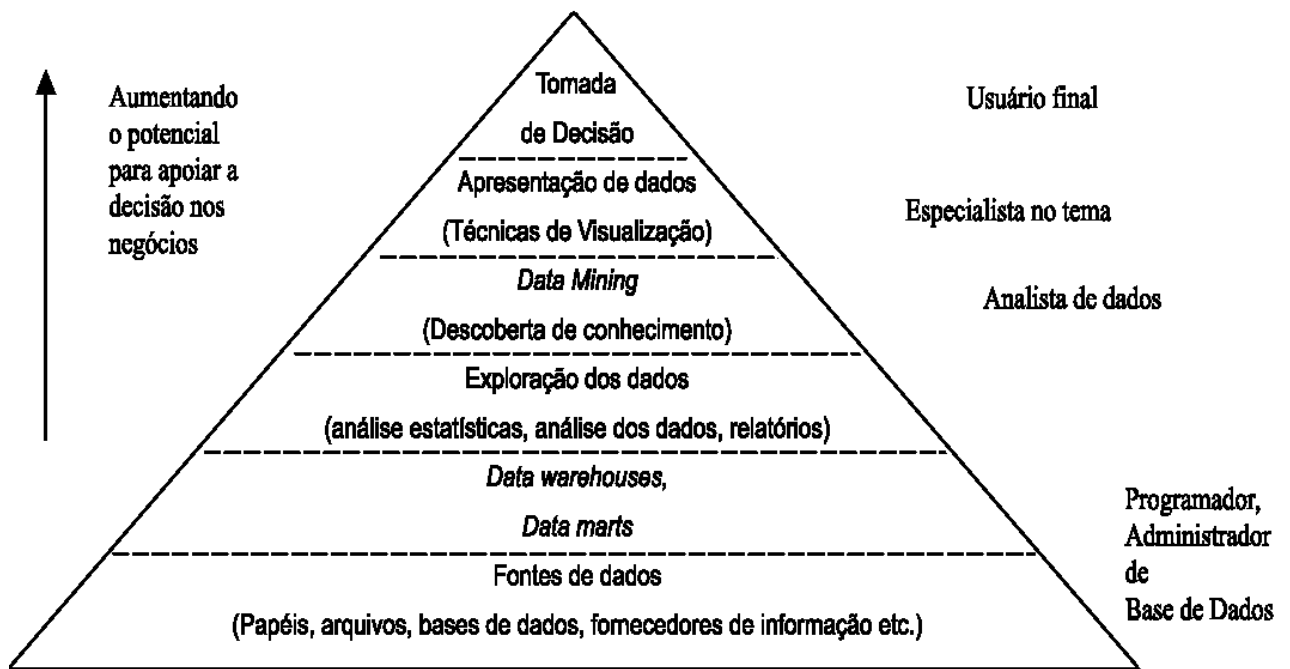


Figura 1: Evolução do valor estratégico de bases de dados (Baseado em [Navega, 2002]).

A Figura 1 mostra o posicionamento lógico de diferentes fases da tomada de decisão com seu valor potencial para as dimensões tática e estratégica de uma organização. Em geral, o valor da informação para apoiar a tomada de decisão aumenta a partir da base da pirâmide. Uma decisão baseada em dados nas camadas mais baixas, onde há tipicamente milhões de registros de dados, não possui muito valor agregado; já aquela apoiada em dados altamente resumidos nas camadas superiores da pirâmide tem probabilidade de alto valor estratégico.

Da mesma forma, encontram-se diferentes usuários nas diferentes camadas. Um administrador, por exemplo, no nível operacional, trabalha primariamente com informações diárias e operações de rotina, encontradas em arquivos e bases de dados, na base da pirâmide informacional. Esses criam dados. Enquanto analistas de negócios e executivos, responsáveis por indicarem direções, formulam estratégias e táticas, supervisionando a sua execução, e estes necessitam de informações de maior qualidade. Preocupam-se com tendências, padrões, ameaças, pontos fortes e

fracos, oportunidades, informação de mercado, entre outros. Necessitam de informações internas e externas. São os que demandam dados analisados com alto valor agregado, as do topo da pirâmide.

1.4 Metodologia de Trabalho

A metodologia utilizada no desenvolvimento deste trabalho se constitui das seguintes fases:

1. **Revisão bibliográfica dos conceitos e etapas do processo de mineração de dados** – buscou-se estudar os conceitos da mineração de dados e o aprofundamento em cada fase do processo;
2. **Revisão bibliográfica das técnicas relacionadas ao processo de mineração de dados** – esta atividade se constituiu na análise de diferentes técnicas da mineração de dados, bem como em um estudo comparativo da relevância e utilidade de cada técnica;
3. **Estudo da utilidade da mineração de dados em diversas áreas de conhecimento da sociedade** – esta etapa do estudo visou analisar as formas que a mineração de dados vem agregando valor a diferentes áreas de nossa sociedade, bem como indicar novas possibilidades de uso da mineração de dados em outras áreas; e
4. **Estudo das ferramentas que auxiliam na mineração de dados** – procurou-se pesquisar e estudar ferramentas que auxiliam no processo de mineração de dados como forma de melhorar a qualidade dos dados em um SGBD.

1.5 Organização da Monografia

Além deste capítulo, esta monografia encontra-se organizada em mais 5 (cinco) capítulos:

O Capítulo 2 irá detalhar todo o processo de mineração de dados, destacando cada fase desse processo, desde a seleção dos dados e sua filtragem, aos algoritmos e técnicas utilizadas para o reconhecimento de padrões.

O capítulo 3 mostrará o uso do processo de mineração de dados em diferentes áreas de conhecimento, destacando o quanto a prática da administração de dados focando na melhoria da qualidade dos mesmos, pode agregar de valor para uma empresa e em particular para a sociedade.

O capítulo 4 apresentará um estudo comparativo de ferramentas de mineração de dados disponíveis atualmente, mostrando as funcionalidades e características de cada uma delas, como também as técnicas que elas utilizam.

O Capítulo 5 irá apresentar uma conclusão para o trabalho, destacando as contribuições providas pelo mesmo, como também apresentando sugestões de possíveis trabalhos futuros.

Finalmente, serão listadas as referências bibliográficas utilizadas neste trabalho.

2 Mineração de Dados

A Mineração de Dados pode ser definida como um conjunto de técnicas automáticas de exploração de grandes massas de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano. De fato, muitas são as técnicas utilizadas, porém a mineração de dados ainda é mais uma arte do que uma ciência. O sentimento do especialista não pode ser dispensado, mesmo que as mais sofisticadas técnicas sejam utilizadas.

Ainda que as técnicas da Mineração de Dados sejam antigas, foi apenas nos últimos anos que passaram a ser usadas como exploração de dados, por vários motivos [Carvalho, 2005]:

- **O volume de dados disponível atualmente é enorme** – Mineração de Dados é uma técnica que só se aplica a grandes massas de dados, pois necessita disto para calibrar seus algoritmos e extrair dos dados conclusões confiáveis. Empresas de telefonia, cartões de crédito, bancos, televisão por assinatura, comércio eletrônico, entre outras, vem gerando a cada dia uma grande quantidade de dados sobre seus serviços e clientes. Estes dados são passíveis de análise por mineração;
- **Os dados estão sendo organizados** - Com a tecnologia do dataware house³, os dados de várias fontes estão sendo organizados e padronizados de forma a possibilitar sua organização dirigida para o auxílio à decisão. As técnicas de

³ De acordo com o Wikipedia: Data Warehouse é uma coleção de dados orientados por assuntos, integrados, variáveis com o tempo e não voláteis, para dar suporte ao processo de tomada de decisão; Data Warehousing é um processo em andamento que aglutina dados de fontes heterogêneas, incluindo dados históricos e dados externos para atender à necessidade de consultas estruturadas e ad-hoc, relatórios analíticos e de suporte a decisão [Wikipedia, 2006].

mineração de dados necessitam de bancos de dados limpos, padronizados e organizados;

- **Os recursos computacionais estão cada vez mais potentes** - A mineração de dados necessita de muitos recursos computacionais para operar seus algoritmos sobre grandes quantidades de dados. O aumento da potência computacional, devido ao avanço tecnológico e à queda dos preços dos computadores, facilita o uso da mineração de dados atualmente. O avanço da área de banco de dados, construindo bancos de dados distribuídos, também auxiliou em muito à mineração de dados;
- **A competição empresarial exige técnicas mais modernas de decisão** - As empresas da área de finanças, telecomunicações e seguro experimentam a cada dia mais competição. Como estas empresas sempre detiveram em seus bancos de dados uma enorme quantidade de informação, é natural que a mineração de dados tenha se iniciado dentro de seus limites. Atualmente, outras empresas buscam adquirir dados para analisar melhor seus caminhos futuros através dos sistemas de apoio à decisão. Para empresas de serviços, a aquisição de dados é importante, pois precisam saber que serviço oferecer a quem. Para outras empresas, até a venda das informações pode ser um produto; e
- **Programas comerciais de mineração de dados já podem ser adquiridos** - As técnicas de mineração de dados são antigas conhecidas da Inteligência Artificial, porém somente recentemente saíram dos laboratórios para as empresas. Alguns pacotes já podem ser encontrados no comércio, contendo algumas destas técnicas. As técnicas mais recentes, no entanto, ainda se encontram no campo acadêmico, sendo necessário que a empresa se dirija a uma universidade que realize pesquisa para obter ajuda.

2.1 Fases da Mineração de Dados

Em 1996, um conjunto de três empresas especializadas no então jovem e imaturo mercado de data mining, desenvolveram um modelo de processos genéricos, com o intuito de padronizar as etapas do processo de mineração de dados, dando início ao denominado projeto CRISP-DM (CRoss Industry Standard Process for Data Mining) [The CRISP-DM Consortium, 2000].

Este projeto desenvolveu um modelo de processo de mineração de dados industrial e livre de ferramenta. Começando pelos embrionários processos de descoberta de conhecimento usados nos primeiros projetos de mineração de dados e respondendo diretamente aos requerimentos do usuário, esse projeto definiu e validou um processo de mineração de dados que é aplicável em diversos setores da indústria. Essa metodologia torna projetos de mineração de dados de larga escala mais rápidos, mais baratos, mais confiáveis e mais gerenciáveis. Até mesmo projetos de mineração de dados de pequena escala se beneficiam com o uso do CRISP-DM. O modelo CRISP, atualmente, é uma referência para que seja desenvolvido um plano de integração para a descoberta de conhecimento.

