CENTRO UNIVERSITÁRIO FEEVALE

SIDINEI PEREIRA GONCHOROSKI

UTILIZAÇÃO DE TÉCNICAS DE KDD EM UM CALL CENTER ATIVO

Novo Hamburgo, junho de 2007.

SIDINEI PEREIRA GONCHOROSKI

UTILIZAÇÃO DE TÉCNICAS DE KDD EM UM CALL CENTER ATIVO

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

Professor Orientador: Juliano Varella Carvalho

Novo Hamburgo, junho de 2007.

AGRADECIMENTOS

Gostaria de agradecer a todos os que, de alguma maneira, contribuíram para a realização desse trabalho de conclusão, em especial:

Aos amigos, família e meu orientador Juliano Varella de Carvalho.

RESUMO

Os *Call Centers* Ativos possuem estrutura para realizar campanhas de *telemarketing* ou vendas através do telefone e são boas alternativas para que as empresas aumentem seu número de clientes. No entanto, as empresas do ramo de atendimento podem não explorar e aproveitar da melhor maneira o conhecimento nas bases de dados. Podem ainda não identificar o melhor perfil de vendas ou ter dificuldade em criar uma estratégia para aumentar o desempenho da produção. Através do processo de descoberta de conhecimento em banco de dados (KDD) é possível identificar regras e padrões válidos aplicando as técnicas e algoritmos de mineração de dados. Sendo assim, esse trabalho apresenta o problema detalhado do *Call Center*, KDD com suas técnicas e algoritmos, dando ênfase principalmente às árvores de classificação e o *software* Weka. Por fim, é feita uma breve introdução do estudo de caso, quais técnicas serão utilizadas e como será realizada validação dos resultados.

Palavras-chave: Descoberta de Conhecimento. Técnicas de Mineração de Dados. Weka. Algoritmos de Mineração de Dados. *Call Center* Ativo.

ABSTRACT

The Outbound Call Center companies have structures to carry out sales or telemarketing campaigns through telephone, which are, good alternatives for companies to increase the number of costumers. However, these companies may not explore and use their databases records in the best way. Also, may not identify the best sales profiles or have difficulty in creating a strategy to increase the production performance. Through the discovery process of knowledge in databases (KDD), it is possible to identify rules and valid patterns applying the techniques and algorithms of data mining. Therefore, this assignment presents the Call Center problem in details, KDD techniques and algorithms, giving emphasis to the classification trees and Weka software. Finally, a brief introduction of the study case is made, which techniques will be used and how it will carry out the validation of the results.

Key words: Knowledge Discovery. Data Mining techniques. Weka. Algorithms of Data Mining. Active Call Center.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

AED	Análise Exploratória de Dados
BD	Banco de Dados
DM	Data Mining
ID3	Iterative Dichotomizer 3
JVM	Java Virtual Machine
KDD	Knowledge Database Discovery
MD	Mineração de Dados
MER	Modelo Entidade Relacionamento
MIS	Management Information Systems
SAC	Serviço de Atendimento ao Consumidor
SGBD	Sistemas Gerenciadores de Banco de Dados
PA	Posição de Atendimento
PVI	Problema com Valores Iniciais

SUMÁRIO

INTRODUÇÃO

O mundo dos negócios sempre explorou os meios de comunicação para aumentar seus índices de vendas e divulgar seus produtos a um público maior. Dentre os meios de comunicação existentes, o telefone continua sendo uma opção para se comunicar com as pessoas. As empresas cada vez mais preocupadas em acompanhar o ritmo constante da concorrência e dos desafios do mercado utilizam o telefone para comercializar seus produtos e aumentar seu público. Um *Call Center* Ativo é um tipo de empresa dotada de toda a estrutura física, humana e tecnológica capaz de realizar atendimentos através do telefone e promover campanhas de *Telemarketing*[1]. É através dos *Call Centers* Ativos que as empresas encontram as soluções mais simples e acessíveis de atingir seus objetivos e explorar um mercado mais abrangente.

Vários setores da sociedade passaram, a partir do crescimento da computação e dos bancos de dados (BD), a armazenar suas operações e produções. Estes dados formaram grandes bases que não possuem um tratamento específico e que não disponibilizam aos profissionais, muitas vezes, informações e conhecimento que possam ser utilizados em seu trabalho de modo eficiente. Através desses dados as empresas puderam criar grandes listas de *prospects*[2], chamadas de *mailing*. O trabalho do *Call Center* Ativo entra em cena no momento que essas empresas decidem realizar uma campanha de *Telemarketing* para recuperar clientes, aumentar vendas ou adquirir novos clientes.

Ao acompanhar as empresas do ramo é possível notar que existem certas limitações e dificuldades de explorar e aproveitar o conhecimento nas bases de dados geradas e atualizadas dentro das empresas de atendimento. Os operadores realizam os contatos corrigindo os dados dos clientes e preenchendo as propostas com os dados ainda não conhecidos.

---

[1] "A utilização planejada de recursos de telecomunicações e informática como forma de se obter lucro direto ou indireto, através da satisfação do mercado consumidor de qualquer bem ou serviço." (DANTAS, 1994, p. 47).
[2] Pessoa que é alvo do atendimento e candidata a adquirir o produto ou serviço oferecido.

As estratégias geradas a partir da experiência dos responsáveis podem falhar, neste caso é preciso realizar uma nova análise até que se descubra, por exemplo, que uma característica não considerada importante antes é o diferencial para o sucesso da venda em questão. Quando não existem indicações sobre quais são as características que classificam um cliente como potencial, realizar a análise pode acarretar perda de tempo e mais demora a se realizar uma venda, já que os clientes serão atendidos aleatoriamente.

A descoberta de conhecimento em banco de dados, do termo em inglês *Knowledge Database Discovery* (KDD) representa, "O processo não-trivial de identificar válidos, novos, potencialmente utilizáveis e por fim, padrões compreensíveis dentro dos dados." (TRADUÇÃO NOSSA) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p. 6). Estas características significam:

Dados: Conjunto de informações que são armazenados em Sistemas Gerenciadores de Banco de Dados (SGBD), em vários registros, com os mesmos atributos e que representam um tipo de coleção. Ex: Todas as vendas de um produto.

Padrão: Conforme (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), padrão é o grupo de itens que tem maior incidência em um conjunto de dados. Por exemplo, analisando um grupo de pessoas para descobrir o tipo de cliente que adquire capitalizações, tendo posse dos valores de rendimento e da quantidade de crediários que cada um possui, chega-se a conclusão que a segunda não influencia na aquisição e somente os que têm rendimentos superiores a certo limiar acabam as adquirindo, independente dos crediários.

Processo: Se trata do processo envolvido na descoberta do conhecimento, que passa pela preparação dos dados, garimpagem e análise. As etapas do processo de KDD, de acordo com (FELDENS, apud CARVALHO; SAMPAIO; MONGIOVI, 1999), são: seleção, pré-processamento, transformação, garimpagem, análise e assimilação.

Válido*:* São aqueles padrões considerados válidos e interessantes ao objetivo traçado.

Novo: Representa todo conhecimento adquirido e que não estava previsto ou não poderia ser deduzido através de hipótese. Um padrão gerado por hipótese não é considerado novo já que poderia ser comprovado através, por exemplo, de estatística. Um destes exemplos é a Análise Exploratória de Dados (AED) do termo em inglês *Exploratory data analysis*, que é uma técnica da área de estatística lançada por (TUKEY, 1977) que visa obter o máximo de informações ocultas entre os dados (HOAGLIN; MOSTELLER; TUKEY, 1992).

Potencialmente Utilizável: Alguns dos padrões encontrados podem acabar não sendo úteis. Para que a descoberta de conhecimento seja relevante, é preciso que o resultado não represente algo totalmente sem sentido para o negócio. Padrões que sejam muito amplos ou que tenham pouca variação em relação a outros acabam não tendo muita utilidade.

Compreensível: É poder criar padrões que possam ser entendidos pelos seres humanos e acrescentem conhecimento útil para a tomada de decisões.

