ESCOLA POLITÉCNICA
DE PERNAMBUCO

d.s.c. UPE
Departamento de Sistemas e Computação

ANÁLISE DA APLICAÇÃO DE ALGORITMOS DE DATA MINING EM BASES DE DADOS DE VENDAS DE PRODUTOS

Trabalho de Conclusão de Curso
Engenharia da Computação

Nome do Aluno: Alexandre da Costa Batista
Orientador: Meuser Valença

ALEXANDRE DA COSTA BATISTA

ANÁLISE DA APLICAÇÃO DE ALGORITMOS DE DATA MINING EM BASES DE DADOS DE VENDAS DE PRODUTOS

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia da Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Recife, novembro 2009

Dedico este trabalho aos meu pais Amaro Romão e Ana Celina,
a meu irmão André Batista,
minha noiva Georgina Marafante
e a meus tios Romão e Lúcia

Agradecimentos

Agradeço primeiramente a Deus por ser a causa de todas as coisas e por transitividade ter sido também a causa de eu ter chegado até a este ponto no curso de Engenharia da Computação.

Aos meus pais Amaro e Ana, por todo o apoio e suporte para que eu pudesse chegar aqui me dando as condições necessárias para que isso acontecesse

À minha noiva Georgina por estar comigo sempre me apoiando como fez em mais esta etapa da minha vida.

Ao professor Meuser por aceitar me orientar me ajudando a atingir os objetivos esperados.

Por fim à Universidade de Pernambuco por ter disponibilizado a estrutura necessária para que eu passasse estes anos da minha vida aprendendo a me tornar um bom profissional.

Resumo

Nas companhias que têm como finalidade a obter lucro, tomar decisões com rapidez e qualidade representa um grande desafio experienciado pela pelo responsável pela gestão do negócio. Para superar este desafio, há a necessidade de tais empresas aperfeiçoarem seus processos de tomada de decisão. Essa necessidade pode ser justificada por tais otimizações possibilitarem reduções de custo ou elevações de receita, conseqüentemente com aumento de lucro.

Como uma das respostas da Tecnologia da Informação para solucionar tais necessidades se baseia nos conceitos de Business Intelligence e Data Mining, este trabalho objetiva, através da pesquisa, entender o conceito e as características em Business Intelligence. Além disso, visa compreender as relações entre BI e Data Mining para, em seguida, aplicar tal conhecimento na construção de um processo de implantação de Data Mining usando-se os conceitos de Business Intelligence obtendo-se como resultado um ambiente de BI com conhecimento capaz de auxiliar a empresa a tomar decisões de negócios com mais garantia e sucesso.

Abstract

In companies that are intended to make a profit, make decisions with speed and quality is a major challenge experienced by the person responsible for managing the business. To overcome this challenge, there is a need for such companies improve their processes of decision making. This need can be justified by such optimizations make possible cost reductions or revenue increases, and consequently with increased profit.

As one of the responses of Information Technology to address those needs based on the concepts of Business Intelligence and Data Mining, this paper aims, through research, understand the concept and features in Business Intelligence. It also seeks to understand the relationships between BI and Data Mining, and then apply this knowledge to build an implementation process of Data Mining using the concepts of Business Intelligence was obtained as a result of a BI environment with knowledge that assist the company to make business decisions with greater accuracy and success.

Sumário

Índice de Figuras

Tabela de Símbolos e Siglas

DM – Data Mining

MD – Mineração de Dados

SGBD – Sistema de gerenciamento de banco de dados

Introdução

Falando-se das empresas com fins lucrativos, o principal desafio vivenciado pelo indivíduo que está responsável pela gestão do negócio é tomar decisões com rapidez e qualidade. Com isto, no meio corporativo, existe a necessidade das empresas aperfeiçoarem seus processos de tomada de decisão. Dentre os motivos para justificar tal necessidade encontra-se o fato de tais otimizações viabilizarem aumento de lucro que, conforme Peter Drucker - respeitado autor da área de Administração de Empresas, representa um dos principais objetivos perseguidos por empresas privadas.

É importante lembrar que estas decisões são baseadas em componentes racionais e emocionais cuja intensidade e direção dependem do fator informação. Nesse contexto, profissionais de Tecnologia da Informação (TI), especificamente de áreas como Banco de Dados e Sistemas de Suporte a Decisão, trabalham com conceitos, técnicas e ferramentas que tanto atendem às necessidades citadas acima quanto organizam e valorizam o patrimônio de informações de negócios das empresas que implantam soluções tecnológicas desta natureza.

Sendo assim temos que destacar os conceitos de Business Inteligence e Data Mining. Business Intelligence, segundo a definição clássica de Howard Dresner considerado o pai do termo, tendo o inventado em 1989 é “o processo com o intuito de explorar e analisar informações estruturadas e específicas de um domínio para enxergar tendências ou padrões e, através disto, produzir percepções e tirar conclusões. Domínios incluem clientes, fornecedores, produtos, serviços e concorrentes”. Data Mining, de acordo com o dicionário é “o processamento de dados usando sofisticadas capacidades de busca de dados e algoritmos estatísticos para descobrir padrões e correlações em grandes bancos de dados preexistentes; uma forma de descobrir novos significados dos dados.”

Podemos exemplificar com um caso bastante conhecido e difundido de aplicação de Data Mining em grandes empresas que é o caso da rede de supermercados norte-americana Wall-Mart. Procurando por relações entre

vendas e dias da semana, perceberam que nas sextas-feiras, o consumo de cerveja aumentava assim como o de fraldas. Logicamente isso não indicava crianças bebendo cerveja, mas depois de uma análise mais profunda revelou que quando os pais iam comprar fraldas para seus filhos, aproveitavam e compravam cerveja para o fim de semana. Desta forma, o supermercado poderia por exemplo, sabendo que homens não costumam demorar muito dentro do supermercado, neste dia colocar estes produtos juntos e próximos à entrada do supermercado, aumentando a quantidade de vendas.

Outro caso interessante é o da PUC do Rio de Janeiro após a análise da informação de vários alunos de vestibular, identificou que boa parte dos candidatos do sexo feminino, que trabalhavam e tiveram aprovação com boa nota no vestibular, não efetivava a matrícula. Refletindo sobre esta situação chegou-se a conclusão que mulheres na idade de vestibular e trabalhavam é porque precisavam trabalhar, sendo assim provavelmente também fez matrícula na universidade pública, e como obteve boas notas, deve ter passado na pública, efetivando a matrícula nesta. Logicamente existem exceções, mas a grande maioria obedece à regra. Neste caso temos a tecnologia para nos ajudar a entender o meio ambiente onde se encontra a empresa.

Objetivos e contexto

O objetivo deste trabalho é, através da pesquisa, entender os conceitos e as características de Business Intelligence e Data Mining e aplicar o conhecimento adquirido para tentar descobrir informações escondidas dentro de uma base dados de uma fabrica de refrigerantes, mostrando que esta tecnologia é acessível e pode ajudar aos tomadores de decisão e gestores de empresas a conseguir melhor desempenho das suas empresas. Com este trabalho, será possível ter uma aplicação prática da técnica ajudando analistas de negócio e de sistemas a fazer uma implantação desta técnica na sua empresa.

Foram escolhidas as ferramentas Microsoft, para a demonstração desta técnica, pois elas apresentam uma boa interface amigável, através da incorporação de bons recursos de análise de dados às ferramentas já conhecidas como a Microsoft Office System onde a complexidade do processo é oculta, apresentando visualizações simples e assistentes que facilitam o

trabalho do usuário levando as informações relevantes diretamente às pessoas responsáveis pela tomada de decisões.

Serão usados neste trabalho os SQL Server Add-ins para Office 2007 que é um conjunto de ferramentas fáceis de usar de mineração de dados que permitem análises preditivas em um desktop. Podendo usar os algoritmos de mineração do SQL Server Analisys Services, que é o serviço responsável por criar modelos de data mining que possibilitarão descobrir informações escondidas nos dados, no ambiente já bastante conhecido do Office, os usuários do negócio, pode ter facilmente dicas valiosas advindas de dados complexos, pois foram desenvolvidos tendo como objetivo o usuário final habilitando-os a ter estas análises avançadas diretamente no Microsoft Excel.

Das ferramentas importantes adicionadas por este Add-in, podemos destacar as duas a seguir [1] :

- Table Analysis Tools for Excel: Possibilita ao usuário final que tem o a ter analises poderosas em dados presentes em planilhas.
- Data Mining Client for Excel: Oferece um modelo complete de mineração dentro do Excel 2007.

Capítulo 1

Conceitos

Aqui serão apresentados os conceitos apresentados neste trabalho e suas respectivas definições para que possamos obter um bom entendimento do trabalho.