O atual processo para mineração de dados propõe uma visão geral do ciclo de vida de um projeto de mineração de dados. Ele contém as fases correspondentes de um projeto, suas respectivas tarefas e relacionamentos entre essas tarefas.

Na Figura 2 é mostrado o ciclo de vida de um projeto de mineração de dados, que consiste de 6 (seis) fases. A seqüência de fases não é obrigatória, ocorrendo a transição para diferentes fases, dependendo do resultado de cada fase, e que etapa particular de cada fase precisa ser executada em seguida. As setas indicam as mais importantes e mais freqüentes dependências entre as fases.

O ciclo externo na figura simboliza o ciclo natural da mineração de dados. Um processo de mineração de dados continua após a solução ter sido desenvolvida. As

lições aprendidas durante o processo podem provocar perguntas novas, freqüentemente mais pertinentes ao negócio. Processos subsequentes se beneficiarão das experiências de processos anteriores.

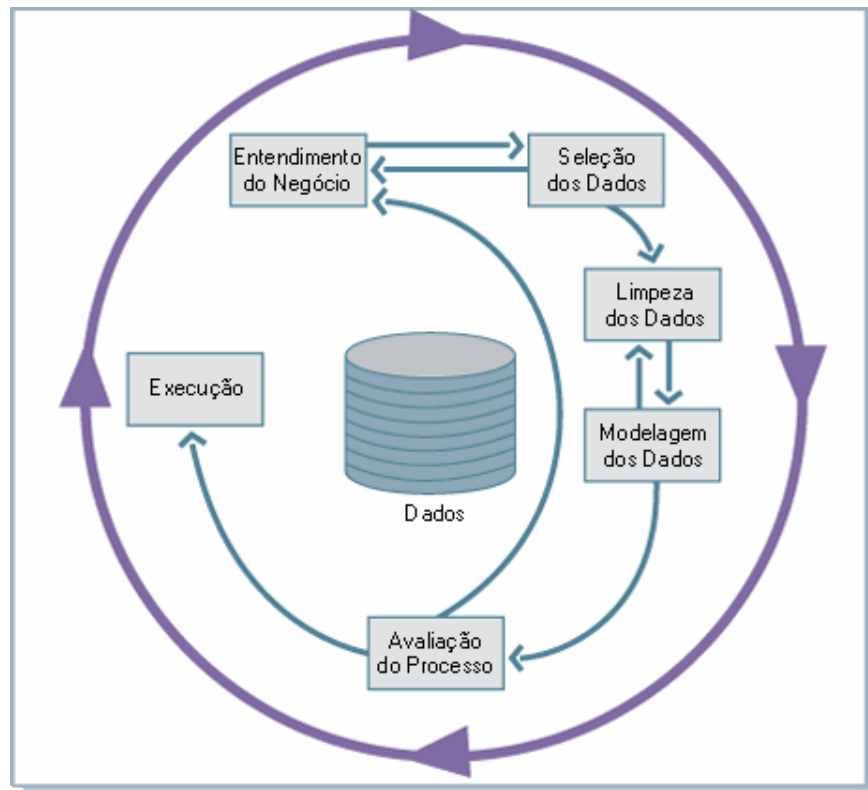


Figura 2: Fases do CRISP-DM Process Model (Baseado em [The CRISP-DM Consortium, 2000])

Segue abaixo uma síntese das etapas pertencentes ao modelo CRISP [The CRISP-DM Consortium, 2000]:

2.1.1 Entendimento do Negócio (Business Understanding)

Essa fase inicial tem o foco no entendimento do negócio que visa obter conhecimento sobre os objetivos do negócio e seus requisitos, e então converter esse conhecimento em uma definição de um problema de mineração de dados, e um plano preliminar designado para alcançar esses objetivos.

2.1.2 Seleção dos Dados (Data Understanding)

Consiste no entendimento dos dados, que visa à familiarização com o banco de dados pelo grupo de projeto, utilizando-se de conjuntos de dados "modelo". Uma vez definido o domínio sobre o qual se pretende executar o processo de descoberta, o próximo passo é selecionar e coletar o conjunto de dados ou variáveis necessárias.

Essa fase se inicia com uma coleta inicial de dados, e com procedimentos e atividades visando a familiarização com os dados, para identificar possíveis problemas de qualidade, ou detectar subconjuntos interessantes para formar hipóteses.

2.1.3 Limpeza dos Dados (Data Preparation)

A fase de preparação de dados consiste na preparação dos dados que visa a limpeza, transformação, integração e formatação dos dados da etapa anterior. É a atividade pela qual os ruídos, dados estranhos ou inconsistentes são tratados. Esta fase abrange todas as atividades para construir o conjunto de dados final (dados que serão alimentados nas ferramentas de mineração), a partir do conjunto de dados inicial.

A utilização de Data Warehouses facilita em muito esta etapa do processo de mineração de dados, que costuma ser a fase que exige mais esforço, correspondendo geralmente a mais de 50% do trabalho. Por isso, é muito importante para uma organização, que ela possua em seus processos habituais boas práticas da administração de dados, como o Data Cleansing, que é uma parte fundamental da cadeia da administração da informação, responsável pelas etapas de detecção, validação e correção de erros em bases de dados [Chapman, 2005].

2.1.4 Modelagem dos Dados (Modeling)

Fase que consiste na modelagem dos dados, a qual visa a aplicação de técnicas de modelagem sobre o conjunto de dados preparado na etapa anterior.

Nessa fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para se obter valores otimizados. Geralmente, existem várias técnicas para o mesmo tipo de problema de mineração. Algumas técnicas possuem requerimentos específicos na forma dos dados. Conseqüentemente, voltar para a etapa de preparação de dados é freqüentemente necessário.

A maioria das técnicas de mineração de dados são baseadas em conceitos de aprendizagem de máquina, reconhecimento de padrões, estatística, classificação e clusterização.

2.1.5 Avaliação do processo (Evaluation)

A avaliação do processo visa garantir que o modelo gerado atenda às expectativas da organização. Os resultados do processo de descoberta do conhecimento podem ser mostrados de diversas formas. Porém, estas formas devem possibilitar uma análise criteriosa para identificar a necessidade de retornar a qualquer um dos estágios anteriores do processo de mineração.

Nesta etapa se construiu um modelo que parece de alta qualidade, de uma perspectiva da análise de dados. Antes de prosseguir, é importante avaliar mais detalhadamente o modelo, e rever as etapas executadas para construir o modelo, para se certificar de que ele conseguirá alcançar os objetivos de negócio.

Deve se determinar se houve algum importante objetivo do negócio que não foi suficientemente alcançado. No fim desta fase, uma decisão sobre o uso dos resultados da mineração deve ser tomada.

2.1.6 Execução (Deployment)

Esta fase consiste na definição das fases de implantação do projeto de Mineração de Dados.

A criação do modelo não é o fim do projeto. Mesmo se a finalidade do modelo for apenas aumentar o conhecimento dos dados, o conhecimento ganho necessitará ser organizado e apresentado em uma maneira que o cliente possa usar. Dependendo das exigências, a fase de execução pode ser tão simples quanto a geração de um relatório, ou tão complexo quanto executar processos de mineração de dados repetidamente.

Em muitos casos será o cliente, não o analista dos dados, que realizará as etapas da execução. Entretanto, mesmo se o analista não se encarregar da execução é importante que ele faça o cliente compreender que medidas deverão ser tomadas a fim de empregar efetivamente os modelos criados.

2.2 Técnicas

Existem 5 (cinco) técnicas gerais de mineração de dados que englobam todas as outras formas de apresentação e permitem uma visão mais global e apropriada ao assunto. São elas a classificação, a estimativa, a previsão, a análise de afinidades e a análise de agrupamentos [Carvalho, 2005].

2.2.1 Classificação

A classificação é uma das mais utilizadas técnicas de mineração de dados, simplesmente porque é uma das mais realizadas tarefas humana no auxílio à compreensão do ambiente em que se vive. O ser humano está sempre classificando o que percebe a sua volta, criando classes de relações humanas diferentes (colegas de trabalho, amigos, familiares, por exemplo...) e dando a cada classe uma forma diferente de tratamento.

A classificação pode ser sintetizada por um processo de discriminação de unidades em classes ou categorias. Assim, classificam-se sabores, amigos, clientes, eventos, entre outros, em categorias, tais como doce / salgado / neutro, bom / mau e legal / ilegal.

Em um processo de mineração de dados, a classificação está especificamente voltada à atribuição de uma das classes pré-definidas pelo analista a novos fatos ou objetos submetidos à classificação. Essa técnica pode ser utilizada tanto para entender dados existentes quanto para prever como novos dados irão se comportar [Euriditionhome, 2004].

Como no mundo físico nada é exatamente igual, por mais semelhante que pareça, para se criar classes é preciso permitir que detalhes sejam desprezados e somente as características principais sejam observadas. A tarefa de classificar geralmente exige a comparação de um objeto ou dado com outros dados ou objetos que supostamente pertençam a classes anteriormente definidas. Para comparar dados ou objetos utiliza-se uma métrica ou forma de medida de diferenças entre eles.

Na mineração de dados são comuns as tarefas de classificação de clientes em baixo, médio ou alto risco de empréstimo bancário; de clientes potencialmente consumidores de um determinado produto a julgar pelo seu perfil; de transações financeiras como legais, ilegais ou suspeitas em sistemas de fiscalização do

mercado financeiro; de ações da bolsa de valores com lucros potenciais baixos, médios e altos, entre outras.

Os algoritmos mais utilizados para este fim são os de árvores de decisão [Pelegri et al., 2005], regressão [Han et al., 2001] e redes neurais [Sousa, 1998].

2.2.2 Estimativa

A estimativa, ao contrário da classificação, está associada a respostas contínuas.

Estimar algum índice é determinar seu valor mais provável diante de dados do passado ou de dados de outros índices semelhantes sobre os quais se tem conhecimento.