A tarefa de descobrir conhecimento não é simples. Os dados recolhidos e armazenados não são preparados de forma que a qualquer momento sejam analisados para que mostrem ao usuário os relacionamentos e padrões. Os dados acabam vindo de várias fontes e necessitando de um tratamento, um pré-processamento para se definir informações importantes e corrigir possíveis imperfeições, já que tiveram origem em outros locais não sendo fruto de técnicas de KDD, conforme WIEDERHOLD (1996). Estes dados podem não estar completos.

A mineração de dados é uma área interdisciplinar que agrupa estatística, BD e inteligência artificial munida de vários algoritmos (FREITAS, 1998) e pode ser definida como:

O termo *data mining* é normalmente utilizado pela comunidade de estatísticos, analistas de dados e os MIS (*Management Information Systems*). [...] a visão de que KDD é todo processo de descoberta de conhecimento útil em dados enquanto *data mining* se refere à aplicação de algoritmos para extração de padrões em dados sem os passos adicionais do processo de KDD. (TRADUÇÃO NOSSA) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p. 3-4).

A KDD utiliza os algoritmos e técnicas de Mineração de Dados para adquirir informações importantes em bases que não estão visíveis ou identificáveis pelos usuários e que podem ser úteis na tomada de decisão. KDD também significa extrair informações de grandes bases de dados sem possuir hipóteses previamente criadas, conforme (CABENA et al., 1997). Através da Mineração de dados e seus algoritmos é possível descobrir regras de associação, classificação e *Clustering* (FAYYAD et al., 1996) (FREITAS, 1998). São exemplos de algoritmos para mineração de dados: *LargeKItemSets* (AGRAWAL; IMIELINSKI; SWAMI, 1993), *Apriori* (AGRAWAL; SRIKANT, 1994), *AprioriTid* (AGRAWAL et al., 1993), *Partition* (SAVASERE; OMIECINSKY; NAVATHE, 1995) e *Multiple Level* (ML-T2L1) (SRIKANT; AGRAWAL, 1996).

Com base nos conceitos citados, a motivação do trabalho é desenvolver uma ferramenta que facilite o tratamento e pré-processamento dos dados, que utilize a ferramenta Weka[3], *open source e freeware* para processar a base de dados com alguns dos algoritmos de mineração e que voltado para um *Call Center* ativo, consiga descobrir padrões e regras que possam auxiliar na seleção dos *prospects* das listas de *mailing* que serão trabalhados, a fim de ter um índice de vendas melhor, com um número menor de tentativas com os clientes. Desta maneira, espera-se melhorar o desempenho e aproveitar eficientemente os recursos e informações disponíveis, mas não utilizadas até o momento, por não existir uma ferramenta especializada para tal.

[3] WEKA 3: Data Mining Software in Java. Nova Zelândia. Universidade de Waikato, 2007. Apresenta todas as características do projeto e do software Weka. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/index.html>. Acesso em: 28 mar. 2007.

1 CALL CENTER ATIVO

Desde o momento da invenção do telefone, por volta da década de 1870, passou a ser um meio de comunicação para aproximar as pessoas, mesmo que esta invenção tenha demorado a romper barreiras e se popularizar. Hoje em dia, uma das formas mais eficientes de se comunicar com as pessoas é utilizando o telefone, móvel ou então residencial.

O foco deste estudo é a atividade dos *Call Center* Ativos que surgiram para explorar essas facilidades na comunicação e localização das pessoas. Os também chamados *Contact Centers* são empresas de atendimento que possuem tecnologia e estrutura telefônica capaz de realizar atendimentos ou campanhas de *telemarketing*. Embora não se tenha registros precisos, em meados da década de 80 é que estes acabaram se espalhando pelo Brasil e hoje já existem muitas empresas especializadas nessa área e que buscam personalizar o relacionamento com seus clientes. Algumas das empresas de *Call Center* do país são Atendebem[4], Meta[5], Atento[6], Ask![7].

Um *Call Center* Ativo é aquele responsável por realizar contatos com *Prospects*[8], ao contrário de Serviço de Atendimento ao Consumidor (SAC) ou central de relacionamento com o cliente que muitas empresas possuem e que é responsável por receber os contatos das pessoas interessadas em produtos ou informações.

Os funcionários destas empresas são chamados de atendentes e cada um possui para realizar suas funções, uma posição de atendimento (PA), que é uma mesa com algum tipo de isolamento lateral (pequenas paredes com cerca de 60 cm) para diminuir o ruído ao seu redor, um computador e um telefone.

---

[4] http://www.atendebem.com.br
[5] http://www.metatmkt.com.br
[6] http://www.atento.com.br
[7] http://www.askcallcenter.com.br
[8] Pessoa alvo do atendimento e candidata a adquirir o produto ou serviço oferecido.

Quanto à estrutura física e de tecnologia das empresas deste segmento, os terminais que cada operador utiliza são ligados em rede para possibilitar a utilização de *softwares* que são responsáveis por disponibilizar consultas e permitir que o atendente faça o registro de seus contatos e de suas vendas. Em relação à telefonia algumas soluções também são utilizadas como gravadores, monitores ou gerenciadores de ligações para controlar todo o fluxo das ligações da empresa.

Para coordenar todos os atendentes, estas empresas mantêm uma estrutura que é composta por supervisores que tem ligação direta com os operadores, coordenadores que são responsáveis por passar as orientações necessárias aos supervisores e ainda acima destes, os gerentes que além de ter contato direto com o contratante, tem responsabilidade sobre os outros níveis desta empresa.

Ainda sobre as pessoas envolvidas nesse trabalho, existem outros setores que não possuem uma denominação específica e que tem como responsabilidade tratar de relatórios, importações e gerenciamento de nomes que os operadores entrarão em contato.

O ambiente de uma empresa de *telemarketing* é preparado para que o operador tenha o máximo de motivação e apoio dos colegas e supervisores para que possa atender sempre melhor um *prospect*. Além de ser um ambiente de trabalho bastante motivacional, os operadores têm que manter uma postura fiel ao contratante e precisam seguir *scripts* de atendimento que são criados e remodelados para que seu atendimento seja o melhor possível e que sua meta seja alcançada através das vendas.

Os atendentes são peças fundamentais em uma empresa de *telemarketing* e para otimizar seus resultados, nos dois últimos anos, conforme ATENDEBEM (2007), as empresas têm investido em treinamento e qualificação para seus operadores. Diversos treinamentos são exemplos de qualificação oferecida ao atendente: reversão de objeções levantadas pelo *prospect*, entonação de voz, aperfeiçoamento e adaptação de *script* e aprofundamento de produto.

Outra característica envolvida em uma empresa de *telemarketing* é a questão das metas de produção. Estas são estipuladas pelas empresas que contratam o serviço e toda a empresa, desde a área de tecnologia até a própria operação – como é chamado o setor onde os operadores e supervisores trabalham – são envolvidos e contribuem para que as metas sejam alcançadas.

A área de tecnologia de um *Call Center* é talvez uma das mais importantes dentro do quadro da empresa. Ela é responsável por manter toda a estrutura de *hardware* e *software* disponível, para que não haja desperdício de tempo, e acabe por contribuir diretamente no sucesso das operações realizadas. Para ilustrar a importância desta área pode-se considerar que se por algum motivo a parte de telefonia for interrompida ou então se a rede não estiver disponível, toda a produção da empresa é comprometida e certamente vai impactar nos resultados.

1.1 Contratante

A contratante é a empresa, órgão ou pessoa que utiliza o serviço de um *Call Center* Ativo para aumentar seu volume de vendas e, por conseguinte, seu público alvo.

Na maioria dos contratos entre a empresa contratante e o *Call Center*, a primeira normalmente é responsável por fornecer as listas com o nome dos clientes e seus dados, chamadas de *mailing*, para quem o *telemarketing* Ativo deverá entrar em contato. Existem outros casos em que a criação dessa base de clientes se dá por indicação de *prospects*, onde cada um ao ser abordado indica outra pessoa para receber o contato.

A empresa contratada fornece toda a estrutura e pessoal para que a contratante possa realizar os devidos treinamentos de produto aos operadores, que a partir do momento do início da campanha. Nesse momento a empresa de *telemarketing* e seus funcionários devem estar preparados para trabalhar com os mais diversos tipos de produtos e serviços. Tanto o setor de tecnologia quanto a operação têm que entender o negócio e produto que são repassados em detalhes. A constante adaptação é uma característica desse tipo de empresa.