Os principais conceitos necessários são:

- Business Intelligence;
- Data Mining;
- Microsoft Office System;
- SQL Server;

Para definirmos ***Business Intelligence (BI)***, é importante que trata-se de um "termo guarda-chuva". Vários autores se referem ao conceito como sendo um "guarda-chuva" em que estariam incluídos outros conceitos como: *Data Warehouse*, D*ata Mart*, D*ata Warehousing,* ETL*, DSS, EIS (Executive Information System), OLAP* e *data mining.*

Cabe agora salientar também que tal conceito apresenta-se definido de formas variadas. A explicação para tal fato é que diferentes pessoas e empresas criaram e adotaram diferentes definições do conceito. Partindo-se finalmente para uma definição, dentre as principais definições da área, destaca-se a definição clássica de *Howard Dresner,* considerado o pai do termo, que em seu trabalho afirma que BI é "*o processo com o intuito de explorar e analisar informações estruturadas e específicas de um domínio para enxergar tendências ou padrões e, através disto, produzir percepções e tirar conclusões. Domínios incluem clientes, fornecedores, produtos, serviços e concorrentes*". Cabe ressaltar que complementarmente à definição usando linguagem de negócios de *Dresner*, mais formada por termos de Tecnologia da Informação, na qual ***Business Intelligence*** é uma categoria ampla de

aplicativos e tecnologias para captar, armazenar, analisar e prover acesso aos dados corporativos de forma a auxiliar os tomadores de decisão a tomarem melhores decisões de negócio.

Também conhecida como Prospecção de Dados ou Mineração de Dados, é o processo de processar grandes quantidades de dados à procura de padrões, como regras de associação, ou seqüências temporais, para detectar relacionamentos entre variáveis, detectando novos subconjuntos de dados.

Trata-se de um conceito recente na área de tecnologia da informação que para atingir seus resultados, utiliza técnicas da estatística, reconhecimento de padrões e inteligência artificial.

Esta mineração se dá pela aplicação de técnicas de ferramentas que através de algoritmos de classificação e aprendizagem baseados em redes neurais e estatísticas, nos permitem explorar os dados nos ajudando na descoberta de conhecimento que se dá pela descoberta de padrões.

Este processo é também conhecido por Descobrimento de Conhecimento em Bancos de Dados (KDD em inglês). Este termo foi criado por Gregory Piatetsky-Shapiro em 1989 para descrever o processo de descobrir dados interessantes, interpretados e úteis. Este processo consiste de três passos que são pré-processamento dos dados, em seguida executar a mineração propriamente dita e a interpretação dos resultados. Estes passos serão explicados mais a frente no trabalho.

Os humanos desde sempre aprenderam observando padrões, criando e testando hipóteses para verificar a efetividade delas. O computador facilitou bastante no que diz respeito ao armazenamento dos dados, que chegou a um ponto de serem necessários novos métodos para que esses dados possam ser processados.

Empresas hoje acumulam uma grande quantidade de dados, e em se tratando de uma empresa de vendas, por exemplo, Business Intelligence pode ser praticada apenas verificando as estatísticas de venda no dia, por exemplo,

que produto saiu mais para que os estoques deste produto sejam repostos com mais freqüência e fique sempre com um nível mais alto para se ter menos dependência do fornecedor, ou identificar que produto gerou mais lucro no dia.

Mas se fizermos uma análise mais aprofundada tentando encontrar a relação entre as variáveis, estaremos praticando a Mineração de Dados, tentando identificar que tipo de cliente leva um produto ou quando um produto é levado, existe muita chance de que um outro também seja comprado.

Podemos então classificar Business Intelligence e Data Mining em dois níveis. O primeiro ajuda, com informações úteis, na tomada de decisões. O segundo ajuda, no plano estratégico, a apresentar aos tomadores de decisão informações novas a respeito do meio ambiente em que a empresa atua.

As tecnologias de mineração de dados possibilitam a análise de dados através da aplicação de algoritmos e análises estatísticas de dados para que seja possível descobrir idéias e oportunidades importantes tais como a determinação de segmentos ou a realização de uma análise de mercado para prever a probabilidade de uma promoção de um certo produto aumentar as vendas de outro.

Pode ser necessário trabalho de processamento dos dados extensivo antes que as informações possam aparecer. Usando ferramentas mais amigáveis pode tornar mais fácil o trabalho do tomador de decisões ajudando-o a prever tendências e relações entre os dados.

Microsoft Office System é uma suíte de aplicativos bastante difundida da fabricante Microsoft que possibilita a integração de vários recursos e compartilhamento de arquivos facilitando o trabalho dos usuários corporativos com aplicativos de escritórios como editor de texto (Microsoft Word), planilha de cálculos (Microsoft Excel) entre outros. Este último será o mais utilizado neste trabalho por apresentar plug-ins que fazem a integração como SQL Server e funciona bem como front-end e fonte de dados para a mineração de dados.

SQL Server é um SGBD (Sistema de Gerenciamento de Bandos de Dados) produzido pela Microsoft que teve sua primeira versão em 1988 passando por várias versões atingindo um alto nível de maturidade para aplicação em empresas de nível mundial respondendo bem aos requisitos de desempenho e segurança que este tipo de corporação exige.

1.1 Data Mining e Ética

O uso de dados, principalmente sobre pessoas, para a mineração de dados tem sérias implicações éticas [2]. Usuários desta técnica, devem sempre tomar cuidado para não se colocar no meio deste tipo de problema.

Quando estas técnicas são aplicadas às pessoas, são freqüentemente usadas para discriminar e alguns tipos de discriminação como racial ou sexual não somente são antiéticas como também são ilegais.

Deve-se sempre levar em consideração o problema exposto, pois, por exemplo, no caso de diagnósticos médicos, o uso de características sexuais e raciais não só é ético como necessário.

Portanto é sempre bom antes de fornecer dados pessoais, saber como eles serão utilizados e como será garantida a confidencialidade

Capítulo 2

Processo de KDD (Knowledge Discovery in Databases)

Este processo em sua definição mais popular proposta em 1996 por um grupo de pesquisadores (Fayyad ET al., 1996a) é “um processo, de várias etapas, não trivial, interativo e iterativo, para a identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados” [3].

O KDD é constituído de três passos básicos que são o pré-processamento, a mineração propriamente dita e o pós-processamento ou interpretação dos resultados. Vemos então que o termos mais conhecido, Mineração de Dados, é na verdade uma das etapas do processo de descobrimento de conhecimento em bancos de dados.

Vejamos mais detalhadamente cada um dos passos do KDD:

2.1 Pré-processamento

Como o processo de data mining pode descobrir informações que já estão nos dados, o conjunto de dados deve ser grande o suficiente para conter estas informações e pequeno o suficiente para que possa ser processado em um tempo aceitável. Este conjunto de dados precisa ser limpo para que sejam removidos ruídos e dados inconsistentes.

Este passo tem etapas intermediárias que precisam ser bem executadas para que tenhamos bons resultados. Vamos ver cada uma delas com sua função.

2.1.1 Seleção

É nesta etapa que identificamos quais dados de que bancos de dados serão utilizadas para o processo de KDD.

Geralmente a origem destes dados são bancos de dados transacionais, de sistemas que estão em produção durante o dia todo, ou seja, estão sendo utilizados e atualizados constantemente com novas informações, novos clientes, clientes alterados, novos pedidos, novos produtos ou quaisquer outras informações, sendo assim, é sempre recomendado que para iniciarmos o processo, tenhamos uma cópia do banco de dados, para que as tarefas de mineração não interfiram de forma alguma nos processos do dia a dia da empresa. Isto se justifica ainda pelo fato de que a maioria dos métodos de mineração [3] pressupõe que os dados estejam em apenas uma única tabela.

É aqui também que escolhemos os atributos e os registros que serão utilizados, eliminando atributos que não sejam relevantes. Essa eliminação deve ser feita com conhecimento do negócio da empresa.

2.1.2 Limpeza

No mundo real com aplicações reais, é muito comum os que os dados presentes bancos de dados apresentem inconsistências, ruídos. Estes dados apresentam erros e podem conter valores divergentes contendo discrepâncias semânticas.

Uma limpeza bem feita nos dados, conseqüentemente leva a uma boa qualidade destes. Esta qualidade é extremamente necessária, pois devemos ter em mente que eles serão intensivamente usados para descobrirmos informações e conhecimento deles.

A melhor forma de evitarmos esta poluição nos dados é sem dúvida a validação da entrada deles no banco de dados do sistema de produção, mas estas validações nem sempre são bem projetadas ou até mesmo estão disponíveis.

2.1.3 Codificação

Na codificação dizemos como os dados serão apresentados. Para a escolha desta apresentação devemos sempre ter em mente as necessidade do algoritmo que será aplicado.

Esta codificação deve ser feita com bastante atenção, pois ela influencia diretamente nos conhecimentos extraídos dos dados.

2.1.4 Enriquecimento

Os dados precisam nos revelar conhecimentos. Por isso, precisamos dotá-los de poder para tal. Isso é feito quando agregamos mais informações aos registros existentes.