Suponha que se deseja determinar o gasto de famílias cariocas com lazer e que para isto se possua índices de gastos de famílias paulistanas com lazer, em função da faixa etária e padrão sócio-cultural. Não se sabe exatamente quanto as famílias cariocas gastam com lazer mas se pode estimar baseando-se nos dados das famílias paulistanas. Certamente que esta estimativa pode levar a grandes erros, uma vez que Rio de Janeiro e São Paulo são cidades com geografias diferentes e que oferecem diferentes opções de lazer a seus habitantes.

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

Os algoritmos de regressão e as redes neurais são bastante utilizados nestes casos.

2.2.3 Previsão

A previsão, como tarefa típica de DM, está associada à avaliação de um valor futuro de uma variável a partir dos dados históricos do seu comportamento passado. Assim, pode-se prever, por exemplo, se o índice bovespa subirá ou descenderá no dia seguinte; qual será o valor de determinada ação daqui a um determinado período de tempo; o número de clientes que serão perdidos por uma empresa, em um dado horizonte futuro de tempo; qual será a população de uma certa cidade daqui a dez anos; entre outras coisas.

A única maneira de avaliar se a previsão foi bem feita é aguardar o acontecimento e verificar o quanto foi acertada ou não a previsão realizada. Sem dúvida, a previsão é uma das tarefas mais difíceis não somente na mineração de dados, mas também no cotidiano das pessoas.

Os algoritmos que podem ser utilizados aqui são, dentre outros, as redes neurais, a regressão, e as árvores de decisão.

2.2.4 Análise de Afinidades

A análise de afinidades preocupa-se em reconhecer padrões de ocorrência simultânea de determinados eventos nos dados em análise. Determinar que fatos ocorrem simultaneamente com probabilidade razoável (co-ocorrência) ou que itens de uma massa de dados estão presentes juntos com uma certa chance (correlação).

O exemplo mais clássico de análise de afinidades é o do carrinho de supermercado, do qual deseja-se conhecer quais os produtos que são comumente comprados em conjunto pelos consumidores. Isto possibilita a otimização do layout interno dos supermercados e a realização de vendas dirigidas nas quais os itens são oferecidos já em conjuntos com preços menores.

Em termos de algoritmos, a utilização das regras de associação constitui-se no procedimento mais utilizado nestes casos [Pelegrin et al., 2005].

2.2.5 Análise de agrupamentos

A análise de agrupamentos visa formar grupos de objetos ou elementos mais homogêneos entre si. Pode ser estabelecido previamente um número de grupos a ser formado, ou então se pode admitir ao algoritmo de agrupamento uma livre associação de unidades, de forma que a quantidade de grupos resultante seja conhecida somente ao final do processo.

Uma clara diferença entre agrupamento e classificação é que na classificação as classes são pré-definidas pelo pesquisador, enquanto que aqui não existe tal requisito. Isto torna esta técnica muito mais complexa do que a classificação. Por exemplo, dadas as classes animal, vegetal e mineral, é relativamente simples classificar a qual dessas classes um certo objeto pertence, porém de posse de uma massa de dados sobre o consumo no Brasil, determinar quantas classes ou padrões de comportamento consumista existem é algo bem diferente. A dificuldade reside no fato de que podem não haver tais classes, ou seja, os dados se distribuem igualmente por todo o espaço possível não determinando nenhuma categoria.

Na análise de agrupamentos, os grupos ou classes são construídos com base na semelhança entre os elementos, cabendo ao analisador das classes resultantes avaliar se estas significam algo útil. Por exemplo, agrupar sintomas pode gerar classes que não representem nenhuma doença explicitamente, uma vez que doenças diferentes podem possuir os mesmos sintomas.

A análise de agrupamentos é normalmente uma técnica preliminar, utilizada quando nada ou pouco se sabe sobre os dados. Segmentar um mercado é uma típica análise de agrupamentos onde consumidores são reunidos em classes representantes dos segmentos deste mercado.

Em geral, a técnica de agrupamento é executada por algoritmos estatísticos específicos para esse fim, porém as redes neurais e os algoritmos genéticos [Han et al., 2001] são também utilizados neste sentido.

2.3 Tipos de Algoritmos

Diversos tipos de algoritmos são utilizados nas diferentes técnicas da mineração de dados. Por serem os mais amplamente utilizados, dois (2) dos principais tipos de algoritmos são brevemente descritos a seguir.

2.3.1 Árvores de Decisão

O método de árvores de decisão representa um tipo de algoritmo de aprendizado de máquina que utiliza uma abordagem dividir-para-conquistar. Consiste em meios de representar resultados de mineração de dados na forma de árvore. Neste método, permite-se ao usuário definir o objeto de saída. Com isto, a partir de um grupo de dados é possível identificar o fator mais importante correlacionado a este objeto. Dado um grupo de dados com numerosas colunas e linhas, uma ferramenta de árvore de decisão pede ao usuário para escolher uma das colunas como objeto de saída, e mostra o único e mais importante fator correlacionado com aquele objeto de saída como o primeiro ramo (nó) da árvore de decisão [Gimenes, 2000].

O conhecimento representado nas árvores de decisão pode ser extraído e representado na forma de regras de classificação do tipo SE-ENTÃO. Sendo assim,

cada nó filho da árvore representa uma condição (antecedente) envolvendo um atributo e um conjunto de valores. Enquanto os nós folhas, correspondem às conclusões (consequente), que indicam a atribuição de um valor ou conjunto de valores a um atributo do problema [Pelegri et al., 2005].

O principal problema relativo a essas árvores é que elas precisam de uma quantidade de dados considerável para desvendar estruturas complexas. Por outro lado, elas podem ser construídas de forma consideravelmente mais rápida do que alguns métodos alternativos de classificação, produzindo resultados com precisão similar [Sousa, 1998].

2.3.2 Redes Neurais

As redes neurais são sistemas paralelos distribuídos, que tentam construir representações de modelos ou padrões achados nos dados. Estruturalmente, uma rede neural consiste em um número de unidades de processamento simples interconectadas (chamadas neurônios), que têm o objetivo de calcular determinadas funções matemáticas (funções de ativação). Os neurônios são dispostos em uma ou mais camadas e interligados por um grande número de conexões. Essas conexões estão associadas a pesos que armazenam o conhecimento representado no modelo e ponderam as entradas recebidas por cada neurônio da rede [Sousa, 1998].

A aquisição de conhecimento a partir do ambiente é feita através de um processo de aprendizagem (treinamento). Na fase de treinamento, os pesos das conexões da rede vão sendo ajustados de forma que o conhecimento extraído dos dados possa ser representado internamente, através de interações repetidas, cada hora ajustando os parâmetros que definem a superfície. Depois de muitas repetições, uma superfície que se aproxima muito dos pontos dentro do grupo de dados pode ser internamente definida [Gimenes, 2000].

Dentre as vantagens dos algoritmos baseados em redes neurais, está a sua robustez ao lidar com erros no conjunto de treinamento, possibilitando uma alta tolerância a dados com ruídos. Também possui boa escalabilidade e como diversos algoritmos têm sido desenvolvidos para extração de regras de classificação de redes neurais, a sua interpretabilidade tem melhorado [Gouvêa, 2005; Sousa, 1998].

O seu ponto fraco está na necessidade de definição de muitos parâmetros como a sua estrutura e valores iniciais dos pesos, além de longos tempos de treinamento. Algoritmos de redes neurais tipicamente necessitam de maior força computacional que, por exemplo, algoritmos de árvores de decisão. Redes neurais aprendidas não são facilmente compreensíveis para pessoas, o que não ocorre com regras, as quais são facilmente interpretadas [Gouvêa, 2005; Sousa, 1998].

3 Aplicações Práticas

Segundo estimativas do The Data Warehousing Institute [TDWI, 2006], a má qualidade sobre os dados dos clientes, custa, só nos Estados Unidos, 611 bilhões de dólares por ano. A realidade é que dados mal-administrados causam mais prejuízos ainda que isso. Dado a importância de negócio atrelada à administração de dados, focando na qualidade dos mesmos, é impressionante a forma causal com a qual muitas empresas encaram e administram este recurso.

Neste capítulo serão mostrados alguns exemplos de empresas de diferentes áreas da sociedade que investiram na administração de dados, focando na melhoria de sua qualidade e geração de conhecimento a partir dos mesmos, e obtiveram retorno empresarial.

3.1 Segmentação de Mercados

Um dos grandes objetivos de uma organização é conhecer seus clientes. Este conhecimento precisa ocorrer em vários níveis, desde o tipo de produto desejado até que tipo de ofertas estão dispostos a aceitar mesmo que os produtos não sejam essenciais no momento. Também é interessante saber o perfil médio do consumidor, sua renda, sexo, idade, tamanho da família, entre outros aspectos. Com estas informações, a empresa poderá ter em estoque o que o cliente médio mais provavelmente precise e fazer ofertas com certo grau de certeza do seu sucesso a clientes específicos. Esta forma de mercado dirigido pode atingir o extremo de uma relação individual com cada cliente à medida que a empresa deseje investir em segmentações (classificações) sucessivas de sua clientela.

Esta típica tarefa de mineração de dados é usada por grandes lojas de departamentos e administradoras de cartão de crédito e se utiliza dos dados das compras dos clientes no passado recente para traçar perfis de consumo. Informações como idade,

sexo, estado civil, salário, moradia própria ou alugada, bairro e cidade também são informações importantes pois permitem a setorização ainda mais fina dos clientes. Se os produtos comprados são avaliados de alguma forma quanto à satisfação do cliente através de pesquisas por telefone ou Internet por exemplo, um quadro ainda maior pode ser traçado.

Em uma determinada rede de lojas, escolheu-se apenas 5 (cinco) categorias de produtos para minerar e melhorar as vendas: vestuário esportivo, aparelhos de ginástica, decoração, mobília e CD. A idéia da empresa era fazer um maior direcionamento à clientela quando fosse enviar catálogos e ofertas.

A rede de lojas possui 3 (três) bases de dados, contendo uma delas os dados dos clientes para fins de crediário e verificação de residência, cheques, entre outros; outra base contendo os dados sobre produtos em estoque, seus preços e descrições; e a terceira base contendo os dados sobre as compras executadas. Uma empresa de data warehouse foi contratada para unificar estes 3 (três) bancos de dados e permitir análises globais dos seus negócios. Após a construção do data warehouse, foi fácil agregar dados sobre os clientes que realizam compras de produtos das 5 (cinco) categorias.