Outros deveres da empresa contratante são definir os tipos de produtos e as regras de negócio envolvidas, além de repassar os *layouts* de transmissão dos arquivos de *mailing* e para a transmissão das propostas. Um dos meios mais utilizados para o recebimento dos nomes no *Call Center* e envio das vendas para o contratante é a troca de arquivos. Portanto é o contratante quem define o layout e formato para esses arquivos. É a maneira mais simples de relacionamento e comunicação entre as partes.

As necessidades da empresa contratante muitas vezes são de urgência e por este motivo a estrutura da empresa contratada tem que ser o mais adaptável possível. É comum para quem trabalha nesse meio receber demandas de cancelamentos, importações e alterações em massa, através de remessas com identificadores particulares da empresa contratante.

1.2 Produto

Assim como em toda a transação comercial, o produto é fundamental para o sucesso da atividade. Nem sempre as campanhas em um *Contact Center* são somente de aquisição de bens e serviços, podem ser pesquisas, *marketing* e qualquer outro tipo de contato possível pelo telefone.

No momento em que a campanha realizada trate da venda ao *prospect,* existem algumas dificuldades que podem ser apontadas, por exemplo, o fato de o contato ser realizado por telefone é uma mudança na cultura das pessoas e uma das principais dificuldades encontradas.

Nesse tipo de comércio não há como o *prospect* apresentar documentos ou assiná-los e por este motivo a gravação do áudio do contato do operador se torna a garantia de ambas as partes. Na necessidade de entrega de documentação, correios, fax ou entregadores são peças incluídas dentro do fluxo de informações de uma venda.

É interessante notar que mesmo não tendo sucesso o contato para venda, pode-se considerar que houve uma divulgação das qualidades e características do produto ofertado. Quebrar barreiras entre o *prospect* e o produto são desafios constantes na rotina de um operador de *telemarketing*, já que a resistência aos contatos por telefones está sempre presente.

Os produtos comercializados por meio de *telemarketing* precisam ser adaptados. No caso de serviços, os problemas apontados anteriormente necessitam que dados extras sejam solicitados, para evitar que aconteçam fraudes. Nos casos de aprovação de crédito, a exigência é muito maior, tanto que certas empresas preferem que o *Call Center* realize apenas contatos para clientes que tenham sido pré-aprovados e já tenham feito parte de seu quadro de clientes. Essa atitude ajuda a garantir perdas em relação à inadimplência.

Outra modalidade de atendimento para venda de produtos é aquela em que é feito apenas um pré-cadastro no *Call Center,* nesses casos é feito apenas um contato para captar apenas os dados principais, ficando para a contratante o contato para efetivar a compra. Essa serve para produtos mais complexos ou então para acabar com uma dificuldade que é a captação de novos clientes. É importante lembrar que o *telemarketing* atinge e entra em contato com um número muito maior de pessoas.

A quantidade de informações sobre o *prospect* no recebimento do *mailing* varia conforme o tipo de produto a ser ofertado. Para produtos como cartões de crédito a existência

de dados como a renda e data de nascimento podem facilitar o direcionamento do produto mais adequado ao perfil do *prospect*.

1.3 Ferramentas e dados disponíveis

Para que o operador de *telemarketing* possa atender com maior agilidade os clientes, ele precisa de uma estrutura tecnológica disponível. Cada operador possui um terminal ou um computador ligado em rede para que possa utilizar o *software* que vai disponibilizar para ele a pesquisa de informações e também que permite que todos os seus contatos e vendas sejam registrados. Juntamente, o telefone, a central telefônica e os equipamentos de gravação do áudio dos contatos são necessários para sejam registradas as conversas, se caso forem necessárias no futuro em questões judiciais ou de controle. Todas estas questões são transparentes para que o operador direcione seu foco para a realização das vendas.

As ferramentas disponíveis para que o operador possa gerenciar e realizar seus contatos é centralizado em um *software* que possui uma *interface* simples ao operador. Este *software* é ligado em rede e utiliza um banco de dados para disponibilizar os dados ao atendente. Este possui ao seu alcance funções como transferir e resgatar um *prospect* para uma agenda pessoal, a fim de utilizá-lo em um futuro atendimento, agendar seu contato para que um outro operador qualquer possa atendê-lo em outro momento, realizar a venda com o preenchimento dos dados obrigatórios, rejeitarem clientes e receber um outro nome da base de clientes.

Um *mailing* possui vários atributos sobre a pessoa que está sendo contatada. Dados como seu nome, data de nascimento, endereço e obviamente o telefone são disponibilizados ao operador no momento do atendimento. Em situações especiais outros dados são adicionados como renda, profissão, tempo de conta e qualquer outro dado que seja conhecido do contratante do serviço de *telemarketing*. Abaixo na figura 1.1 é mostrado um exemplo dos tipos de atributos presentes em um *mailing*.

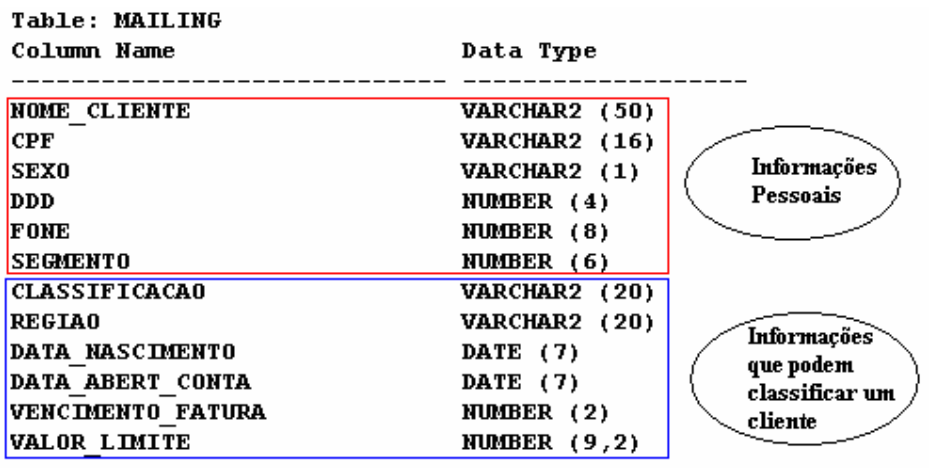

Figura 1.1 – Tipos de atributos de uma tabela *mailing* em um Banco de Dados
Fonte: FIGURA NOSSA

As empresas que realizam campanhas de *telemarketing* com seus clientes, podem ter atributos que classificam estes dentro da empresa. Em uma campanha de venda de títulos de capitalização, os clientes alvo podem ser aqueles que possuem cartão de crédito, desta maneira podem ser classificados como da categoria *internacional*, *gold*, *premium*, etc.

O recebimento dos nomes é totalmente transparente ao atendente e não há como este buscar um cliente que não esteja disponível ou em sua agenda particular. Para tal o *software* precisa de algum módulo que faça a busca destes nomes e entrega até o atendente. Essa busca de nomes pode ser simples através de uma consulta ao banco de dados em todos os computadores ou complexa com a utilização de servidor centralizando as buscas. Simplificando esse processo, o *software* poderia ir buscando na base de dados os nomes na mesma seqüência em que foi inserido cada *prospect* em cada lista. Os nomes vão sendo trabalhados e acabam sendo efetivados como vendas ou são sinalizados como uma rejeição à oferta. Os nomes podem ser também agrupados conforme suas características em listas distintas e estas serem disponibilizadas para trabalho separadamente. Ao fim destes nomes, uma nova carga é necessária para que os atendentes continuem seus trabalhos.

Esta carga de nomes na base é realizada ou pelo coordenador ou pela pessoa que tem como função exclusiva de cuidar da base de nomes. Para gerenciar estes dados, os coordenadores possuem ferramentas para que possam incluir e retirar de trabalho certas listas, além de possuírem relatórios que permitam acompanhar o rendimento de cada operador ou cada campanha.

1.4 Dificuldades

Em uma base de clientes, muitos podem não ter o perfil mínimo estipulado para a campanha e como os dados que o classificam como tal não são classificadores de prioridade de atendimento ou apenas foram captados no momento da ligação obrigam o atendente a fazer a rejeição do contato e um precioso tempo é perdido.