2.1.5 Normalização

Atributos com valores diferentes do normal podem influenciar negativamente no processo. Sendo assim os registros com estes valores discrepantes devem ser removidos do conjunto de dados submetido ao processo para que não influenciem tendenciosamente nos resultados.

2.1.6 Construção de atributos

Com esta operação podemos gerar novos atributos tomando como base os já existentes. Estes atributos derivados nos ajudam a expressar relações entre os atributos existentes, podendo inclusive diminuir o conjunto de dados a ser processado pois pode resumir mais de um atributo. Trata-se de uma operação muito comum no processo de KDD.

2.1.7 Correção de Prevalência

É uma tarefa muito útil em procedimentos de classificação, corrigindo desequilíbrios na distribuição dos registros. Diferentemente do processo de normalização dos dados que se limita a excluir registros com atributos discrepantes, esta correção os exclui semanticamente. Por exemplo, se temos um produto que tem poucas vendas efetuadas e momentaneamente não é foco principal da empresa, as vendas deste produto devem ser excluídas para que não influenciem nos percentuais de vendas dos produtos foco.

2.1.8 Partição do Conjunto de dados

Devemos sempre garantir a qualidade dos dados e dos resultados que podemos obter deles, sendo assim, e pela necessidade dos algoritmos de serem treinados para poderem gerar informações, se faz necessário que tenhamos um conjunto de dados para treinamento. Este conjunto é subconjunto do todo que temos para a aplicação da mineração. Portanto devemos particionar o conjunto que inicial que temos para submetermos ao algoritmo para que ele seja treinado. Alem disso, devemos em seguida, testar o treinamento do algoritmo com outro subconjunto dos dados. Estes dois subconjuntos devem necessariamente ser diferentes para que a avaliação seja isenta.

2.2 Mineração dos dados

Trata-se da principal etapa do processo de KDD. É aqui que efetivamente o conhecimento é buscado a partir dos dados aplicando-se algoritmos. Estes algoritmos procuram explorar só dados de forma a produzir conhecimento [3], baseados em determinados paradigmas.

Temos alguns exemplos de algoritmos que podem ser aplicados e o que podemos encontrar com a aplicação de cada um:

- Classificação – arruma os dados em grupos predefinidos.
- Clustering – arruma os dados em grupos não predefinidos encontrando padrões de semelhança entre os dados
- Regressão – tenta encontrar uma função que modele os dados com o menor erro possível.
- Aprendizado de regras de associação – procura por relacionamentos entre as variáveis.

Vamos ver mais detalhes sobre os algoritmos aplicados neste trabalho no capítulo 3, onde são mostrados como eles agem para a extração do conhecimento.

2.3 Validação dos resultados / interpretação das informações

Aqui podemos verificar quais padrões produzidos pelos algoritmos de data mining é válido, ou seja, acontece em todo o conjunto de dados. Nem todos os padrões encontrados são necessariamente válidos. É comum encontrar padrões que não estão presentes no conjunto total de dados o que é chamado de overfitting, por isso o resultado deve ser testado com o conjunto de dados de teste sobre os quais os algoritmos ainda não foram aplicados para verificarmos a validade dos padrões.

Capítulo 3

Algoritmos de mineração

Os algoritmos de mineração de dados são os mecanismos que criam os modelos de mineração. Para criar este modelo, o algoritmo analisa o conjunto de dados e procura por padrões e tendências. Sendo assim, o algoritmo usa os resultados desta análise para definir os parâmetros de mineração. Então, estes parâmetros são aplicados ao conjunto completo de dados para extrair padrões e estatísticas detalhadas.

Os modelos que os algoritmos podem ser de vários tipos, dentre eles destacamos:

- Um conjunto de regras que descreve como produtos estão agrupados
- Uma árvore de decisão que pode dizer se um cliente em particular comprará um produto
- Um modelo matemático mapeando previsões de venda

A ferramenta onde o trabalho foi desenvolvido disponibiliza vários algoritmos para uso. Algoritmos de terceiros também podem ser usados desde que sejam compatíveis com sua tecnologia. Dentre os disponibilizados por padrão, podemos destacar os tipos a seguir:

- Algoritmos de classificação
- Algoritmos de regressão
- Algoritmos de segmentação
- Algoritmos de associação
- Algoritmos de análise de seqüências

Neste trabalho serão usadas implementações de algoritmos de associação e de

3.1 Algoritmo de Associação

Este algoritmo é bastante utilizado para fazer recomendações de compra aos clientes baseado em compras feitas por eles mesmos ou por intenções de compra deles. Por estes motivos, ele é bastante utilizado para analisar cesta de compras.

Os modelos de associação são construídos em conjuntos de dados que contém identificadores tanto para os casos individuais quanto para os itens que cada caso contém [4]. Aqui, o caso, é a transação que para o nosso caso pode ser mapeado para o pedido. Ou seja, temos a identificação do pedido e dos itens que participam do pedido. Sendo assim, o modelo de associação são séries de conjuntos de dados e as regras que descrevem como estes itens são agrupados dentro dos pedidos. As regras que o algoritmo identifica podem ser usadas para prever as compras futuras de um cliente

3.1.1 Funcionamento do algoritmo

O algoritmo varre dos dados procurando por itens que aparecem juntos em um número mínimo transações. Este número é dado por um parâmetro de entrada do algoritmo que define este valor. O algoritmo recebe também como entrada, o campo da tabela ou da consulta que define o identificador da transação e o identificador do item. Desta forma ele gera as regras para os conjuntos de dados. Estas regras são usadas para prever a presença de um produto baseado na presença de um outro produto identificado como importante pelo algoritmo, ou seja, com maior relevância.

A implementação que iremos analisar deste algoritmo, é baseado no algoritmo Apriori. O algoritmo Apriori não analisa padrões, em vez disto, ele gera e conta conjuntos de itens candidatos.

Tipos comuns de variáveis lógicas representando Sim ou Não, ou Existente ou Não Presente, são atribuídos a cada atributo, como o descrição do produto. A análise de cesta de compras é um exemplo da aplicação deste algoritmo que usa variáveis lógicas para representar a presença ou ausência de um produto na cesta de compras de um cliente

Para cada conjunto de itens, o algoritmo cria pontuações para representar o suporte e a confiança que são usados para produzir um ranking e novas e interessantes regras.

Suporte, Probabilidade, e Importância

Também referido como freqüência [5] representa a porcentagem de transações de um banco de dados de transações onde a regra se verifica. Somente os itens que tem no mínimo a quantidade especificada, podem ser incluídos no modelo.

Podemos chamar de conjunto de itens freqüente, uma porção de itens que tem suporte acima do limite especificado no parâmetro MINIMUM_SUPPOR. Por exemplo, se o conjunto itens é {A,B,C} e o parâmetro MINIMUM_SUPPORT é 10, cada item A, B, C individualmente, deve aparecer em pelo menos 10 casos e a combinação deles, também deve aparecer em no mínico 10 casos para que esta regra possa ser incluída no modelo.

Também podemos limitar a quantidade de regras baseado na probabilidade ou confiança de uma regra. Por exemplo, se o conjunto {A, B, C} aparece em 50 casos, o conjunto {A, B, D} em outros 50 e o conjunto {A, B} não podemos afirmar que {A, B} necessariamente leva a {C} então podemos calcular a probabilidade de {A, B} -> {C} dividindo o suporte de {A, B, C} pelo suporte de todos os conjuntos de dados relacionados. O valor de probabilidade pode ser configurado no parâmetro MINIMUM_PROBABILITY.

Portanto, seja I = {i1, i2, i3, ..., in} um conjunto de atributos binários chamados de itens, e D = {t1, t2, t3, ..., tn} um conjunto de transações, cada transação em D tem um identificador que contém um subconjunto de itens tem I. Uma regra é definida como uma implicação lógica de X -> Y onde X, Y contém I e X interseção Y = conjunto vazio.

3.2 Algoritmo de Naive Bayes

Este é um algoritmo de classificação usado na modelagem preditiva. Este nome deriva do fato de que este algoritmo usa o teorema de Bayes, mas não

leva em conta as dependências que possam existir entre os dados e sendo assim é chamado de Naive, ou seja, ingênuo.

Não se trata de um algoritmo tão intenso computacionalmente [6] quanto os outros, por isso, pode ser usado para termos modelos de mineração rápidos para descobrir relacionamentos entre colunas de entrada e colunas onde tentamos fazer uma previsão. É usado para fazer uma mineração inicial dos dados. De acordo com as saída deste algoritmo, podemos aplicar outro que seja mais preciso.

3.3 Algoritmo de Cluster

Este é um algoritmo de segmentação de dados. [7] Ele usa técnicas iterativas para agrupar os casos no conjunto de dados porções que contém características similares. Este agrupamento pode ser usado para explorar os dados e identificar anomalias nos dados.