A empresa decidiu pela utilização de redes neurais para a análise de agrupamentos – e no fundo a segmentação de mercados nada mais é que uma análise de agrupamentos. Uma média dos dados de todos os clientes foi realizada para efeitos de comparação com o perfil médio dos clientes de cada agrupamento. Em média, os consumidores das 5 (cinco) categorias de produtos escolhidos pela rede de lojas para análise têm 42 anos, ganham US\$ 35.000,00 por ano, são em maioria (58%) do sexo feminino e 50% casados contra 35% solteiros, possuindo 40% deles casa própria. Estes clientes gastaram em média no último ano US\$ 500,00 em vestuário esportivo, US\$ 1.000,00 em equipamentos de ginástica, US\$ 1.250,00 em decoração, US\$ 780,00 em CD e US\$ 1.100,00 em mobília.

A rede neural criou um grupo de clientes contendo 48% do total com um perfil bastante semelhante ao perfil médio descrito acima, com exceção de gastarem o dobro em decoração. O segundo maior grupo possui 29% dos clientes e tem 52 anos em média, gastando a metade que a média total em esportes e aproximadamente US\$ 500,00 a mais da média em mobília. O terceiro grupo possui 20% dos clientes e consome aproximadamente o mesmo que a média geral com a diferença de gastar apenas um quarto da média em decoração. O último grupo tem apenas 26 anos de idade em média e representa 11% do total gastando o dobro da média em esportes.

Com base nesses resultados a empresa decidiu dividir sua mala-direta em grupos com diferentes ofertas: Em artigos de decoração para o primeiro grupo, em mobília para o segundo segmento de clientes, e em material esportivo para o quarto perfil de clientela. Certamente que os analistas de mineração de dados poderiam ter caminhado um pouco além e procurado dividir os clientes em mais subgrupos de forma a permitir uma visão mais específica. Porém, os quatro grupos descritos foram suficientes para as necessidades da empresa naquele momento [Carvalho, 2005].

O processo de análise de agrupamentos pode ser facilmente estendido para dividir as, aproximadamente, mil filiais da rede de lojas em todo os EUA em categorias e facilitar a distribuição de produtos, assim como a realização de compras nos fornecedores e indústrias. Assim, ao invés de analisar o que cada uma das mil filiais está necessitando para seus estoques e montar uma estrutura logística enorme, pode-se agrupar lojas componentes de cada um dos grupos. O gerenciamento da rede de lojas fica, então, bastante facilitado.

3.2 Varejo

Vários fatores podem contribuir para a necessidade de previsão de vendas tais como a manutenção do cliente que não se frustra ao encontrar na loja o que deseja, o menor custo com estoques pela manutenção de estoques mais ajustados às vendas futuras, a melhor alocação de vendedores em função da previsão das vendas para o futuro período, entre outros.

Os parâmetros importantes a serem considerados quando se analisa a disponibilidade de produtos em uma loja são a capacidade de produção e distribuição da indústria produtora do item, a existência ou não de propaganda realizada pelo produtor do item e o período do ano ou mês dependendo do produto tratado.

Uma grande revendedora de automóveis de vários fabricantes nos seus diversos modelos observando sua perda de venda e de clientes a cada vez que não possuía o desejado carro em seus estoques e, contrapondo este fato com o alto custo de manutenção de grandes estoques deste produto durável e caro, resolveu desenvolver um sistema de previsão de vendas. A empresa possuía um banco de dados de vendas de carros nos últimos 5 (cinco) anos e desejava um sistema de previsão capaz de avaliar as vendas 15 dias a frente pois este era o tempo necessário para encomenda e transporte de novos itens [Carvalho, 2005].

Além da informação contida no banco de dados, é necessário contextualizar cada dado de venda com outras informações como a existência de propaganda realizada pelo fabricante, se a venda foi realizada em certos períodos do ano mais propícios à compra de automóveis e também ao fim de cada mês quando há um natural

aquecimento das vendas. Como prever o futuro não é nada fácil, a maior quantidade de informação pertinente possível deve ser considerada em qualquer metodologia.

Foi escolhido pelo uso de uma rede neural. O treinamento da rede neural foi feito com dados de quatro anos e meio deixando os últimos seis meses do período de 5 (cinco) anos de vendas para a testagem da capacidade de previsão do sistema. O aprendizado mostrou-se eficiente tendo um erro máximo de previsão em algumas semanas de 20%, porém o erro médio se manteve dentro dos desejados 10%.

Desta forma, o sistema passou a prever as vendas dos próximos 15 dias fornecendo mais tempo para a encomenda e transporte do produto. A cada quatro semanas, a rede neural era ensinada de novo, incluindo-se os dados de mais 4 (quatro) semanas ocorridas seis meses antes e testando-se o erro de previsão utilizando-se sempre os últimos 6 (seis) meses de vendas, agora incluindo as últimas quatro semanas recentemente terminadas.

3.3 Mercado Financeiro

A empresa norte-americana LBS já há algum tempo (desde 1986) investe em tecnologia por achar que os enfoques tradicionais no gerenciamento de investimentos não fornecem resultados que superem expressivamente o lucro médio de mercado. A empresa acredita que novas técnicas podem capturar relações de causa e efeito não lineares comuns no funcionamento do mercado financeiro.

Com a utilização de sistemas de previsão a LBS esperava aumentar o retorno e minimizar os riscos de perda. Seu desejo era poder, baseado em sinais de hoje, prever as tendências do mercado após um determinado tempo, comprando e vendendo os papéis corretos, nos tempos mais apropriados. A empresa tenta extrair dos dados passados e presentes alguma visão das variações futuras.

Prever o mercado financeiro não é tarefa fácil, pois se no curto prazo os boatos e as guerras psicológicas alteram preços muito rapidamente, no longo prazo ocorrem mudanças políticas nacionais e internacionais assim como catástrofes e outros fatos que alteram completamente as tendências. As relações micro e macro-econômicas não são bem entendidas, sendo muito comum diferentes explicações para eventos já ocorridos. A massa de dados a ser analisada é muito grande e a evolução das relações com o tempo pode ser muito rápida, requerendo sistemas de análise adaptativos.

A LBS concluiu que a melhor solução seria o uso de redes neurais. Decidiu construir uma rede neural para cada papel ao invés de uma única rede neural para os milhares de papéis existentes. Cada uma delas foi treinada com dados dos últimos 3 (três) meses e a cada nova semana era re-treinada. As redes neurais se mostraram muito adaptáveis às flutuações do mercado financeiro. Vários experimentos se fizeram necessários para a determinação das variáveis importantes para a previsão das tendências de cada papel específico.

De fato, a LBS foi uma das mais importantes empresas do mercado financeiro nos EUA durante 7 (sete) anos consecutivos. Sua carteira cresceu durante este período de 25% a 100% ao ano e nunca houve uma perda maior que 7% durante este tempo. Se a mineração dos dados foi a responsável por este sucesso não se pode afirmar, mas certamente alguma importância tiveram, pois recursos consideráveis foram destinados a este projeto [Carvalho, 2005].

4 Análise de Ferramentas

Este capítulo tem como objetivo analisar 3 (três) ferramentas disponíveis no mercado que auxiliam no processo de mineração de dados, descrevendo uma visão geral das mesmas e detalhando suas principais funcionalidades. Também serão apresentadas algumas telas capturadas das ferramentas para melhor ilustrar a demonstração da mesma.

4.1 SAS Enterprise Miner

Presente em 110 países, com mais de 40 mil instalações, a SAS no Brasil é o parceiro de negócios líder de mercado na oferta de soluções para a gestão corporativa baseada em inteligência. Fornece propostas de valor que englobam soluções e serviços de consultoria para o gerenciamento e a previsibilidade dos negócios. O SAS mapeia, identifica e formula processos que oferecem soluções específicas para as corporações, tornando-as mais eficientes e competitivas [SAS, 2007].

Entre outros sistemas corporativos, a SAS possui uma ferramenta de mineração de dados chamada Enterprise Miner, cujo site da ferramenta [SAS, 2007] afirma ser um software de mineração de dados avançado, combinando um extenso pacote de ferramentas integradas de mineração de dados com uma facilidade de utilização nunca vista, capacitando os usuários a tirarem proveito de dados corporativos para obter vantagem estratégica, tudo em um único ambiente.

O site da ferramenta sugere que o Enterprise Miner se destina, dentre outros, a mineradores de dados, analistas de mercados, analistas de risco, investigadores de fraudes, engenheiros e cientistas que desempenham papéis importantes na resolução de problemas de negócio críticos ou pesquisas científicas.

O software integra diferentes técnicas da mineração de dados, sendo uma avançada ferramenta para predição e descrição de dados, utilizando diversos algoritmos incluindo algoritmos de árvores de decisão, algoritmos de redes neurais, entre outros.

Entre outras funcionalidades, o sistema também disponibiliza uma tela que demonstra graficamente todas as fases do processo de mineração de dados, desde o acesso aos dados, até a análise dos resultados, dando todo suporte necessário a cada etapa em uma única solução integrada. Na Figura 3 é mostrado o fluxo entre as etapas da mineração de dados utilizadas para a análise de um empréstimo.