Estatisticamente quanto maiores forem as tentativas, da mesma maneira maiores serão as chances de realizar vendas. No momento em que os nomes das listas são trabalhados na ordem que estão no banco de dados, despreza-se a possibilidade de trabalhar primeiramente os nomes com um perfil igual ao de outras vendas já realizadas. Dessa forma, ao não trabalhar os nomes conforme seu perfil o mesmo desempenho atingido com uma lista completa, poderia ser atingida apenas com uma parte desta lista, se fosse possível ordenar pelo perfil do cliente.

Obrigar que cada estação fosse responsável por buscar um novo nome para atendimento, além de realizar várias consultas ao banco de dados deixa o controle descentralizado. É preciso considerar que quanto maior esta base de dados, maior será o tempo gasto para que seja realizada a busca desse novo nome para trabalho.

Conhecer os atributos que podem classificar ou identificar um possível comprador e não poder utilizar estas informações para focar os contatos também se torna uma das grandes perdas de desempenho.

Com o passar do tempo notou-se que alguns dos operadores de *telemarketing* tinham maior afinidade ao telefone com pessoas de maior idade, outros tinham facilidade com pessoas da região nordeste do país, outros ainda tinham facilidade em falar com pessoas que possuíam uma renda superior. Além desses dados também puderam ser associados determinados tipos de produtos a certo tipo de clientes e outros dados que auxiliavam no aumento das possibilidades de se conseguir contato de sucesso com os clientes.

A descoberta de conhecimento nesse contexto está mais voltada à conquista de experiência. As pessoas responsáveis pelo estudo do desempenho das vendas acabam baseando suas decisões e análises conforme o que aprenderam com o tempo. Não existe nenhum método cientifico utilizado para tal além da estatística. Nessa área perda de tempo é perda de oportunidades e isso certamente influencia nos resultados.

O próprio fato de com o tempo se conhecer os perfis que tem os melhores resultados, mesmo assim ter controle total para direcionar estes nomes para o atendimento deixa muito

impotente o processo da busca de nomes. Apenas dividir tipos de clientes diferentes e agrupá-los em listas com cada perfil é uma saída, mas demanda tempo para fazer a divisão e necessita de processos especiais para fazer a divisão da lista de *mailing* recebida. Ter que tratar os dados antes da carga de nomes de *prospects* no banco de dados vai ocupar tempo e vai gerar várias listas diferentes. Já que muitos são os tipos de cliente gerados pela combinação de atributos são muitos os perfis, por exemplo, o perfil de clientes mais velhos e com tempo de conta maior que dois anos formarão um perfil, da mesma maneira a faixa etária combinada com 5 tipos de faixa de tempo de conta formarão 5 perfis. A quantidade de perfis será igual ao produto da quantidade Q1 de variações de um atributo A pela quantidade Q2 de atributo B (A(Q1) * B(Q2)). Ainda sobre esta questão, a definição de quais atributos é importante é mais uma dificuldade que somente é resolvida através da realização de testes e analise dos acertos e erros.

Centralizar as informações e decisões nas pessoas pode prejudicar o processo de venda em uma empresa de *telemarketing*. É necessário que os melhores caminhos sejam conhecidos ou estarem ao alcance de qualquer um e não apenas a quem esta acompanhando diariamente a operação de *telemarketing*. Na primeira troca de funcionário podem ser perdidas importantes informações sobre o trabalho. Matemática e estatística podem ajudar a definir os melhores perfis, mas existem informações importantes dentro de uma base de dados que não podem ser identificadas matematicamente.

Através da quantidade das vendas e dos totais de cada tipo de cliente é possível encontrar os tipos de cliente que mais compram, mas a relação de mais de uma característica não é tão clara. Por exemplo, pode-se saber que quem tem renda maior compra mais, mas não é explícito que quem tem renda alta somente compra uma capitalização determinada e quem adquire alguma com maior valor são aqueles que possuem uma estabilidade profissional há mais de 2 anos. Nesse exemplo quem tem estabilidade menor apenas adquire capitalizações de valores menores.

1.5 Soluções e problemas não resolvidos

Como uma das primeiras soluções para as dificuldades, a busca dos nomes acabou sendo retirada de cada estação e criou-se um *software* servidor que centralizaria a busca e a distribuição dos nomes disponíveis.

Para resolver o problema para disponibilizar os nomes foram criadas tabelas dentro do banco de dados a fim de que cada atributo dentro do banco de dados possuísse um

identificador e que tivesse um cadastrado para o mesmo. Por exemplo, no caso da data de nascimento foi criado um cadastro de faixas etárias, dessa forma através da data era calculado a que faixa o cliente pertencia. Essa solução não foi nada mais do que fazer a normalização dos dados.

Neste momento já é possível classificar os clientes através dos novos atributos que são chaves estrangeiras para as tabelas do banco de dados que possuem as faixas de atributos e o cadastro dos atributos. Assim, já poderia ser definido que, por exemplo, o melhor cliente era o que possuía a faixa de renda 2, faixa etária 3, etc.

Com os dados normalizados fica possível acrescentar ao sistema novas características. Se necessário, agora pode ser criado uma forma de todos os nomes serem importados juntos em lista única e a distribuição ser feita conforme cada atributo. Se o usuário que é responsável pela gerência definir que um atributo qualquer é responsável por aumentar o rendimento das vendas, os clientes com este atributo deverão ter prioridade em relação a outros e dessa forma o usuário terá um controle maior sobre os nomes que serão disponibilizados.

Nota-se que dessa forma, a divisão dos nomes que chegam até a empresa de *telemarketing* é dispensável visto que os nomes possuem uma forma de serem identificados de maneira mais eficiente e classificados.

Algumas das dificuldades ainda ficam por conta dos administradores do sistema, as tomadas de decisão ainda são feitas baseadas no conhecimento adquirido com o tempo e a escolha dos atributos mais importante ou do perfil melhor dos vendedores ainda é realizada com base no tempo de experiência dos gerenciadores da operação e por muitas vezes no conhecimento adquirido a cada erro ou insucesso.

Estas soluções já trazem para a campanha novas características e já pode ser mais bem aproveitada classificando os clientes e definindo prioridades no atendimento. O tempo que antes era perdido com os atendimentos a clientes com perfil inferior, agora fica para um segundo momento e aqueles que possuem as características comuns aos mais vendidos tem prioridade. Na busca da meta exigida, os melhores caminhos já estão disponíveis e certamente serão utilizados.

O problema atual está em conseguir adquirir mais conhecimento através do banco de dados. Através dos resultados obtidos no passado, conseguirem definir regras e associar

clientes com atributos diferentes, mas que formam uma classe de clientes com grande propensão de venda para os operadores.

Atualmente com o cenário comentado é preciso que além da utilização do conhecimento das pessoas, que existam maneiras de se definir níveis mínimos de desempenho para que o sistema aponte as principais combinações que permitam encontrar os melhores grupos de clientes.

Muitas informações importantes dentro da base de dados podem ou não estar visíveis ao usuário. Ter um mecanismo que traga ao usuário esta informação permitem que o sistema fique automatizado e que apenas seja necessário definir quais regras geradas pelo *software* serão utilizadas.

Apoiar as decisões em métodos científicos de busca de informações deve ser realizado para que os resultados sejam mais consistentes e que erros ou enganos sejam evitados.

Precisam-se criar diversas opções de ações que podem ser tomadas e que estejam disponíveis para que o usuário possa fazer comparações e optar pelas regras mais eficientes. É interessante que se tenham mais regras para que sempre seja possível ter uma segunda opção.

Não existe como fazer planejamento da utilização das listas de clientes. Embora se saiba que certo atributo é fundamental para definir que o *prospect* pode ser um possível comprador, disponibilizar todos os nomes apenas apoiado nessa informação pode deixar escapar outros dados que em combinação trazem um resultado diferente.

As próprias escolhas de quais atributos são importantes não tem fundamentação alguma e a combinação de atributos é realizada testando durante o período de trabalho. Sendo ineficiente resultará em tempo perdido e muitos clientes em potencial não serão contatados. Essa perda tem causado um custo para empresa e deveria ser evitado ou diminuído.