Este algoritmo nos ajuda a identificar relações entre os dados que não são facilmente visualizados em uma observação simples

Capítulo 4

Estudo de Caso

Neste trabalho, foi utilizada a ferramenta Microsoft Excel 2007. Esta ferramenta foi escolhida para que pudesse ser mostrado que este tipo de tecnologia está disponível a qualquer pessoa com bons conhecimentos de informática em um aplicativo difundido mundialmente em qualquer organização, desde as de pequeno até as de grande porte.

Isto possibilita ao usuário comum, com um breve treinamento, ter acesso a informações que podem ajudar a tomar decisões mais acertadas. Para o caso da fábrica de refrigerantes, pode ajudar a desenvolver, por exemplo, uma ação de mercado que intensifique a venda de um produto que ultimamente não está sendo vendido em grandes quantidades, para determinado perfil de cliente.

O autor deste trabalho teve acesso apenas às tabelas do banco de dados da fábrica que seriam relevantes ao desenvolvimento deste e ainda assim com informações significativas, mas apenas do período do último quadrimestre do ano de 2008, o que nos dá uma boa quantidade de informação pra ser trabalhada.

As tabelas utilizadas foram carregadas a partir do banco de dados da fábrica (SQL Server 2000) para um servidor de testes (SQL Server 2008). Neste último também estava instalado o serviço de Analisys Services, responsável por armazenar os modelos criados no Excel. O Excel, por sua vez, para que tivesse acesso aos algoritmos do Analisys Services e ao seu suporte para a análise dos dados, tinha instalado um Add-in chamado SQL Server Data Mining Add-ins for Office 2007. Este add-in adiciona duas abas às já existentes do Excel possibilitando a análise dos dados.

Foi dada ênfase em tabelas que apresentassem informações de pedidos e de perfil, tanto de clientes como de produtos. A tabela de pedidos apresenta um campo identificando unicamente cada pedido. Por essa tabela ter uma modelagem na qual armazena o cabeçalho e os detalhes do pedido, sempre

que fizermos alguma análise, precisaremos filtrá-la para que tenhamos apenas os itens, pois neles temos todas as informações necessárias. Na figura 1 podemos ver as tabelas utilizadas no processo.

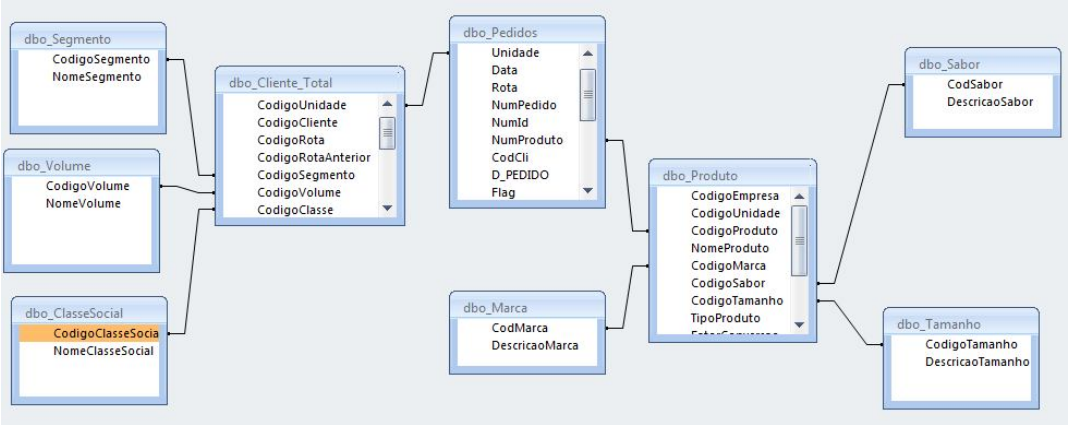

Figura 1 Modelagem do banco de dados (apenas tabelas importantes)

Vejamos como foram feitos os passos de KDD para o caso da fábrica de refrigerantes:

4.1 Pré-processamento

4.1.1 Seleção de Dados

Identificamos como informações a serem consideradas, informações de vendas da fábrica aos seus clientes. Neste ponto geralmente faz-se necessário a cópia dos dados atualmente em produção na empresa, geralmente em seu banco de dados relacional, para um banco de testes onde estes dados possam ser consultados e alterados sem interferir nos processos da empresa. E assim, foi feita uma cópia das tabelas envolvidas para um servidor de testes. Limpeza

Foram encontrados produtos de clientes que não tinham informações de perfil. Estes pedidos foram excluídos da análise por não poderem ser representativos para o nosso objetivo.

4.1.2 Codificação

Temos disponíveis informações sobre pedidos de compras realizados por clientes à fábrica de refrigerantes. É interessante dizer que também dispomos de um perfil, ainda que resumido, dos clientes e dos produtos vendidos. Portanto a análise foi feita considerando as vendas e os perfis disponíveis. E neste ponto de codificação escolhemos como os dados serão representados. Não foi necessária nenhuma transformação de dados, pois no estado em que se encontram, já estão prontos para serem usados pelos algoritmos escolhidos.

4.1.3 Enriquecimento

Os dados de vendas foram enriquecidos com as informações de perfis do cliente e do usuário. Adicionalmente foi incluída a informação de Rota de venda do cliente. Esta rota constitui a separação geográfica dos clientes já que clientes que se encontram próximos participam da mesma rota de vendas. Infelizmente, a informação de que rota representa cada bairro ou localidade não nos foi disponibilizada. As rotas são necessárias para guiar os vendedores da empresa durante suas atividades do dia. Sendo assim ele, no começo do dia, já sabe por que clientes passar apenas seguindo sua rota.

4.1.4 Normalização de Dados

Para esta nossa análise, os dados já se encontravam normalizados não sendo necessária nenhuma tarefa neste sentido.

Foi apenas necessário um filtro na tabela de pedidos, já que para esta modelagem apresentada, esta tabela contém tanto informações de cabeçalho como de detalhes dos produtos vendidos em um pedido. Sendo assim foram filtrados para que tivéssemos apenas as informações dos itens vendidos. O cabeçalho não apresenta informações importantes para este tipo de análise.

4.1.5 Construção de Atributos

Este processo consiste em gerarmos atributos a partir dos já existentes nos dados. Para o nosso caso, apenas um atributo novo foi gerado que se trata do preço de venda do produto. Esta informação foi encontrada através da divisão

do preço de venda do produto e sua quantidade vendida, que por sua vez estavam incluídas em um campo texto que continha várias outras informações que não serão consideradas.

4.1.6 Correção de Prevalência

Para o nosso caso, esta empresa tem a venda de refrigerantes de cola muito superior que a de qualquer outro sabor. Neste passo de correção, pede-se uma correção de um eventual desequilíbrio entre os dados com certas características. Logicamente veríamos um desequilíbrio grande entre a venda de refrigerantes de cola e de outros sabores, então foi decidido não aplicar nenhuma técnica neste sentido, para que pudéssemos ver esta diferença e, quem sabe poder sugerir ações de venda para impulsionar a venda dos outros sabores.

4.1.7 Partição do Conjunto de Dados

A partição dos dados em conjuntos distintos de treinamento e testes é feita aleatoriamente pelo algoritmo em uso que tem parâmetros para configurar a porcentagem do total dos dados que será utilizada para os dois conjuntos.

4.1.8 Extração de dados

Para termos informações pertinentes para a aplicação dos algoritmos de mineração, a extração dos dados foi feita com a consulta SQL abaixo:

```
1.  select p.NumPedido Pedido, p.Data, pr.NomeProduto Produto,
    t.DescricaoTamanho,
2.  p.Rota, cs.NomeClasseSocial, s.NomeSegmento Segmento, v.NomeVolume
    Volume
3.   from Pedidos p
4.   inner join Cliente_Total ct
5.    on ct.CodigoUnidade = p.Unidade
6.    and ct.CodigoCliente = p.CodCli
7.   inner join ClasseSocial  cs
8.    on cs.CodigoClasseSocial = ct.CodigoClasse
9.   inner join Segmento s
10.   on s.CodigoSegmento = ct.CodigoSegmento
11.  inner join Volume v
12.   on v.CodigoVolume = ct.CodigoVolume
13.  inner join Produto pr
14.   on pr.CodigoUnidade = p.Unidade
```

```
15.    and pr.CodigoProduto = p.NumProduto
16.  where p.NumId = 7
17.  and pr.CodigoMarca in
     ('d','A','B','C','D','E','F','G','K','P','Q','R','8','g','i','k','
     l')
18.  order by Data, NumPedido
```

O resultado desta consulta foi carregado diretamente em uma planilha Excel. Quando a selecionávamos tínhamos acesso às abas especiais de Data Mining.

A linha 16 corresponde ao filtro, já comentado anteriormente, para termos apenas os detalhes ou itens de cada pedido.