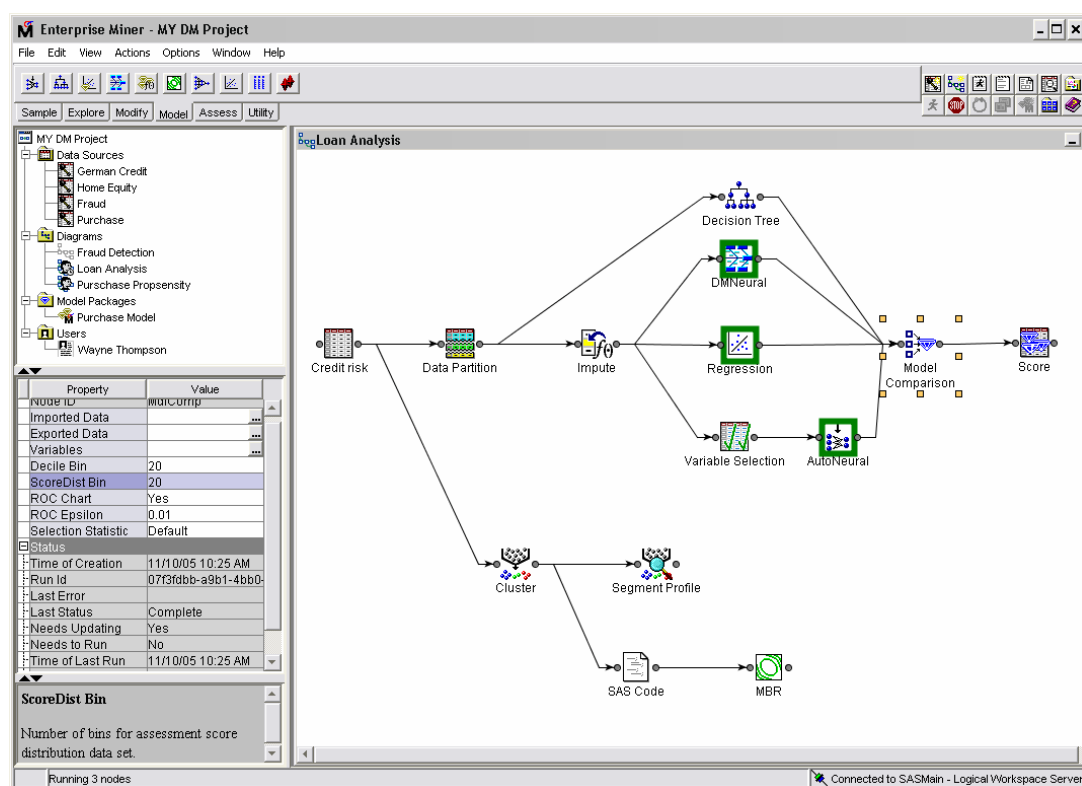


Figura 3: Fluxo da mineração de dados em análise de empréstimo [SAS, 2007].

O Enterprise Miner também possui funcionalidades ligadas ao pré-processamento dos dados. Relatórios com resumos dos resultados são gerados e a interface gráfica Java possui gráficos estatísticos flexíveis. Na Figura 4 é mostrada uma tela de relatório gerada a partir de uma análise de agrupamentos. O sistema provê visualizações para ajudar a determinar quais variáveis são mais importantes para distinguir os agrupamentos.

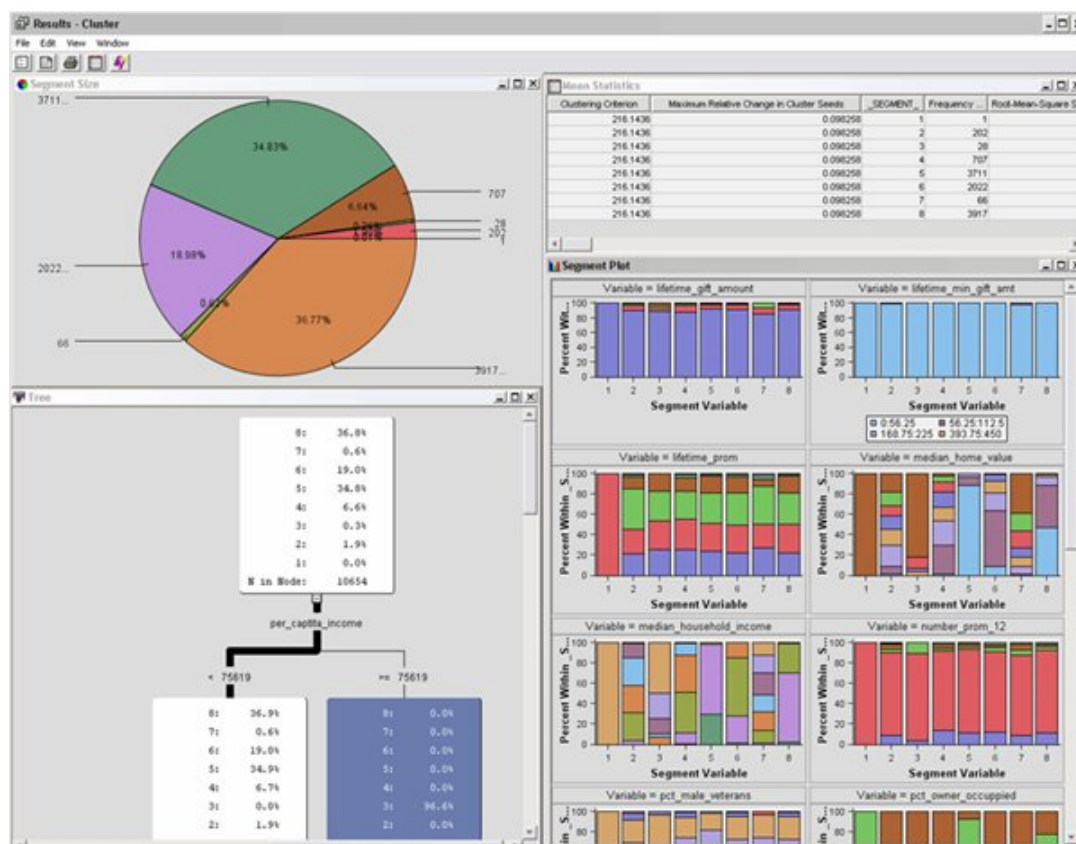


Figura 4: Análise de Agrupamentos [SAS, 2007].

Oferecendo um modelo de comparação integrada, utilizado para comparar os resultados de técnicas diferentes em visões de negócio como também em termos estatísticos, o software fornece a habilidade original de poder comparar a eficácia de cada técnica em relação ao total do conhecimento gerado, permitindo que estatísticos e analistas do negócio discutam facilmente os resultados, para analisar as técnicas mais eficazes, como mostrado na Figura 5.

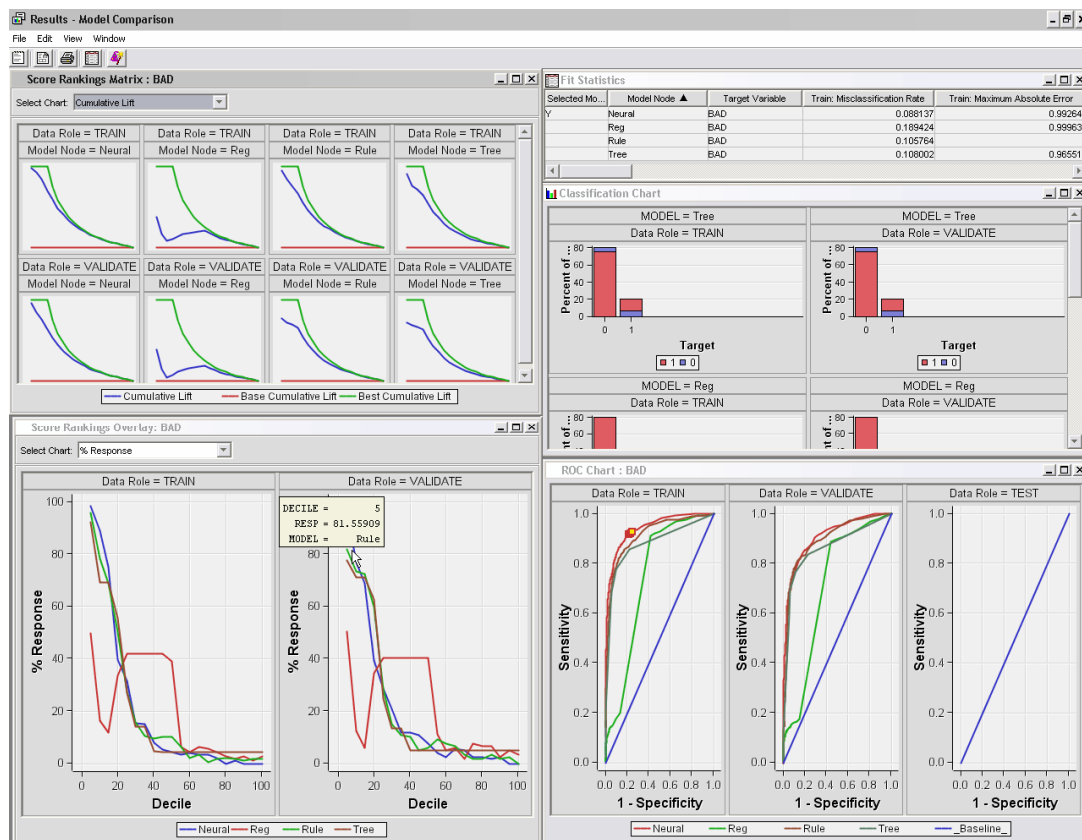


Figura 5: Comparação de técnicas [SAS, 2007].

4.2 IBM Intelligent Miner

O DB2 Data Warehouse Edition (DWE), da IBM, é uma suíte de produtos que combinam a administração de dados, com uma poderosa infra-estrutura de inteligência corporativa. O DWE pode ser usado para construir uma completa solução de data warehousing e gerenciá-la com uma solução única que integra componentes núcleos com funcionalidades distintas.

Entre os componentes que compõem o DWE está o Intelligent Miner, cujo site da ferramenta [IBM, 2007] o classifica com uma poderosa ferramenta para análise de dados integrada. As tradicionais técnicas de mineração de dados (análise de

agrupamentos, análise de afinidades, classificação, estimativa e previsão) são suportadas. Adicionalmente, ricos componentes de apresentação estão disponíveis para possibilitar uma análise visual dos resultados.

Na Figura 6, pode-se ver uma representação visual do resultado de uma análise de agrupamentos entre clientes, feita através da mineração da base de dados de um banco. A figura mostra que existem 4 (quatro) agrupamentos, o maior contendo 33,71% do total de clientes, e o menor contendo 13,56% do total.