2 A DESCOBERTA DE CONHECIMENTO EM BANCO DE DADOS (KDD)

Em vários setores de nossa sociedade houve um crescimento dos dados das atividades do comércio, da medicina, dos negócios e da ciência que passaram a ser armazenados. Tanto foi o crescimento que estes dados saíram do controle daqueles que os armazenavam. Até o momento que a quantidade se mantinha em um volume pequeno, ficava fácil fazer uma análise dos dados através da visualização, os próprios SGBD disponibilizavam ferramentas que permitiam a visualização e a manipulação dos dados.

O nível da informação adquirida dentro dessas bases de dados era superficial e gerado com muito esforço. Todo o conhecimento adquirido ou disponível ao usuário é aquele que está diante de seus olhos e da sua capacidade de interpretar grande quantidade de registros. Entretanto, informações valiosas se perdem em meio aos grandes volumes de dados, e por isto novas tecnologias se mostram necessárias ao processo de recuperação de informação.

A descoberta do conhecimento em banco de dados, conforme introduzido, é definido pelo processo de identificar padrões válidos e novos que possam ser utilizados e compreendidos (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). A busca desses padrões significa encontrar relacionamentos e classificações que tornam possíveis encontrar informações surpreendentes dentro das bases. A área de descoberta de conhecimento passou a ser uma das áreas mais estudadas e mais desejadas da computação, conforme WIEDERHOLD (1996).

Não é difícil encontrar exemplos de situações em que muitos dados são armazenados e acabam não sendo aproveitados. Grandes redes de comércio acumulam milhares de registros de suas vendas. Um exemplo muito citado nesse ramo são as grandes redes que possui números muito elevados de vendas diárias. Na ciência, a área do estudo do corpo humano como os códigos genéticos humanos, gera um banco de dados com *gigabytes* de informações e combinações. A medicina ao recolher informações sobre pacientes com a mesma doença a

fim de encontrar características que possam determinar algo em comum entre os pacientes ou alguma fragilidade.

Registros da exploração de minério ou de petróleo são importantes para a continuidade do trabalho, da mesma forma se tornam de difícil análise pelos gestores, pela diversidade e quantidade de dados sobre toda a exploração.

Na área de tecnologia estudos de interpretação de imagens, podem utilizar as ferramentas de descoberta de conhecimento para aperfeiçoar processos de análise.

A grande quantidade de dados proporciona aos profissionais duas faces da situação. Com o auxílio dos bancos de dados muito mais informações podem ser armazenadas, maior é o histórico das transações e são guardados muito mais dados sobre o negócio. Por outro lado, por maior que seja a qualidade dos profissionais, chega um momento que a análise direta e por visualização se torna muito lenta ou então deixa de ser eficiente.

Percebe-se que existem pelo menos duas falhas que devem ser corrigidas. Primeiramente, esses dados devem ser processados de maneira mais rápida, para que os esforços possam ser direcionados a outras tarefas. A segunda melhoria seria incluir um método científico no processo de análise, garantindo, dessa forma, maior precisão e a descoberta de conhecimento contido dentro de milhares de registros.

Conforme WIEDERHOLD (1996), a tarefa de descobrir conhecimento não é simples. Os dados recolhidos e armazenados não são preparados de forma que a qualquer momento sejam analisados para que mostrem ao usuário os relacionamentos entre eles. Técnicas de KDD normalmente não são aplicadas em dados que já são alguns resultados de outro processo de KDD, mas aplicados em dados que podem ser de outros setores e áreas distintas[9]. Esses fatores são cruciais no momento em que é necessário atingir um nível alto de sucesso. Segundo (FREITAS, 1998), o tipo de método utilizado para que seja resolvida uma tarefa é o paradigma da descoberta de Conhecimento e para estes métodos é desejável eficiência, flexibilidade e generalidade.

Informações importantes podem não estar presentes na base ou podem ter sido registradas de forma incompleta. Tomando como exemplo a área comercial, ao se registrar os

[9] Essa origem, em geral não possui dados preparados para a mineração ou não pode ser utilizado sem uma nova transformação.

dados das vendas de um ano inteiro, muitos dos dados pessoais dos clientes podem ter sido alterados e com o passar do tempo isto contribui para que esses dados estejam desatualizados.

Outra situação que pode ocorrer é o caso da área coletora dos dados ter deficiência durante a coleta e por conseqüência, no fim da captação muitos atributos podem ser perdidos ou não identificáveis. A união de mais de uma base de dados pode ser utilizada para permitir que o modelo se torne completo e rico de informações.

Torna-se muito importante ter um bom modelo que possa proporcionar ao usuário a visualização e relacionamento entre todas as variáveis consideradas candidatas, inclusive adicionar aquelas que não estão disponíveis. Dessa maneira é possível definir qual é modelo que será utilizado para auxiliar nos possíveis problemas de falta de informação, que pode prejudicar o resultado, conforme (WALKER, apud WIEDERHOLD, 1996).

A falta de informação dentro da base a ser utilizada se torna um risco para alguns estudos (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992), por exemplo, embora se chegue à conclusão que o poder aquisitivo do cliente é fator fundamental para a aquisição de um título de capitalização, ainda pode existir outro fator que não foi considerado e que influencia fortemente a compra dos títulos.

O processo de KDD possui alguns passos que são altamente interativos e “em princípio, de qualquer passo talvez necessite voltar para um passo anterior” (TRADUÇÃO NOSSA), conforme (FREITAS, 1998, p. 41).

Os dados precisam ser tratados a fim de retirar as imperfeições e informações irrelevantes para fazer a busca de padrões e após, serão avaliados, para definir se todo o processo precisa ser refinado para obter maior qualidade. KDD com o auxílio de algoritmos pode gerar alguns padrões que o usuário pode interpretar e utilizar.

Uma característica muito importante para levar em consideração é que a descoberta do conhecimento não se dá exclusivamente por algoritmos e métodos, é preciso que exista a intervenção humana a fim de definir quais são os níveis e interpretar se as respostas geradas são úteis.

Para ilustrar o processo de KDD, segue a Figura 2.1, representando as suas várias etapas.

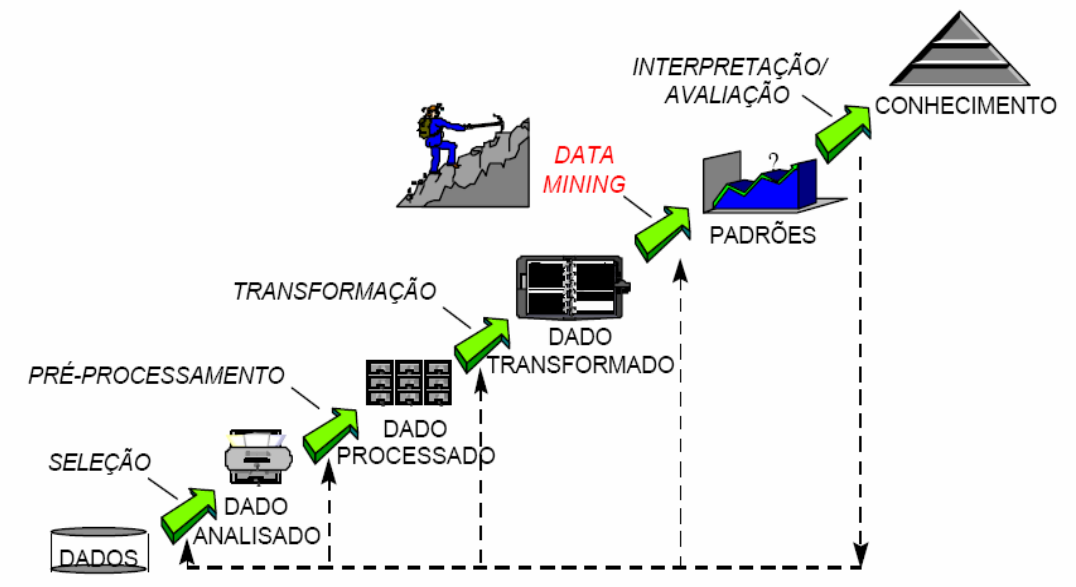

Figura 2.1 – O processo de KDD
Fonte: OLIVEIRA et al., 2002, p. 2

Antes mesmo de se aplicar as técnicas escolhidas, é preciso que a pessoa defina qual será seu foco e que possua algum conhecimento dentro da área onde os dados forem captados para que possa fazer a sua interpretação. O usuário deve eleger as principais variáveis para montar um modelo para testes.