Na linha 17 do código acima, existe um filtro para que sejam apenas mostrados produtos refrigerantes, já que esta fábrica também distribui cervejas e sucos e foram escolhidos apenas os produtos refrigerantes já que são os principais e além de distribuí-los ela também fabrica. Estes produtos têm uma alta representatividade no total de produtos.

Para conseguirmos os dados necessários

Na figura 2 a representação dos dados carregados em uma planilha. Estes dados foram carregados utilizando a consulta explicada mais acima. Detalhes de como esta importação de dados foi feita serão omitidos por não serem foco do trabalho.

	A	B	C	D	E	F	G	H
1	Pedido	Data	Produto	Tamanho	Rota	ClasseSocial	Segmento	Volume
2	185512	20081111	COLA	260 ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME
3	185512	20081111	COLA	290 ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME
4	185512	20081111	LARANJA	260 ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME
5	185512	20081111	LARANJA	290 ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME
6	185512	20081111	COLA	2000ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME
7	185512	20081111	LARANJA	2000ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME
8	185512	20081111	LIMAO	2000ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME
9	185512	20081111	GUARANA	2000ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME
10	193631	20081111	COLA	260 ML	B81	NAO PARCEIRO	INDIRETO	BAIXO
11	193631	20081111	COLA	290 ML	B81	NAO PARCEIRO	INDIRETO	BAIXO
12	193631	20081111	COLA	2000ML	B81	NAO PARCEIRO	INDIRETO	BAIXO
13	193631	20081111	COLA	1000ML	B81	NAO PARCEIRO	INDIRETO	BAIXO
14	193631	20081111	LARANJA	1000ML	B81	NAO PARCEIRO	INDIRETO	BAIXO
15	221736	20081111	COLA	260 ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME
16	221736	20081111	COLA	290 ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME
17	221736	20081111	COLA	2000ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME

Figura 2 Dados do banco na planilha

Temos também uma outras visão que pode ser usada para alguns algoritmos incluindo o valor do produto e agregando os campos de descrição

do produto e tamanho em apenas um campo e temos os dados como mostra a figura 3.

	A	B	C	D	E	F	G	H
1	NumPedido	Data	Produto	Rota	NomeClasseSocial	NomeSegmento	NomeVolume	Valor
2	185512	20081111	COLA - 260 ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME	15.88
3	185512	20081111	COLA - 290 ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME	15.88
4	185512	20081111	LARANJA - 260 ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME	15.88
5	185512	20081111	LARANJA - 290 ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME	15.88
6	185512	20081111	LARANJA - 2000ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME	12.05
7	185512	20081111	LIMAO - 2000ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME	12.05
8	185512	20081111	GUARANA - 2000ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME	10.66
9	193631	20081111	COLA - 260 ML	B81	NAO PARCEIRO	INDIRETO	BAIXO	15.88
10	193631	20081111	COLA - 290 ML	B81	NAO PARCEIRO	INDIRETO	BAIXO	15.88
11	193631	20081111	COLA - 2000ML	B81	NAO PARCEIRO	INDIRETO	BAIXO	16.01
12	193631	20081111	COLA - 1000ML	B81	NAO PARCEIRO	INDIRETO	BAIXO	10.5
13	193631	20081111	LARANJA - 1000ML	B81	NAO PARCEIRO	INDIRETO	BAIXO	9.39
14	221736	20081111	COLA - 260 ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME	17.48
15	221736	20081111	COLA - 290 ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME	17.48
16	221736	20081111	COLA - 2000ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME	16.43
17	221736	20081111	LARANJA - 2000ML	B82	PARCEIRO	INDIRETO	ALTO VOLUME	12.47
18	222885	20081111	COLA - 260 ML	B81	NAO PARCEIRO	INDIRETO	BAIXO	17.48
19	222885	20081111	COLA - 290 ML	B81	NAO PARCEIRO	INDIRETO	BAIXO	17.48
20	222885	20081111	COLA - 2000ML	B81	NAO PARCEIRO	INDIRETO	BAIXO	16.43
21	297600	20081111	COLA - 260 ML	C19	NAO PARCEIRO	TRADICIONAL LOW	BAIXO	16.31
22	297600	20081111	COLA - 290 ML	C19	NAO PARCEIRO	TRADICIONAL LOW	BAIXO	16.31
23	297600	20081111	LARANJA - 200 ML	C19	NAO PARCEIRO	TRADICIONAL LOW	BAIXO	10.63
24	297600	20081111	LARANJA - 473 ML	C19	NAO PARCEIRO	TRADICIONAL LOW	BAIXO	10.63
25	297600	20081111	COLA - 1000ML	C19	NAO PARCEIRO	TRADICIONAL LOW	BAIXO	6.42

Figura 3 Outra visão dos dados na planilha

4.2 Aplicação do Cluster

Com este algoritmo podemos ver a distribuição das características dos registros de venda agrupados por semelhança. O que na verdade se torna bastante interessante quando se tem um objetivo já traçado, por exemplo, diversificar os produtos vendidos para um determinado segmento de clientes.

Para a execução deste algoritmo, foi escolhido o algoritmo de K-means. Este parâmetro é escolhido através da tela de parâmetros visualizada na figura 4 Este algoritmo foi escolhido por conseguir resumir a quantidade de subdivisões de categorias (clusters) onde temos informações bem definidas possibilitando uma boa análise dos dados.

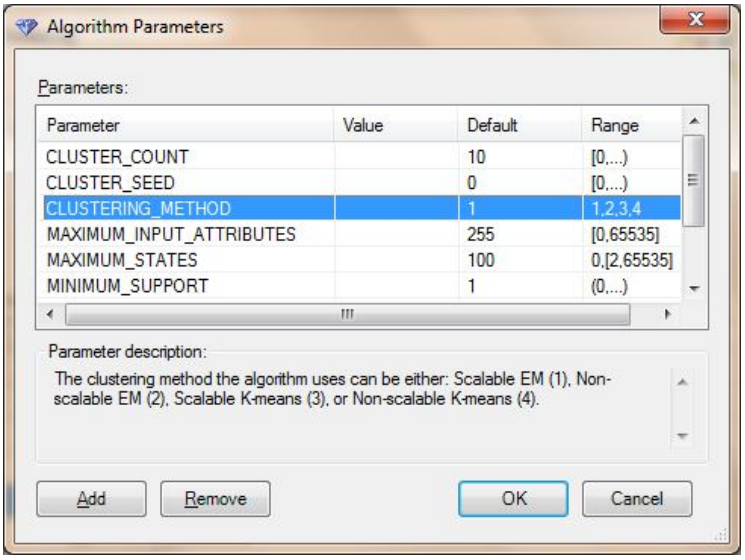

Figura 4 Parâmetros do algoritmo de Cluster

Quando é pedido que o algoritmo seja executado, também é pedido que se escolha a porcentagem dos dados que serão escolhidos aleatoriamente a base de treinamento do algoritmo. Sendo assim, ele mesmo já efetua o treinamento e o teste em seguida, nos dando como saída um gráfico dividindo as categorias pelos atributos escolhidos, que neste caso foram apenas, o produto, tamanho da embalagem do produto e classe social, segmento e volume do cliente. Na figura 5 podemos ver resultado gerado sobre os dados de pedidos.

Figura 5 Resultados do algoritmo de Cluster

Analisando o resultado gerado, podemos inferir algumas informações importantes para a empresa. Isto pode levar à criação de ações e promoções a

fim de atingir os objetivos. Devemos lembrar que os clusters estão ordenados pela quantidade de vezes que são encontrados nos dados.

Nos números gerais, disponíveis na coluna “Population”, podemos ver que a empresa tem uma clara divisão de vendas para parceiros e não parceiros, com uma boa quantidade de produtos diferentes, sempre dando ênfase no refrigerante de cola, mas com uma boa variedade de segmentos e tamanhos. Também existe uma divisão entre alto e baixo volume de vendas. Entretanto analisando clusters individualmente podemos encontrar algumas informações interessantes. Vejamos algumas delas:

- Em se tratando do mix de produtos (quantidade de produtos diferentes) podemos notar que no cluster 2 temos um baixo mix, dando muita ênfase ao refrigerante de cola. Vendo os outros atributos, vemos que os clientes do cluster 2 em sua maioria são não parceiros e são de baixo volume. Em se tratando do cluster 2, vemos que existe uma grande quantidade, ou pelo menos a segunda mais significativa, de clientes não parceiros, de baixo volume e com baixo mix de produtos, ou seja, exatamente o que a empresa precisa mudar nos três atributos. Temos que notar que para este cluster temos uma grande quantidade de Segmento Bar-Lanchonete, que tradicionalmente, não apresenta um alto volume. Portanto temos que ter em mente que quando uma ação ou promoção foi feita para que estes índices apresentados no cluster sejam alterados, provavelmente o atributo de volume pode não ser alterado. Sem dúvida com o marketing agindo, pode-se aumentar o mix de produtos, o que interessa ao fabricante, pois terá mais produtos diferentes sendo consumidos por clientes e seu nome mais conhecido. Também deve focar em tornar os clientes em parceiros, pois desta forma, o cliente tem mais descontos na compra dos produtos e conseqüentemente, aumentando o volume vendas.
- Comparando os clusters 4 e 7, vemos que tendo eles as mesmas características de classe social e volume, apresentam uma diferença entre os outros atributos. Existe uma informação interessante nestes dados. Vejamos que no cluster 4 com tamanhos variados principalmente menores que 2000ml e uma grande quantidade de refrigerante de cola. Já no cluster 7, vemos que existe uma predominância de refrigerante de guaraná e uma grande quantidade de vendas de produtos de 2000ml. Daqui podemos ver uma boa

separação entre refrigerante de cola com baixos tamanhos e de guaraná com altos tamanhos. Neste ramo dizemos que baixos tamanhos são essencialmente para consumo individual enquanto os de grandes tamanhos, a partir de 1000ml, são de consumo familiar e vemos que é necessário dar mais atenção a este tipo de cliente incentivando tanto a venda de produtos familiares de cola quanto a produtos individuais de guaraná.