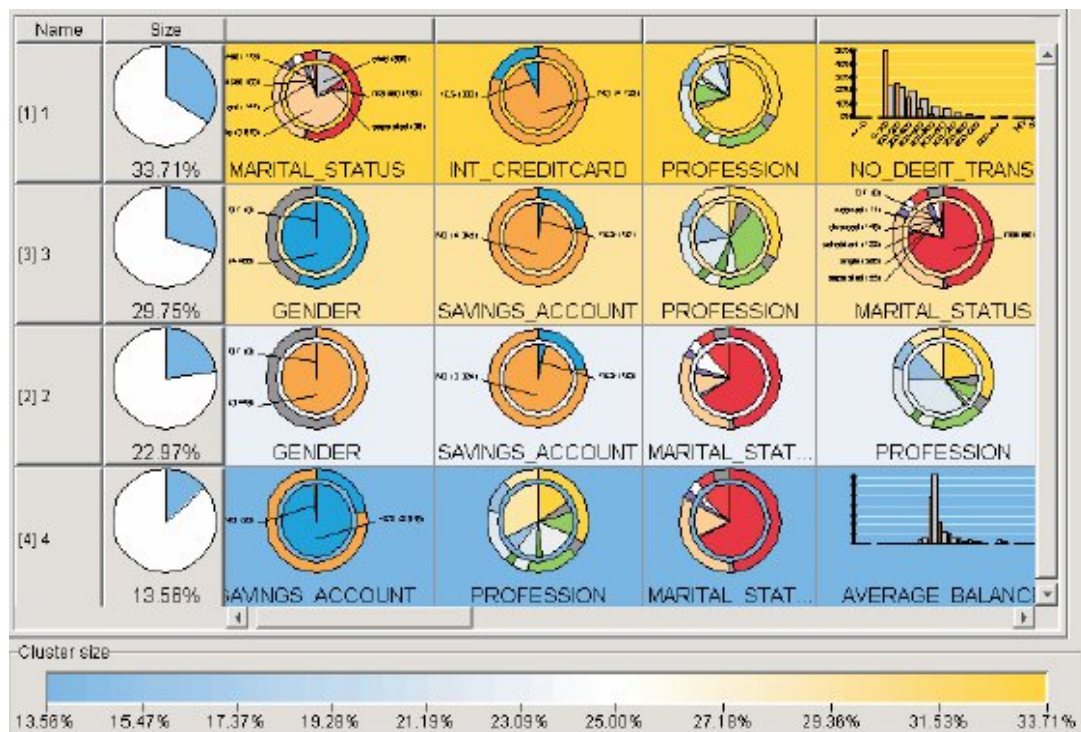


Figura 6: Análise de agrupamentos de clientes bancários [IBM, 2007].

Os gráficos de setores circulares e os gráficos de barras mostram a distribuição dos valores das colunas nos agrupamentos comparados com o total de clientes. Nos gráficos de setores circulares, o círculo interno representa a população de um agrupamento. O círculo externo representa a população total. Por exemplo, o gráfico de setores circulares INT_CREDITCARD na Figura 7 mostra que apenas

alguns clientes no agrupamento 1 possuem um cartão de crédito internacional comparados à quantidade total de clientes.

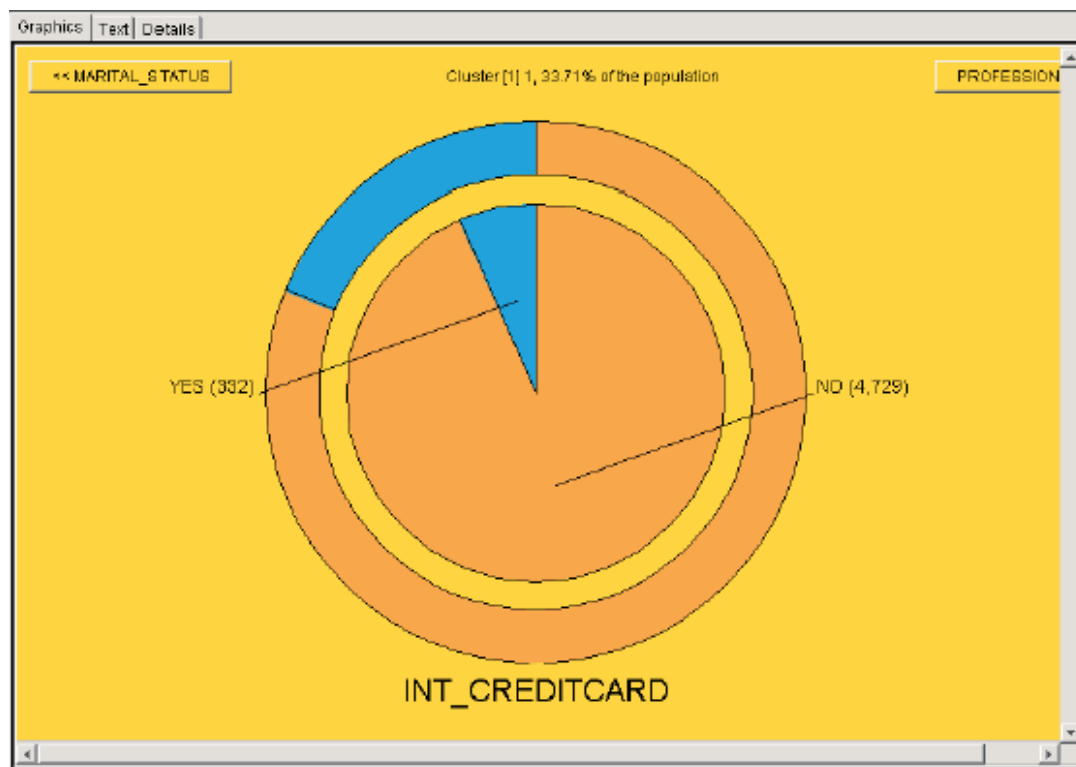


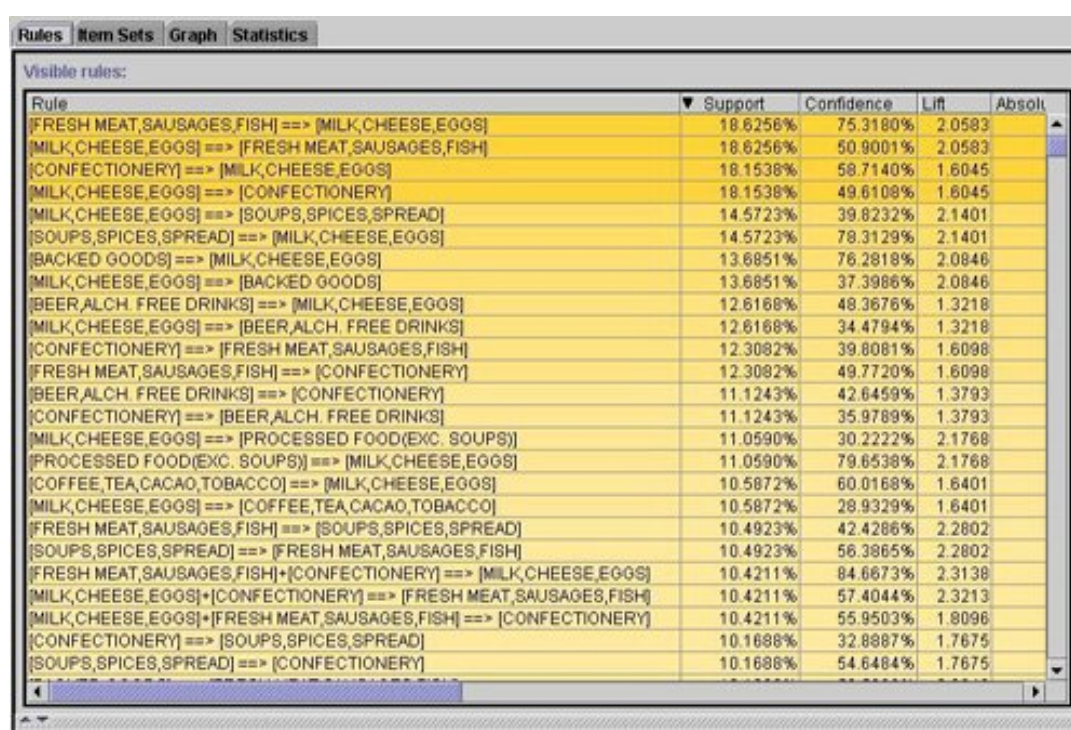
Figura 7: Gráfico de setores circulares INT_CREDITCARD [IBM, 2007].

CLIENT_ID	NBR	AVERAGE	INT	PROFESS	AGE	GEND	MARITAL	SAVING	NO_DEB	NO_CRE	CLUS	QUALITY	CONF
01044550	3	1,693.00	NO	inactive	24	F	single	NO	84	34	1	0.95	0.736
00926635	3	2,036.00	NO	employee	29	M	single	NO	84	34	1	0.837	0.633
01054766	2	80.00	NO	farmer	57	F	separated	NO	84	37	1	0.639	0.527
00710052	22	4,776.00	NO	pensioner	73	M	single	YES	84	37	4	0.752	0.54
00978861	16	6,580.00	YES	employee	34	M	married	NO	84	39	3	0.705	0.596
00054676	11	1,746.00	NO	inactive	19	F	single	YES	84	41	1	0.816	0.726
00112266	14	8,347.00	NO	pensioner	67	F	single	YES	84	45	4	0.701	0.504
00597861	24	63,084.00	YES	craftsman	65	M	single	NO	84	47	4	0.47	0.417
01277071	23	18,855.00	YES	inactive	72	F	widowed	YES	84	48	4	0.623	0.546
00363650	3	1,465.00	NO	inactive	20	F	single	NO	84	55	1	0.912	0.712
00040999	4	18,191.00	NO	inactive	83	F	widowed	NO	84	70	1	0.691	0.549
01261950	3	786.00	NO	employee	37	F	single	NO	85	12	1	0.783	0.598
01059830	2	5,039.00	NO	employee	23	M	single	NO	85	12	1	0.83	0.699
00814430	17	2,207.00	NO	intermedia	37	M	single	NO	85	13	1	0.722	0.623
00577519	11	3,443.00	YES	inactive	29	M	single	NO	85	13	1	0.793	0.663
00360629	4	1,950.00	NO	inactive	23	F	single	NO	85	16	1	0.97	0.771
01058354	2	6,105.00	NO	inactive	19	F	single	NO	85	15	1	0.97	0.775
00696890	29	10,619.00	NO	pensioner	84	F	widowed	NO	85	15	1	0.593	0.502
00592091	0	1,114.00	NO	inactive	23	F	single	NO	85	17	1	0.944	0.758
00147662	6	-435.00	NO	inactive	26	F	single	NO	85	18	1	0.945	0.739
00567620	26	3,935.00	YES	inactive	55	F	married	NO	85	20	2	0.647	0.549
00160392	5	5,041.00	NO	inactive	24	M	single	NO	85	21	1	0.981	0.77
01233651	17	20,697.00	YES	employee	43	F	single	YES	85	22	4	0.584	0.486
01136296	0	511.00	NO	inactive	21	F	single	NO	85	23	1	0.942	0.756
00363849	3	977.00	NO	inactive	21	F	single	YES	85	23	1	0.883	0.807
01209425	4	3,165.00	NO	inactive	33	M	married	YES	85	23	1	0.751	0.611
00662796	5	3,566.00	NO	executives	43	M	divorced	NO	85	23	1	0.687	0.523
00932852	5	3,960.00	NO	inactive	23	M	single	NO	85	25	1	0.979	0.769
01239446	2	3,723.00	NO	worker	39	M	married	NO	85	26	3	0.765	0.558
01233669	32	18,825.00	NO	pensioner	73	M	married	NO	85	27	3	0.675	0.577
01217691	19	964.00	NO	inactive	19	F	single	YES	85	28	1	0.795	0.699

Figura 8: Tabela de clientes, com os agrupamentos informados [IBM, 2007].