O pré-processamento e limpeza dos dados se tornam essenciais para que seja removido tudo aquilo que não é necessário ou importante. Dificilmente se encontrará alguma base de dados que esteja totalmente preparada para ser processada por alguma técnica de KDD e, por este motivo, que WIEDERHOLD (1998) descreve como maior barreira para a aquisição de conhecimento os próprios dados armazenados.

Na análise de um público investidor em títulos de capitalização, não se pode afirmar que o perfil da base analisada é o mesmo para todos os investidores: mudanças financeiras e de confiança podem interferir. Nessa situação, podem existir outras variáveis que interferem nos números, porém ao analisar são nulas ou indiferentes. Se o número de variáveis envolvidas for muito grande pode ser necessário que seja feita uma redução da quantidade destas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Ao se definir um número muito grande de variáveis pode-se criar uma diversidade muito grande de resultados, todos com as características semelhantes e que não trazem informações úteis.

BRACHMANN e ANAND (1996) reforçam a afirmação que o processo de KDD não deve ser automático e deve ser assistido e orientado por um humano, para que os sistemas de descoberta nos auxiliem a entender melhor como adquirir mais conhecimento em grandes

bancos de dados. O ser humano sozinho não possui capacidade e velocidade o bastante para processar tantas informações e gerar conhecimento não explícito. Os sistemas por sua vez resolvem esse problema, mas não possuem ainda o bom senso necessário para interpretar os dados e apenas retornar aquilo que é interessante e que seja significante de fato. O que é importante considerar no resultado de um processo de KDD é que a informação gerada é tão pessoal, pelo fato do usuário determinar quando é relevante e acionável, que o conhecimento deve ser estruturado para que o sistema ou outras pessoas possam utilizar.

2.1 Arquitetura de KDD

Os passos do processo de KDD envolvem preparação dos dados e a mineração dos dados que é a parte fundamental da arquitetura de KDD, onde será feita a escolha dos algoritmos utilizados e qual técnica será aplicada. Após ser realizada a busca dos padrões (aplicação da técnica), chega o momento de fazer a interpretação. Como o processo de KDD é realizado de forma que se não atingir os objetivos propostos pode ser necessário reiniciar a rotina. Pode ser refeito até que a resposta possa ser considerada sólida e agregue algum tipo de conhecimento que não seria possível identificar simplesmente pela visualização dos dados.

Uma peça fundamental do processo de KDD, certamente é a figura humana. Não há uma maneira de medir exatamente se os objetivos foram bem definidos, se as metas são corretas e avaliar se o resultado atingiu o que era esperado. O amadurecimento dos resultados se dá exclusivamente com o intensa busca por melhorias. Em relação ao envolvimento de pessoas durante o processo de KDD, a partir das experiências de trabalho de estudiosos, BRACHMAN e ANAND (1996, p. 39-40) consideram que:

> Descoberta de Conhecimento é uma tarefa de aprendizado intensivo consistindo de complexas interações, protegido todo o tempo, entre um humano e uma grande base de dados, possivelmente auxiliado por uma heterogênea suíte de ferramentas. [...] Para o maior sucesso do desenvolvimento de ferramentas de suporte a descoberta de conhecimento, é necessário entender a exata natureza das relações entre humanos e os dados que levam a descoberta de conhecimento. (TRADUÇÃO NOSSA).

A arquitetura do processo de KDD pode ser composta de cinco fases, conforme figura 2.1, que auxiliarão na difícil tarefa de transformar os dados em informações (FAYYAD et al., 1996). Conforme já comentado, o processo de KDD normalmente inicia com os dados armazenados sem nenhum tipo de tratamento ou modelagem especial para que seja aplicada a descoberta de conhecimento. Estes dados armazenados de maneira “bruta” e representam o início do processo. Nesse momento, o problema é definido e são estipulados os objetivos a serem alcançados. É muito importante levar em consideração o custo/benefício da análise dos

dados, apenas realizar o processo de descoberta de conhecimento, sem um motivo ou sem um foco, traz uma chance muito grande de que apenas se perca tempo e não se consiga nenhum resultado expressivo.

A primeira etapa trata da seleção dos dados, onde são identificados os dados a serem utilizados e que bases de dados estão localizadas. Em relação aos dados, é importante levar em consideração algumas características: nesse ponto do processo de KDD pode ser necessária a junção de mais de uma base de dados e serão definidos quais os dados dessas ou dessa única base serão utilizados. As empresas acabam armazenando informações importantes para seus negócios e que podem ser aproveitadas no processo de KDD, mas isso não livra da necessidade ou priva esta de buscar informações de fontes externas.

Como exemplo de seleção, pode ser considerado a mineração de informações sobre as vendas de produtos importados em um site de compras pela internet quando os clientes adquiriram somente produtos que custaram mais de R$ 100, durante um mês, mas além desses dados deve-se incluir a cotação do dólar comercial de cada dia do período selecionado. Esta informação da cotação do dólar neste exemplo não estava presente junto aos dados e dessa forma foi obtida externamente agregando um outro dado que pode ser muito relevante na descoberta de conhecimento. Segundo BRACHMAN e ANAND (1996, p. 42), é o "analista o responsável pelos objetivos, realizando as consultas na base de dados para extrair dados relevantes ao objetivo" (TRADUÇÃO NOSSA).

Na etapa do pré-processamento dos dados é realizada a adaptação da base para que possa ser aplicada a mineração de dados. Essa etapa se torna necessária quando a base não está preparada ou a integração de duas bases acarrete na falta de consistência dos dados. Se a aplicação da descoberta de conhecimento fosse realizada através do tempo de conta no banco[10], esse dado do cliente possivelmente estaria armazenado em forma de data, sendo necessário que fosse feita a conversão para meses. Outra situação que pode ocorrer é que o formato das datas de duas bases, por exemplo, seja diferente ou uma delas já estar no formato de meses. A limpeza dos dados é uma tarefa que garante qualidade ao processo e nesse ponto cabe a pessoa que está realizando o pré-processamento remover os casos em que falta alguma informação ou que os dados estejam corrompidos.

A próxima etapa é a transformação, onde os dados de entrada serão recebidos do pré-processamento, de maneira que estarão formatados diferentemente de quando não haviam

[10] O tempo de conta representa a quantidade de meses que o cliente possui vínculo com a instituição.

sido pré-processados. A transformação será exclusivamente para que os dados já formatados sejam organizados de maneira que a ferramenta e/ou técnica escolhida possa realizar a garimpagem. Cada ferramenta de mineração de dados e/ou técnica pode ter uma maneira especial de receber os dados. As etapas de seleção, processamento e transformação formam a preparação dos dados em um processo de KDD.

As próximas etapas serão compostas pela própria garimpagem dos dados, onde serão escolhidos os algoritmos e finalmente pela análise dos resultados a fim de identificar se o conhecimento é relevante (CARVALHO; SAMPAIO; MONGIOVI, 1999). No momento da análise é que vai ser definido se vai ser necessário que o processo seja reiniciado no caso do resultado não esteja dentro dos objetivos definidos.

2.2 Mineração dos dados (MD)

A mineração ou garimpagem de dados, segundo CABENA et al. (1997, p. 12) é o "processo de extrair previamente informação não conhecida, válida e acionável de grandes bases de dados e então utilizar a informação para realizar cruciais decisões no mundo dos negócios" (TRADUÇÃO NOSSA). A mineração de dados (MD), muitas vezes confundida com KDD, na verdade trata apenas de uma das etapas de todo o processo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Data Mining (DM) além de ser uma área interdisciplinar (FREITAS, 1998), segundo DWBRASIL! (2007) nasceu da estatística, inteligência artificial e outra área que mistura essas duas, chamada de aprendizagem de máquina onde os programas adquirem conhecimento utilizando os conceitos da estatística e os algoritmos e técnicas da inteligência artificial. Deixando claro o objetivo de DM no processo de KDD, "a mineração busca selecionar o algoritmo ou os algoritmos para processar os dados" (TRADUÇÃO NOSSA) (CABENA et al., 1997, p. 55).

Uma característica importante de DM, são seus algoritmos, cada um deles têm específicas entradas e saídas, também conhecidas como suas técnicas (CABENA et al., 1997). Como mais conhecidas e utilizadas pode-se citar a descoberta de regras de associação, classificação, *clustering* e regressão.