4.3 Aplicação do Shopping Basket (Associação)

Com este algoritmo podemos ver as associações entre os produtos em pedidos de compras. Com ele podemos ver que produto mais é comprado em conjunto com outros. Isso nos dá uma visão que nos ajuda a por exemplo fazermos com que o produto carro chefe da empresa se associe a vários outros fazendo com que eles também sejam vendidos em boa quantidade, aumentando o lucro da empresa. Com uma das planilhas geradas, podemos ver como está a distribuição dos produtos e quantidade de ocorrências nos nossos dados de cada conjunto de dados.

Para iniciarmos o processamento deste algoritmo precisamos preencher alguns valores iniciais (figura 6) para o número da transação, ou pedido no nosso caso, em que campo encontra-se o item vendido e o valor. Este último é opcional mas quando presente temos a possibilidade de ver sua relevância levada em consideração no resultado do algoritmo.

Uma necessidade deste algoritmo é que os dados estejam ordenados segundo o campo de número do pedido.

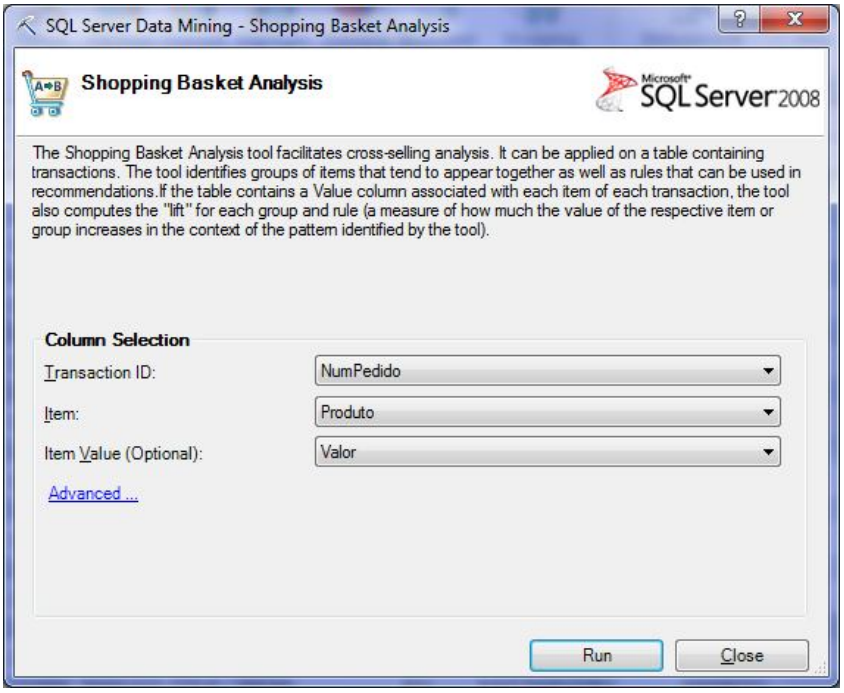

Figura 6 Parâmetros do algoritmo de Associação

Também nos é permitido escolher o valor de suporte e probabilidade (probability) que serão aplicados ao algoritmo. Neste caso foram escolhidos os valores de 10 itens para suporte e 50% para probabilidade. Quanto maior a probabilidade menos regras serão geradas e teremos mais facilidade de entender os resultados.

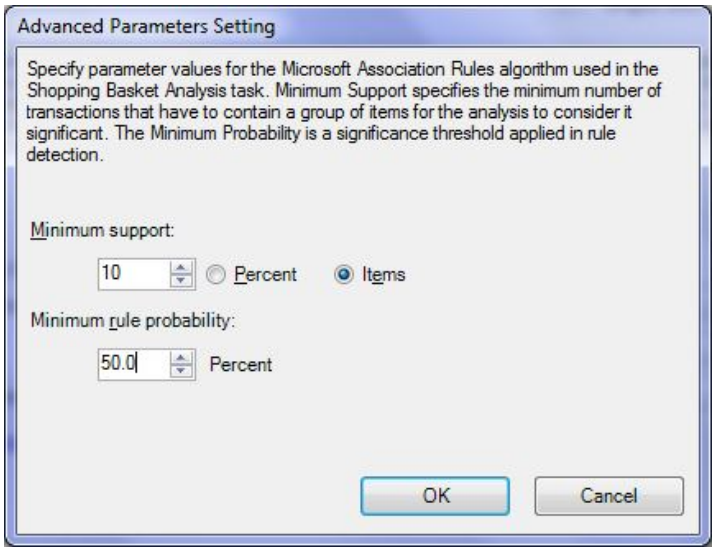

Figura 7 Parâmetros do algoritmo de Associação

Quanto terminamos de selecionar os parâmetros o algoritmo pode ser executado nos mostrando cada passo da execução como vemos abaixo.

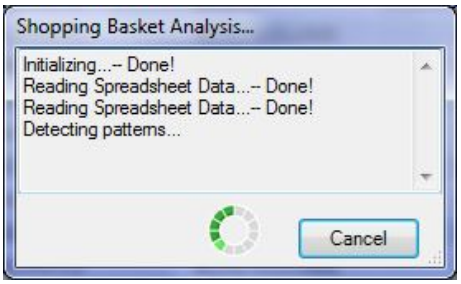

Figura 8 Progresso de execução do algoritmo

Logo após a execução ele nos apresenta 2 planilhas recém criadas com informações interessantes.

A primeira delas é para nos mostrar a divisão das vendas baseadas nas regras recém encontradas, como vemos na figura 9.

	A	B	C	D	E
1	Shopping Basket Bundled Items				
2					
3					
4	Bundle of items	Bundle size	Number of sales	Average Value Per Sale	Overall value of Bundle
5	COLA - 290 ML, COLA - 260 ML	2	12810	31.64884952	405421.7624
6	COLA - 290 ML, COLA - 260 ML, COLA - 1000ML	3	7398	40.2023888	297417.2723
7	LARANJA - 290 ML, LARANJA - 260 ML, COLA - 290 ML	3	4338	47.31655472	205259.2144
8	LARANJA - 290 ML, LARANJA - 260 ML, COLA - 260 ML	3	4338	47.31655472	205259.2144
9	LARANJA - 290 ML, COLA - 290 ML, COLA - 260 ML	3	4338	47.18658108	204695.3887
10	LARANJA - 260 ML, COLA - 290 ML, COLA - 260 ML	3	4338	47.18658108	204695.3887
11	COLA - 290 ML, COLA - 1000ML	2	7398	24.27093419	179556.3712
12	COLA - 260 ML, COLA - 1000ML	2	7398	24.27093419	179556.3712
13	COLA - 473 ML, COLA - 290 ML, COLA - 260 ML	3	3996	41.88961584	167390.9049
14	COLA - 200 ML, COLA - 290 ML, COLA - 260 ML	3	3996	41.88961584	167390.9049
15	COLA - 2000ML, COLA - 290 ML, COLA - 260 ML	3	3296	49.44749822	162978.9541
16	LARANJA - 290 ML, LARANJA - 260 ML	2	4817	31.64452564	152431.68
17	COLA - 473 ML, COLA - 200 ML, COLA - 260 ML	3	3996	36.32011197	145135.1675
18	COLA - 473 ML, COLA - 200 ML, COLA - 290 ML	3	3996	36.32011197	145135.1675
19	COLA - 473 ML, COLA - 200 ML	2	7058	20.38762256	143895.84
20	COLA - 2000ML, COLA - 1000ML	2	5288	26.31663607	139162.3716
21	LARANJA - 290 ML, COLA - 290 ML	2	4338	31.50104526	136651.5344
22	LARANJA - 260 ML, COLA - 290 ML	2	4338	31.50104526	136651.5344
23	LARANJA - 260 ML, COLA - 260 ML	2	4338	31.50104526	136651.5344
24	LARANJA - 290 ML, COLA - 260 ML	2	4338	31.50104526	136651.5344

Figura 9 Divisão de vendas algoritmo de Associação

Vemos que, sem dúvida, o refrigerante de cola é o mais vendido. Na verdade as vendas dele são praticamente o dobro do próximo segundo produto que é o de laranja. Logo em seguida começamos com uma variação maior dos produtos nos pedidos com o aparecimento de mais de um produto (coluna “size”) por pedido. Existe uma boa divisão entre os pedidos de tamanho 2 e 3.