A Figura 8 mostra a tabela de clientes, exibindo em uma coluna a qual agrupamento gerado, cada cliente pertence.

Na Figura 9 é mostrado o exemplo de uma análise de afinidades, utilizada para resolver o tradicional problema do carrinho de compras de supermercado. Nela, é mostrada a co-ocorrência de alguns itens em uma mesma compra, e um percentual de probabilidade de os conjuntos de itens serem comprados simultaneamente.



Rule	Support	Confidence	Lift	Absol.
[FRESH MEAT, SAUSAGES, FISH] ==> [MILK, CHEESE, EGGS]	18.6256%	75.3180%	2.0583	
[MILK, CHEESE, EGGS] ==> [FRESH MEAT, SAUSAGES, FISH]	18.6256%	50.9001%	2.0583	
[CONFECTIONERY] ==> [MILK, CHEESE, EGGS]	18.1538%	58.7140%	1.6045	
[MILK, CHEESE, EGGS] ==> [CONFECTIONERY]	18.1538%	49.6108%	1.6045	
[MILK, CHEESE, EGGS] ==> [SOUPS, SPICES, SPREAD]	14.5723%	39.8232%	2.1401	
[SOUPS, SPICES, SPREAD] ==> [MILK, CHEESE, EGGS]	14.5723%	78.3129%	2.1401	
[BACKED GOODS] ==> [MILK, CHEESE, EGGS]	13.6851%	76.2818%	2.0846	
[MILK, CHEESE, EGGS] ==> [BACKED GOODS]	13.6851%	37.3986%	2.0846	
[BEER, ALCH. FREE DRINKS] ==> [MILK, CHEESE, EGGS]	12.6168%	48.3676%	1.3218	
[MILK, CHEESE, EGGS] ==> [BEER, ALCH. FREE DRINKS]	12.6168%	34.4794%	1.3218	
[CONFECTIONERY] ==> [FRESH MEAT, SAUSAGES, FISH]	12.3082%	39.8081%	1.6098	
[FRESH MEAT, SAUSAGES, FISH] ==> [CONFECTIONERY]	12.3082%	49.7720%	1.6098	
[BEER, ALCH. FREE DRINKS] ==> [CONFECTIONERY]	11.1243%	42.6459%	1.3793	
[CONFECTIONERY] ==> [BEER, ALCH. FREE DRINKS]	11.1243%	35.9789%	1.3793	
[MILK, CHEESE, EGGS] ==> [PROCESSED FOOD (EXC. SOUPS)]	11.0590%	30.2222%	2.1768	
[PROCESSED FOOD (EXC. SOUPS)] ==> [MILK, CHEESE, EGGS]	11.0590%	79.6538%	2.1768	
[COFFEE, TEA, CACAO, TOBACCO] ==> [MILK, CHEESE, EGGS]	10.5872%	60.0168%	1.6401	
[MILK, CHEESE, EGGS] ==> [COFFEE, TEA, CACAO, TOBACCO]	10.5872%	28.9329%	1.6401	
[FRESH MEAT, SAUSAGES, FISH] ==> [SOUPS, SPICES, SPREAD]	10.4923%	42.4286%	2.2802	
[SOUPS, SPICES, SPREAD] ==> [FRESH MEAT, SAUSAGES, FISH]	10.4923%	56.3865%	2.2802	
[FRESH MEAT, SAUSAGES, FISH] + [CONFECTIONERY] ==> [MILK, CHEESE, EGGS]	10.4211%	84.6673%	2.3138	
[MILK, CHEESE, EGGS] + [CONFECTIONERY] ==> [FRESH MEAT, SAUSAGES, FISH]	10.4211%	57.4044%	2.3213	
[MILK, CHEESE, EGGS] + [FRESH MEAT, SAUSAGES, FISH] ==> [CONFECTIONERY]	10.4211%	55.9503%	1.8096	
[CONFECTIONERY] ==> [SOUPS, SPICES, SPREAD]	10.1688%	32.8887%	1.7675	
[SOUPS, SPICES, SPREAD] ==> [CONFECTIONERY]	10.1688%	54.6484%	1.7675	

Figura 9: Análise de afinidades [IBM, 2007].

4.3 Oracle Darwin Data Mining Software

O Darwin Data Mining Software da Oracle, é descrito em seu site oficial [Oracle, 2007] como uma poderosa ferramenta de mineração de dados que ajuda a transformar gigantes massas de dados em inteligência corporativa. Darwin ajuda a encontrar padrões significativos e correlações em dados corporativos. Padrões que permitem um melhor entendimento e previsão do comportamento de clientes.

Também segundo o site, utilizando o Darwin pode-se traçar estratégias para vendas conjuntas para clientes existentes, pode-se obter novos clientes, detectar fraudes, identificar clientes mais lucrativos e traçar perfis de clientes com maior eficácia.

O software é baseado em algoritmos de aprendizagem, com a implementação de algoritmos de árvores de decisão, redes neurais, entre outros. Além disso, possui também algoritmos para otimização, avaliação e comparação de modelos. Oferece funções básicas de estatística necessárias para obter alguma compreensão dos dados iniciais e para avaliar as previsões feitas pelos modelos obtidos.

Um dos pontos fortes da ferramenta é sem dúvida a facilidade de utilização do sistema através de um conjunto de *wizards*, que ajudam e guiam o usuário na criação de modelos automaticamente e na seleção dos melhores e mais apropriados, como pode ser visto em alguns exemplos na Figura 10.

Na Figura 11 é mostrado o resultado de uma segmentação de clientes, através de uma visualização interativa em forma de árvore.

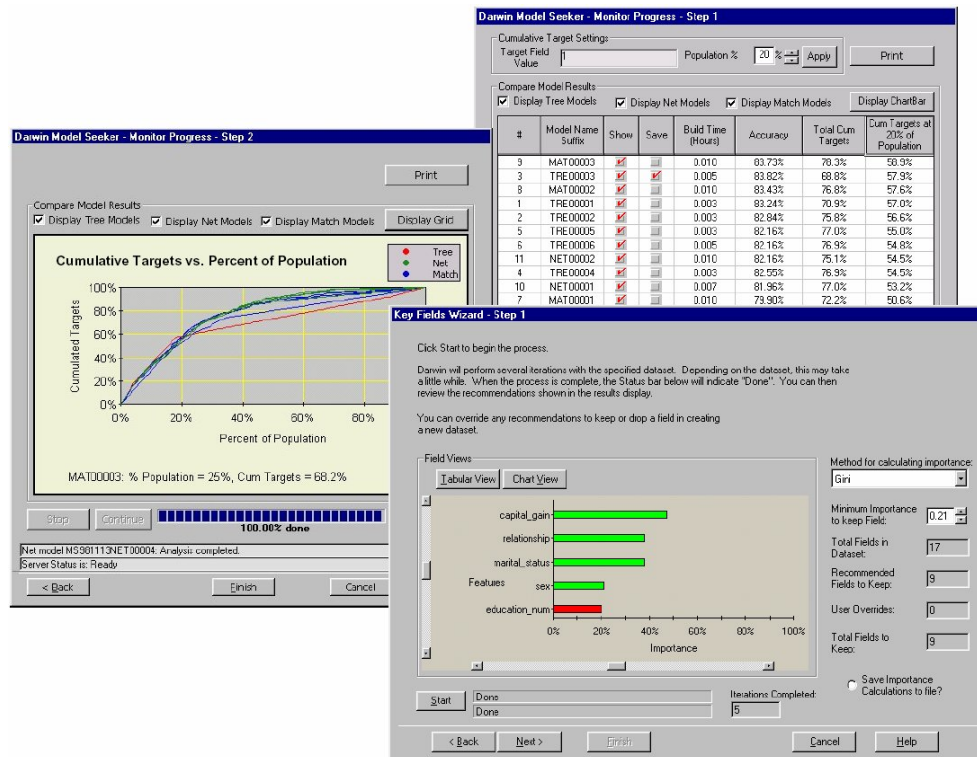


Figura 10: Wizards para criação de modelos de mineração de dados [Oracle, 2007].

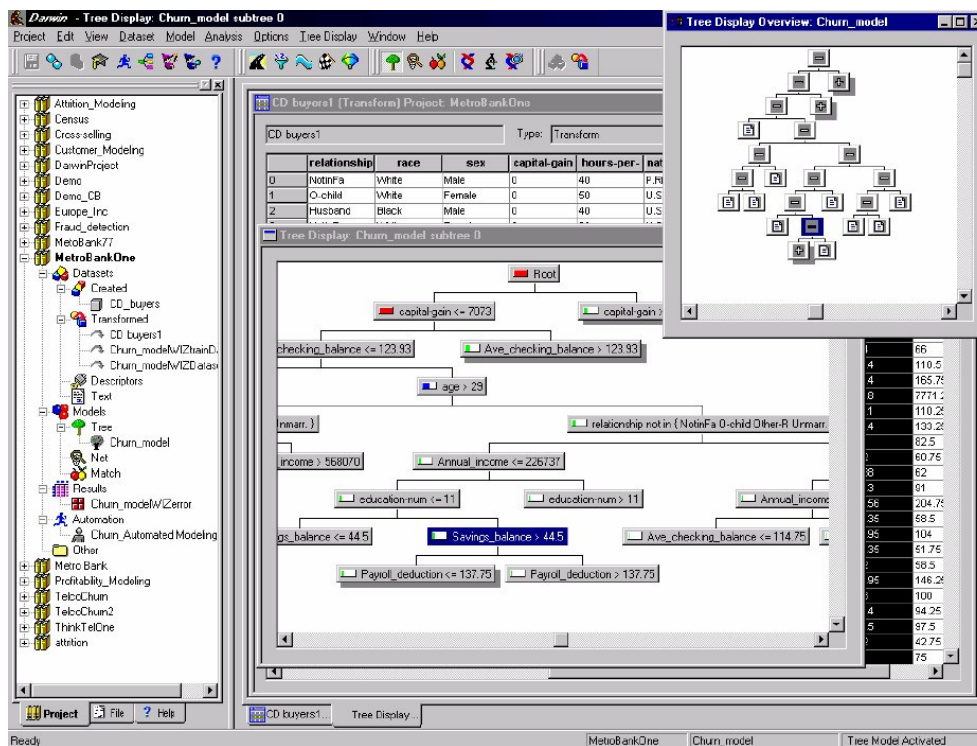


Figura 11: Visualização em forma de árvore para segmentação de clientes [Oracle, 2007].