2.2.1 Regras de associação

A descoberta de regras de associação, como o próprio nome resume, baseia-se na descoberta de regras que possam associar itens dentro da base e probabilisticamente, através do cálculo da freqüência de um item, definir as melhores regras.

As regras que podem ser consideradas melhores são aquelas onde o percentual de acerto é maior em relação aos outros. Por exemplo, uma regra para identificar as vendas de títulos de capitalização define que o melhor cliente é o que possui 22 anos, renda maior que R$800 e que mora na região sul. Essa regra se aplica a 81% do montante das vendas. Para que outra regra possua maior qualidade deve possuir um percentual maior que 81% dos registros da base.

Segundo AGRAWAL e SRIKANT (1994, p. 487), "o problema de mineração de regras de associação é gerar todas as regras de associação que tenham suporte e confiança maiores que o mínimo suporte especificado pelo usuário [...] e o mínimo de confiança [...]" (TRADUÇÃO NOSSA).

A regra de associação é definida de maneira que se "X então Y" ou "$X \Rightarrow Y$", onde X e Y são conjuntos de itens e $X \cap Y = 0$. Para essas regras X é definido como o antecedente e Y é o conseqüente (CARVALHO; SAMPAIO; MONGIOVI, 1999). Para que uma regra possa ser avaliada, os dois fatores são considerados: Suporte e Confiança. O suporte é o número de transações que contém o conjunto de item, divido pelo total de transações, ou seja, o suporte é a freqüência em determinado conjunto de itens, já que identifica a quantidade dos registros que possuem os atributos analisados em relação a todos os presentes na amostra.

Suporte = nº. registros com X e Y / nº. total de registros

A confiança é o cálculo para a regra "Se X então Y", onde o total de registros X e Y são divididos pelo total de registros X

Confiança = nº. registros com X e Y / nº. registros com X

O índice chamado de confiança representa a freqüência com que um conjunto de itens é comercializado em relação às vendas antecedentes, ou seja, é medida a confiança que se pode depositar em uma análise, de que registros de um conjunto possuirão características associadas a outras, em quantidade maior que o mínimo definido e esperado por quem aplica a técnica.

Para descobrir todas as regras de associação, existem alguns problemas envolvidos que podem ser caracterizados como:

- “Procurar todos os grupos de item (*itemsets*) que tenham suporte da transação acima do suporte mínimo” (TRADUÇÃO NOSSA) (AGRAWAL; IMIELINSKI; SWAMI, apud AGRAWAL; SRIKANT, 1994, p. 488). Conforme já discutido, o suporte é o número de tansações analisadas que possuem os *itemsets*, sendo que os grandes grupos são os que possuem o mínimo, ao contrário são pequenos grupos (AGRAWAL; SRIKANT, 1994).

- “Usar os grandes grupos de itens para gerar as desejáveis regras” (TRADUÇÃO NOSSA) (AGRAWAL; SRIKANT, 1994, p. 488). Para um grande grupo G, são geradas as regras para cada um dos subgrupos S no formato $S \Rightarrow (G - S)$, removendo S quando possuir valor inferior de confiança mínima (AGRAWAL; SRIKANT, 1994).

Como exemplo de algoritmo de regras de associação pode ser citado o algoritmo Apriori, um dos mais conhecidos e utilizados na área. Nesse tipo de técnica é mais comum reunir os registros de venda e selecionar os atributos para identificar as características que formam um perfil de comprador com potencial.

2.2.2 Classificação

A classificação é uma técnica de aprendizado supervisionado, ou seja, os resultados precisam ser analisados por um especialista para fazer a avaliação de relevância. A classificação gera modelos a partir de exemplos dentro de uma base, que são chamados de conjunto de treinamento, que deve ser uma amostra dos registros que serão analisados (CABENA et al., 1997). A especialização da classificação apresentada nesse trabalho é a indução de árvores (*tree induction*), técnica que “constrói um modelo preeditivo na forma de árvores de decisão” (TRADUÇÃO NOSSA) (CABENA et al., 1997, p.51).

As árvores de decisão representam uma árvore de forma invertida, onde as raízes passam a ser folhas e essa hierarquia é disposta de forma que ao seguir a estrutura é possível tomar as decisão e executar a tarefa da maneira que foi proposta, já que em cada nível as opções a serem tomadas são os nós do nível seguinte e as decisões são tomadas até que sejam atingidos os nós terminais (MONGIOVI, 1998). Na classificação todos os registros fazem parte de classes que são identificadas pelo atributo objetivo e os registros vão conter esse atributo e um conjunto de atributos “previsores”. Através das classes que já são conhecidas no conjunto de teste (relacionamentos encontrados), o objetivo então é descobrir o atributo que é

"meta[11]" naqueles registros que ainda não foram classificados (base a ser analisada) (FREITAS, 1998).

Como exemplo de classificação, podemos separar 100 registros de pessoas que foram contatadas pela operação de *telemarketing* e definir que nosso atributo meta serão os clientes que compraram ou não o título de capitalização. Nosso objetivo com essa classificação é encontrar em outros 2000 clientes ainda não contatados àqueles que através da árvore criada são prováveis compradores. Os métodos automáticos de classificação podem ser indutivos ou de indução neural (também conhecidos como conexionistas por alguns autores). Os algoritmos indutivos são aqueles que geram as árvores de decisão e os conexionistas são as redes neurais (CABENA, 1997).

2.2.3 Clustering

Clustering, ou ainda agrupamento, é uma técnica que visa criar classes e agrupa os registros com atributos semelhantes. É um tipo de aprendizagem não supervisionada, o que quer dizer que o resultado não requer avaliação do usuário. "*Clustering* é uma tarefa descritiva comum onde uma semente identifica um grupo finito de categorias ou *cluster* para descrever os dados" (TRADUÇÃO NOSSA) (JAIN; DUBES, apud FAYYAD et al., 1996, p.14).

Essa técnica pode ser utilizada para realizar uma análise inicial dos dados e assim obter uma visão geral dos agrupamentos existentes na base de dados e então, analisar a fim de definir quais grupos serão utilizados ou levados em consideração. Pode-se, após essa técnica, utilizar outra como a classificação. Na figura 2.2, pode-se observar um gráfico fictício de que mostra um *clustering* aplicado a algum tipo de dado, onde foram identificados 3 clusters.

[11] É o que se deseja descobrir dos registros com base no resultado do conjunto de treinamento.

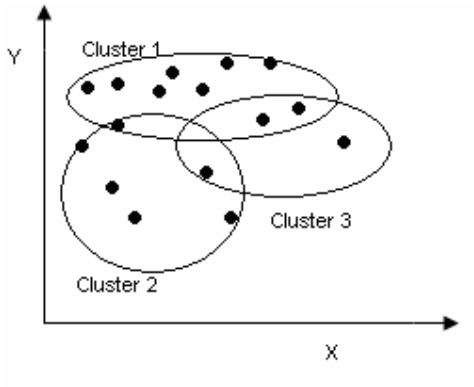

Figura 2.2 – Representação de 3 *Clusters* gerado com a técnica
Fonte: Adaptado de FAYYAD et al., 1996, p. 14

Como característica de *clustering* pode-se destacar que não há um atributo meta e todos possuem a mesma importância. A técnica é utilizada especialmente para a exploração e sumarização dos dados (FREITAS, 1998).

Como exemplos, as características dos clientes que já adquiriram títulos de capitalização podem ser analisadas. A idéia de aplicar *clustering* nesse caso é descobrir grupos ou categorias de clientes. Essas categorias são geradas através dos atributos como idade, renda, residência, tempo de relacionamento com a instituição financeira, sexo, etc. Ao ser feita uma média simples de idade e renda, chega-se a uma média de clientes com 37 anos e renda de 2000 reais, porém existem alguns clientes com idade muito elevada e com renda menor que 2000 reais. Dessa maneira, apenas através de uma média tem-se a impressão de que a maioria dos registros possui realmente essa idade e renda.

Ao aplicar o *clustering*, se deseja conhecer quais são os grupos de clientes conforme características semelhantes existentes na base. Continuando o exemplo, é encontrado um *cluster* de idade próxima 24 anos e renda aproximada R$5000 reais, segundo a quantidade de registros que se enquadram nesse *cluster,* torna-se possível identificar que é grande o bastante, em relação ao total, para ser dado foco aos *prospects* com essas características no momento do atendimento.