Também é possível ver que os maiores valores alcançados são necessariamente pelos refrigerantes de cola, mas que os de laranja,

especialmente os de tamanhos individuais tem grande potencial. Entretanto os de guaraná não aparecem na parte de cima da lista. Isso se dá principalmente pelo fato de que existe um grande concorrente neste sabor e vemos que é necessário alguma ação no sentido de alavancar as vendas deste refrigerante.

	A	B	C	D	E	F	G
4	Selected Item	Recommendation	Sales of Selected Items	Linked Sales	% of linked sales	Average value of recommendation	Overall value of linked sales
5	COLA - 290 ML	COLA - 260 ML	12810	12810	100.00 %	15.82442476	202710.8812
6	COLA - 260 ML	COLA - 290 ML	12810	12810	100.00 %	15.82442476	202710.8812
7	LARANJA - 260 ML	LARANJA - 290 ML	4817	4817	100.00 %	15.82226282	76215.84
8	LARANJA - 290 ML	LARANJA - 260 ML	4817	4817	100.00 %	15.82226282	76215.84
9	LARANJA - 2000ML	COLA - 2000ML	5737	4226	73.66 %	12.80411782	73457.22394
10	COLA - 200 ML	COLA - 473 ML	7058	7058	100.00 %	10.19381128	71947.92
11	COLA - 473 ML	COLA - 200 ML	7058	7058	100.00 %	10.19381128	71947.92
12	LARANJA - 260 ML	COLA - 290 ML	4817	4338	90.06 %	14.12577421	68043.85435
13	LARANJA - 290 ML	COLA - 290 ML	4817	4338	90.06 %	14.12577421	68043.85435
14	LARANJA - 260 ML	COLA - 260 ML	4817	4338	90.06 %	14.12577421	68043.85435
15	LARANJA - 290 ML	COLA - 260 ML	4817	4338	90.06 %	14.12577421	68043.85435
16	COLA - 473 ML	COLA - 290 ML	7058	3996	56.62 %	8.956580823	63215.54745
17	COLA - 473 ML	COLA - 260 ML	7058	3996	56.62 %	8.956580823	63215.54745
18	COLA - 200 ML	COLA - 260 ML	7058	3996	56.62 %	8.956580823	63215.54745
19	COLA - 200 ML	COLA - 290 ML	7058	3996	56.62 %	8.956580823	63215.54745
20	COLA - 260 ML	COLA - 1000ML	12810	7398	57.75 %	4.816195941	61695.47
21	COLA - 290 ML	COLA - 1000ML	12810	7398	57.75 %	4.816195941	61695.47
22	GUARANA - 2000ML	COLA - 2000ML	4201	2990	71.17 %	12.36098635	51928.50364
23	COLA - 2000ML	COLA - 1000ML	9087	5288	58.19 %	5.200851766	47260.14
24	COLA ZERO - 350 ML	COLA - 350 ML	4121	3169	76.90 %	11.05921081	45575.00774
25	LARANJA - 350 ML	COLA - 350 ML	4228	3246	76.77 %	10.70473529	45259.62079
26	COLA - 600 ML	COLA - 2000ML	4591	2339	50.95 %	8.916341045	40934.92174
27	LIMAO - 2000ML	COLA - 2000ML	3158	2346	74.29 %	12.84867511	40576.116
28	GUARANA - 2000ML	LARANJA - 2000ML	4201	2994	71.27 %	9.265338608	38923.68749

Figura 10 Recomendações do algoritmo de Associação

O algoritmo também nos fornece uma boa quantidade de recomendações. Em sua maioria, levam o refrigerante de cola, pois fica claro que este sabor, pode alavancar as vendas de vários outros. Mas devemos atentar para o potencial do de guaraná em tamanhos familiares. Ou seja, uma ação no sentido de favorecer as vendas do guaraná em tamanhos familiares atrelados aos de cola pode fazer com que este produto fique cada vez mais conhecido e aceitado no mercado, levando indiretamente também a vendas mais altas nos tamanhos individuais. Isto se dá pelo fato conhecido do perfil de consumo de clientes de refrigerantes que se acostumam com o sabor do já conhecido e evitam provar outros sabores inclusive outras marcas. Desta forma pode-se difundir o refrigerante em tamanhos familiares para aumentar as suas vendas como um todo.

Vemos também que a maioria das regras sugere a venda associada de produtos de mesmo tamanho ou próximo, ou seja, apenas individuais ou apenas familiares, mas devemos dar atenção às sugestões para tamanhos diferentes também para que possam ser difundidos mais tamanhos para a o mercado, diversificando o mercado e atingindo mais consumidores.

4.4 Aplicação do “Influenciadores Chave” (Naive Bayes)

Aqui veremos quais são os atributos que mais influenciam na compra de determinados valores.

Conseguimos com este algoritmos dados muito mais informativos sobre como estão as vendas da empresa influenciada por perfis de clientes por exemplo.

Neste caso, foi aplicado o algoritmo escolhendo o alvo como campo “Produto + Tamanho” e veremos como esse campo é influenciado pelos campos de Segmento, Volume e Classe Social.

Podemos ver como esses valores são passados para o algoritmo na figura 11

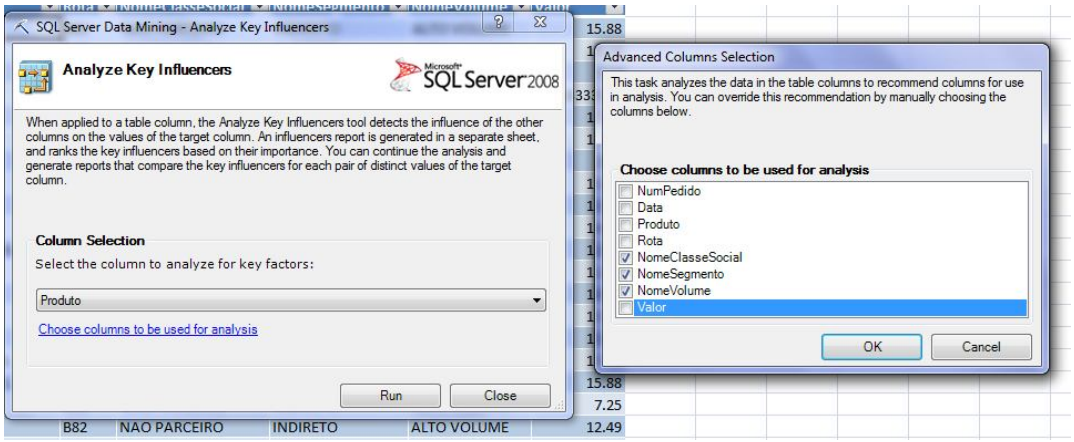

Figura 11 Parâmetros do algoritmo de Naive Bayes

Após a execução do algoritmo temos disponíveis filtros que nos ajudam no entendimento dos resultados. Isto pode ser visto na figura 12

	A	B	C	D
1	Key Influencers Report for 'Produto'			
2				
3	Key Influencers and their impact over the values of 'Produto'			
4	Filter by 'Column' or 'Favors' to see how various columns influence 'Produto'			
5	Column	Value	Favors	Relative Impact
6	NomeSegmento	BAR BOTECO	COLA - 260 ML	
12	NomeSegmento	BAR BOTECO	COLA - 290 ML	
21	NomeSegmento	BAR BOTECO	COLA - 1000ML	
48	NomeSegmento	BAR BOTECO	LARANJA - 260 ML	
53	NomeSegmento	BAR BOTECO	LARANJA - 290 ML	
142	NomeSegmento	BAR BOTECO	GUARANA - 300 ML	
169	NomeSegmento	BAR BOTECO	LIMAO - 300 ML	
250	NomeSegmento	BAR BOTECO	COLA ZERO - 260 ML	
255	NomeSegmento	BAR BOTECO	COLA ZERO - 290 ML	
287	NomeSegmento	BAR BOTECO	COLA LIGHT - 260 ML	
291	NomeSegmento	BAR BOTECO	COLA LIGHT - 290 ML	
332	NomeSegmento	BAR BOTECO	UVA - 260 ML	
336	NomeSegmento	BAR BOTECO	UVA - 290 ML	
343	NomeSegmento	BAR BOTECO	GRFFT 290M - 260 ML	
346	NomeSegmento	BAR BOTECO	GRFFT 290M - 290 ML	
395				

Figura 12 Resultado do algoritmo de Naive Bayes (Bar Boteco)

Fazendo o filtro pela coluna “values” para “BAR BOTECO” vemos que a maior parte das vendas é feita para produtos individuais o que caracteriza o negócio do ponto de venda, que como o próprio nome já diz trata-se de um ponto onde os clientes se servem neles durante um lanche ou refeição.