O Darwin também possui relatórios que são integrados com o Microsoft Excel, como pode ser visto na Figura 12.

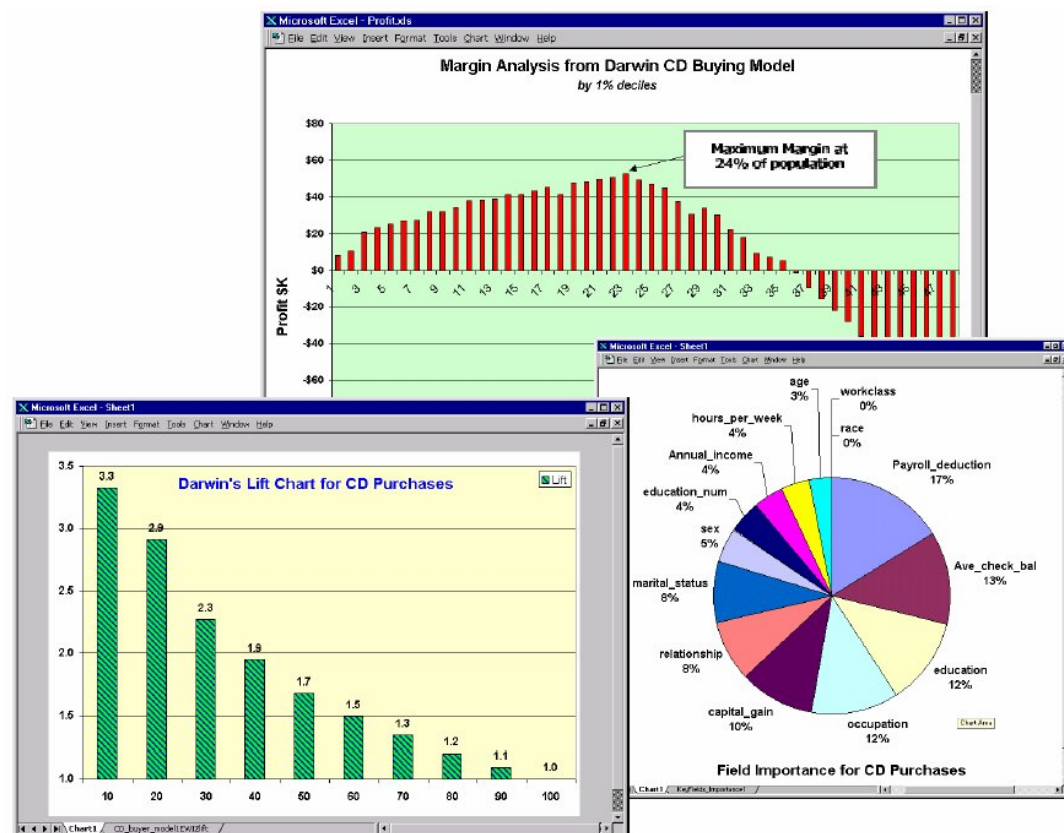


Figura 12: Gráficos para visualização de resultados em formato MS Excel [Oracle, 2007].

5 Conclusões

Neste trabalho foi abordado o tema da mineração de dados (*data mining*), como importante área da administração de dados que visa à melhoria da qualidade dos mesmos e à geração de conhecimentos novos a partir deles.

Inicialmente, foram abordados os principais conceitos e características da mineração de dados. Uma visão geral das etapas do ciclo de vida de um projeto de mineração de dados foi mostrada. Observou-se a existência de diversas técnicas para se obter tipos diferentes de relações e conhecimento para propósitos diferentes, sendo detalhadas as 5 (cinco) principais delas. Também foram abordados alguns conceitos relacionados a 2 (dois) dos principais tipos de algoritmos utilizados para a aplicação dessas técnicas.

Ao longo do trabalho foi enaltecida a importância da mineração de dados em diversas áreas de conhecimento da nossa sociedade. Foram detalhados 3 (três) exemplos práticos de aplicações reais da mineração de dados em empresas, e foram analisados os resultados obtidos a partir da utilização das práticas da mineração de dados por estas empresas.

Finalmente, algumas ferramentas que auxiliam no desenvolvimento e aplicação do processo de mineração de dados em uma organização foram brevemente descritas.

5.1 Trabalhos Futuros

Com a conclusão deste trabalho, pode-se identificar alguns estudos futuros que podem ser feitos a partir do tema. Mais especificamente, pode-se realizar um estudo de caso da aplicação de um processo de mineração de dados em uma empresa, desde o entendimento do negócio, até à avaliação do processo.

Também pode ser feita uma análise de retorno de investimento de empresas reais que investiram na mineração de dados com o objetivo de aumentar os seus lucros.

5.2 Considerações Finais

Este trabalho teve como um de seus objetivos, estimular as empresas para que elas dêem cada vez maior importância a seus dados, obtidos em suas operações diárias. Mostrou-se o quanto uma empresa pode agregar de valor para seus negócios, a partir de conhecimentos escondidos em seus dados. Com a concorrência cada vez mais forte, empresas que não adotarem estratégias não só para manter a qualidade dos dados, como também para a geração de conhecimento a partir deles, estarão reduzindo cada vez mais suas chances no mercado em que atuam.

6 Referências

[Carvalho, 2005] Luís Alfredo Vidal de Carvalho. Data Mining – A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. 2005

[Chapman, 2005] Chapman, A. D. 2005. Principles and Methods of Data Cleaning – Primary Species and Species- Occurrence Data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

[DAMA International, 2006] The Data Management Association. Disponível em: <http://www.dama.org/public/pages/index.cfm?pageid=1>. Acessado em Janeiro de 2007.

[Euriditionhome, 2004] Data Mining Tutorials, Resources. Disponível em: <http://datamining.eruditionhome.com>. Acessado em Janeiro de 2007.

[Fayyad et al., 1996] Fayyad, Usama; Piatetski-Shapiro, Gregory; Smyth, Padhraic; Uthurusamy, Ramasamy. Advances In Knowledge Discovery And Data Mining. Novembro de 1996.

[Gimenes, 2000] Eduardo Gimenes, A Importância da Mineração de dados em tomadas de decisões. Disponível em: http://geocities.yahoo.com.br/dugimenes/arquivos/data_mining.zip. Acessado em Fevereiro de 2007.

[Gouvêa, 2005] Custódio Gouvêa, Introdução a Técnicas de Data Mining – DM. Disponível em: <http://arquivosevt.incc.br/pdfs/Introducao%20Data%20Mining%203.pdf>. Acessado em Fevereiro de 2007.

[Han et al., 2001] Han, Jiawei; Kamber, Micheline. Data Mining – Concepts and Techniques. 2001

[IBM, 2007] IBM DB2 Data Warehouse Edition. Disponível em: <http://www-306.ibm.com/software/data/db2/dwe/>. Acessado em Março de 2007.

[Navega, 2002] Sergio Navega, Princípios Essenciais do Data Mining. Disponível em: <http://www.intelliwise.com/reports/i2002.htm>. Acessado em Dezembro de 2006.

[Oracle, 2007] Oracle Darwin Data Mining Software. Disponível em: <http://www.oracle.com/technology/documentation/darwin.html>. Acessado em Março de 2007.

[Pelegrin et al., 2005] Diana Colombo Pelegrin, Diego Paz Casagrande, Merisandra Côrtes de Mattos, Priscyla Waleska Targino de Azevedo Simões, Rafael Charnovski, Jane Bettiol. As Tarefas de Associação e de Classificação na Shell de Data Mining Orion. Disponível em: <http://www.dcc.unesc.net/sulcomp/artigos/sessaoOral/22103.pdf>. Acessado em Fevereiro de 2007.

[SAS, 2007] SAS Enterprise Miner. Disponível em: <http://www.sas.com/technologies/analytics/datamining/miner/>. Acessado em Março de 2007.

[Sferra e Corrêa, 2003] Heloisa Helena Sferra, Ângela M. C. Jorge Corrêa. Conceitos e Aplicações de Data Mining. Disponível em: <http://www.unimep.br/phpg/editora/revistaspdf/rct22art02.pdf>. Acessado em Janeiro de 2007

[Sousa, 1998] Mauro Sérgio Ribeiro de Sousa, 1998. Mineração de Dados: Uma implementação fortemente acoplada a um sistema gerenciador de banco de dados paralelo. Disponível em: <http://www.cos.ufrj.br/~marta/papers/TeseMauroS.pdf>. Acessado em Janeiro de 2007.

[TDWI, 2006] The data warehousing institute. Disponível em: <http://www.tdwi.org/>. Acessado em Março de 2007.

[The CRISP-DM Consortium, 2000] CRoss Industry Standard Process for Data Mining. Disponível em: <http://www.crisp-dm.org>. Acessado em Janeiro de 2007.

[Wikipedia, 2006] Sistema de Gerenciamento de Banco de Dados. Disponível em: <http://pt.wikipedia.org/wiki/SGBD>. Acessado em Dezembro de 2006.

Assinaturas

Recife, 29 de março de 2007

Fernando da Fonseca de Souza (orientador)

Thiago Miranda Amorim Silva (aluno)