4	Filter by 'Column' or 'Favors' to see how various columns influence 'Produto'			
5	Column	Value	Favors	Relative Impact
18	NomeSegmento	TRADICIONAL LOW	COLA - 1000ML	
24	NomeSegmento	TRADICIONAL LOW	COLA - 2000ML	
26	NomeSegmento	TRADICIONAL HI	COLA - 2000ML	
36	NomeSegmento	TRADICIONAL HI	COLA - 350 ML	
37	NomeSegmento	TRADICIONAL LOW	LARANJA - 2000ML	
41	NomeSegmento	TRADICIONAL HI	LARANJA - 2000ML	
43	NomeSegmento	TRADICIONAL LOW	LIMAO - 2000ML	
44	NomeSegmento	TRADICIONAL HI	LIMAO - 2000ML	
59	NomeSegmento	TRADICIONAL LOW	GUARANA - 2000ML	
62	NomeSegmento	TRADICIONAL HI	GUARANA - 2000ML	
68	NomeSegmento	TRADICIONAL HI	LARANJA - 350 ML	
77	NomeSegmento	TRADICIONAL HI	LIMAO - 350 ML	
110	NomeSegmento	TRADICIONAL HI	LIMAO - 250 ML	
116	NomeSegmento	TRADICIONAL HI	LIMAO - 750 ML	
122	NomeSegmento	TRADICIONAL HI	COLA ZERO - 250 ML	
127	NomeSegmento	TRADICIONAL HI	COLA ZERO - 750 ML	
131	NomeSegmento	TRADICIONAL LOW	LARANJA - 1000ML	
138	NomeSegmento	TRADICIONAL LOW	GUARANA - 1000ML	
153	NomeSegmento	TRADICIONAL LOW	COLA - 2500 ML	

Figura 13 Resultados algoritmo Naive Bayes (segmento)

Já na figura 13 um filtro foi feito apenas trazendo os segmentos TRADICIONAIS que são supermercados e vemos que agora esse tipo de

segmento não é definitivo termos apenas um tipo de produto. Vemos com mais freqüência os produtos familiares, que é o normal vermos quando vamos a supermercados mas sabemos que este tipo de estabelecimento também vende produtos individuais com a presença dos tamanhos de 350ml que são as latas de metal não retornáveis.

Agora analisando por volume de vendas podemos ver que na figura 14 vemos que mais uma vez, este atributo por si só não define nem sabor nem tamanho dos produtos entretanto vemos que pela primeira vez vemos a presença do sabor UVA que apareceu apenas nos clientes de ALTO VOLUME

5	Column	Value	Favors	Relative Impact
27	NomeVolume	ALTO VOLUME	COLA - 2000ML	
40	NomeVolume	ALTO VOLUME	LARANJA - 2000ML	
45	NomeVolume	ALTO VOLUME	LIMAO - 2000ML	
61	NomeVolume	ALTO VOLUME	GUARANA - 2000ML	
130	NomeVolume	ALTO VOLUME	LARANJA - 1000ML	
135	NomeVolume	ALTO VOLUME	GUARANA - 1000ML	
160	NomeVolume	ALTO VOLUME	COLA LIGHT - 2000ML	
163	NomeVolume	ALTO VOLUME	LIMAO - 1000ML	
194	NomeVolume	ALTO VOLUME	UVA - 350 ML	
208	NomeVolume	ALTO VOLUME	COLA - 600 ML	
214	NomeVolume	ALTO VOLUME	COLA ZERO - 1000ML	
230	NomeVolume	ALTO VOLUME	COLA ZERO - 600 ML	
264	NomeVolume	ALTO VOLUME	LIMAO - 600 ML	
269	NomeVolume	ALTO VOLUME	GUARANA - 600 ML	
273	NomeVolume	ALTO VOLUME	LARANJA - 500 ML	
304	NomeVolume	ALTO VOLUME	COLA LIGHT - 1000ML	
320	NomeVolume	ALTO VOLUME	COLA LIGHT - 600 ML	
327	NomeVolume	ALTO VOLUME	LARANJA - 1500ML	
334	NomeVolume	ALTO VOLUME	UVA - 260 ML	
338	NomeVolume	ALTO VOLUME	UVA - 290 ML	

Figura 14 Resultados algoritmo Naive Bayes (volume)

Já com relação aos clientes de BAIXO volume, podemos ver (figura 15) que predominantemente temos a presença de produtos individuais. Ou seja, é necessário transformar estes clientes em alto volume, principalmente tornando-os parceiros com isso disponibilizando descontos para que eles passem a aumentar o seu volume de compras e quantidade de tamanhos diferentes.

5	Column	Value	Favors	Relative Impact
8	NomeVolume	BAIXO	COLA - 260 ML	
14	NomeVolume	BAIXO	COLA - 290 ML	
33	NomeVolume	BAIXO	COLA - 350 ML	
84	NomeVolume	BAIXO	LARANJA LIGHT - 250 ML	
91	NomeVolume	BAIXO	LARANJA LIGHT - 750 ML	
99	NomeVolume	BAIXO	GUARANA - 250 ML	
106	NomeVolume	BAIXO	GUARANA - 750 ML	
144	NomeVolume	BAIXO	GUARANA - 300 ML	
199	NomeVolume	BAIXO	COLA - 200 ML	
204	NomeVolume	BAIXO	COLA - 473 ML	
296	NomeVolume	BAIXO	COLA - 3000ML	

Figura 15 Resultado algorimo Naive Bayes (volume)

Como já havia sido dito antes, este algoritmo nos dá apenas informações fazendo com que tenhamos uma boa visão da distribuição de produtos entre seus perfis

Uma análise nestes resultados feita por uma funcionário de marketing da empresa inevitavelmente traria mais informações que podem ajudar a empresa a melhorar suas vendas.

Capítulo 5

Conclusão

Com este trabalho ficou claro que este tipo de tecnologia está acessível para as empresas e que se trata de uma técnica que realmente pode trazer resultados muito interessantes, ajudando a aumentar os lucros.

Foi muito importante a demonstração em um aplicativo já muito difundido e de grande aceitação, tornando ainda menos problemático a assimilação dos conceitos aqui definidos, podendo mostrar a profissionais de vendas, financeiro, marketing que sempre é possível melhorar os índices da empresa em que trabalha em que nem sempre isso significa gastos extraordinários com ferramentas computacionais.

É importante lembrar que estas análises foram feitas pelo autor do trabalho apenas com a visão adquirida nos dois anos de serviços prestados a esta fábrica, passando por projetos de vendas e marketing, tendo inclusive contato com funcionários de marketing dos quais alguns conhecimentos foram adquiridos.

Não temos como objetivo dizer que este tipo de aplicação da técnica de Mineração de Dados pode resolver todos os problemas de lucro, penetração de produtos e outros, mas sem dúvida norteia as ações para que estes problemas sejam sanados.

O uso desta aplicação não exclui alguma outra mais especializada, flexível e adaptável ao negocio do usuário, mas também aqui pode direcionar análises feitas por esta outra aplicação mais complexa poupando tempo, já que a aplicação foco do trabalho é muito simples de ser implementada e pode ser utilizada de várias maneiras para que se encontre os caminhos para atingir os objetivos da empresa.

Bibliografia

[1] MICROSOFT. **Introducing the SQL Server 2005 Data Mining Add-ins for Office 2007**. Disponivel em: <http://www.sqlserverdatamining.com/ssdm/Home/DataMiningAddinsLaunch/tabid/69/Default.aspx>. Acesso em: 14 ago. 2009.

[2] WITTEN, H. I.; EIBE, F. **Data Mining - Practical Machine Learning Tools and Techniques**. 2ª Edição. ed. San Francisco: Elsevier, v. 1, 2005.

[3] RONALDO, G.; EMMANUEL, P. **Data Mining - Um Guia Prático**. Rio de janeiro: Elsevier, v. 1, 2005.

[4] MICROSOFT. **Microsoft Association Algorithm Technical Reference**. Disponivel em: <http://msdn.microsoft.com/en-us/library/cc280428.aspx>. Acesso em: 20 out. 2009.

[5] AMO, S. D. Técnicas de Mineração de Dados, Uberlândia.

[6] MICROSOFT. **Microsoft Naive Bayes Algorithm Technical Reference**. Disponivel em: <http://msdn.microsoft.com/en-us/library/cc645902.aspx>. Acesso em: 21 out. 2009.

[7] MICROSOFT. **Microsoft Clustering Algorithm**. Disponivel em: <http://msdn.microsoft.com/en-us/library/ms174879.aspx>. Acesso em: 15 out. 2009.