A aplicação da técnica dá ao usuário uma visão geral de todos os registros da base, já que vários *clusters* podem ser identificados e conforme sua quantidade é possível definir os mais importantes. Outra característica do resultado da técnica é que fica possível entender a tendência de aquisição do produto conforme as características. Ao colocar as dados em um gráfico, por exemplo, nota-se que a quantidade clientes compradores cresce conforme vai

aumentado um atributo X e diminuindo outro Y, ao contrário ou ainda ao aumentar e diminuir ambos.

2.2.4 Regressão

"A regressão é uma função de aprendizado que mapeia um atributo para uma variável preeditiva real-validada" (TRADUÇÃO NOSSA) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996, p. 13). A regressão serve para estimar a probabilidade de certo fenômeno.

Nessa técnica é gerada uma linha de regressão que representa a relação entre duas variáveis representadas na figura 2.4 em forma de gráfico. A linha desenhada na diagonal no gráfico representa a função linear onde se classificam os registros. Para os próximos dados a serem analisados, aqueles que estiverem próximos dessa linha tem maior probabilidade de possuírem o mesmo resultado.

Exemplificando, sendo X a renda e Y o valor do título comprado, se obtém a função linear que determina o valor de título adquirido conforme o valor da renda do cliente. Na figura 2.3, é apresentada a representação de uma linha de regressão.

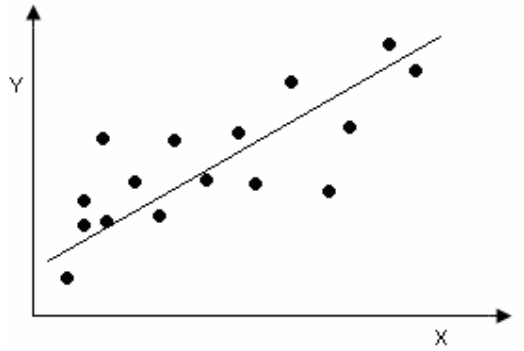

Figura 2.3 – Representação da Linha de Regressão
Fonte: Adaptado de FAYYAD et al., 1996, p. 14

A forma mais simples de regressão é a linear, utilizada na figura 2.3. São utilizadas duas variáveis, onde uma delas é aleatória (y), que é a função linear de outra deste mesmo tipo (x), formando a equação (y = a + bx) (LÓPEZ; HERRERO, 2004). "Nesta equação a variação de *y* se assume que é constante, e *a* e *b* são os coeficientes de regressão que especificam a intersecção com a linha de ordenadas, e a inclinação da reta" (TRADUÇÃO NOSSA) (LÓPEZ; HERRERO, 2004, p. 56). O coeficiente de regressão é medido através do método

dos mínimos quadrados, que utilizam as equações abaixo, para diminuir os erros dos dados e da estimativa da linha (PRESS et. al, apud LÓPEZ; HERRERO, 2004).

$$b = \frac{S_{xy}}{S_x^2}$$

$$a = y - bx$$

Ao obter S registros de exemplo com seus pontos (X1, Y1), (X2, Y2) ... (Xs, Ys) é possível obter os coeficiente com essas equações, onde Sxy é a covariância de x e y, enquanto a variância de x é representada por Sx^2.

Na figura 2.4, tem-se um exemplo de regressão linear que representa vendas de capitalização, estudando a relação do valor da renda e do valor dos títulos adquiridos:

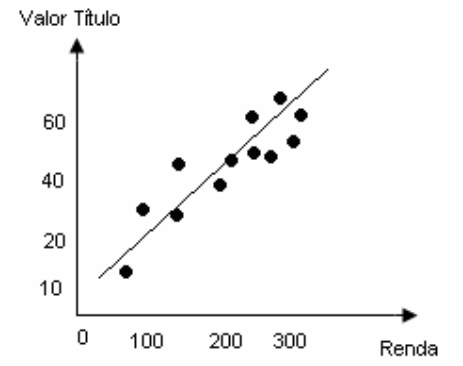

Figura 2.4 – Regressão linear do valor de renda e valor de título adquirido
Fonte: FIGURA NOSSA

No exemplo da figura 2.4, pode-se perceber esta relação, onde a regressão identificou que a aquisição do valor do título está ligada com o valor da renda. O aumento desse total sugere que a aquisição de título de capitalização se dá conforme o aumento do valor da renda.

A regressão linear múltipla segue os mesmo princípios da regressão linear, porém utilizando mais de uma variável preditiva (x) e a variável (y) como uma função linear de um vetor multidimensional ($y = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$).

Segundo LÓPEZ e HERRERO (2004, p. 60-61), ainda existem as regressões lineares ponderadas localmente, que "geram modelos durante o processo preeditivo" (TRADUÇÃO NOSSA), dando peso diferente para os exemplos de treinamento e a regressão não linear utiliza uma função polinomial para os dados que o resultado depende do valor das variáveis independentes da função polinomial.

2.3 Interpretação e avaliação dos dados

Esta última etapa, embora seja a mais simples, é uma das mais importantes. Nesse momento do processo de KDD é que o usuário, que deve sempre acompanhar o processo, define se a resposta da mineração é útil.

É algo comum na mineração de dados, que os algoritmos utilizados tragam até o usuário alguma informação que não pode ser considerada como relevante. Por exemplo, ao aplicar uma regra de associação se obtém uma regra indicando que as pessoas com renda muito alta compram títulos de valor maior, logo, para este caso, a informação já era conhecida.

Quando o usuário interpretar o resultado do processo de KDD, ele vai identificar a necessidade ou não de reiniciar o processo e gerar outro tipo de regra ou informação, se as obtidas não forem acionáveis. Após a avaliação, se o conhecimento gerado for considerado relevante é momento então de consolidar o conhecimento gerado e incorporar este dentro dos sistemas, documentar ou então utilizar na tomada de decisões (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

3 WEKA

A ferramenta Weka[12] é um *software open source* implementado com a linguagem Java, desenvolvido pela Universidade de Waikato (WITTEN; FRANK, 2005) e disponível em (WEKA, 2007). O *software* implementa diversos algoritmos de mineração de dados, possibilitando ao usuário gerar arquivos de texto (.arff) para serem analisados. Na figura 3.1 pode-se verificar o formato do arquivo .arff:

```
%Informações dos registros de venda na base de teste
@relation perfil-vendas

@attribute CLASSIFICACAO        {BASE M2,CAPITALIZACAO,CAPITALIZACAO M2,EXCLUSIVO,INSTITUCIONAL,PLATINUM,UNICO,NONE}
@attribute COD_REGIAO_MAILING   {1,2,3,4,5,6,7,8,9,10,11}
@attribute COD_CLASSIFICACAO    {1,2,3,4,5}
@attribute DIA_VENC_FATURA      {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,24,25,26,27,28,29,30}
@attribute COD_TEMPO_CONTA      {1,2,3,4,5,6,7,8}
@attribute COD_FAIXA_ETARIA     {1,2,3,4,5,6}
@attribute COD_VALOR_LIMITE     {1,2,3,4,5,6,7,8,9,10,11,12,13}
@attribute COD_SEXO_ATB         {1,2}

@data
%
% 12 instances
%
INSTITUCIONAL,4,4,1,1,1,1,3
UNICO,11,2,1,1,1,1,3
PLATINUM,7,3,1,1,1,1,3
PLATINUM,8,3,1,1,1,1,3
EXCLUSIVO,7,5,1,1,1,1,3
PLATINUM,7,3,1,1,1,1,3
EXCLUSIVO,7,5,1,1,1,1,3
UNICO,4,2,1,1,1,1,3
UNICO,8,2,1,1,1,1,3
UNICO,7,2,1,1,1,1,3
UNICO,7,2,1,1,1,1,3
INSTITUCIONAL,6,4,1,1,1,1,3
```

Figura 3.1 – Arquivo .arff do Weka
Fonte: WEKA, 2007

O arquivo arff é separado por *tags* que identificam a estrutura do arquivo. A *tag* @relation identifica o nome do arquivo. Cada *tag* @attibute identifica no arquivo os atributos presentes e seus tipos. A seqüência das *tags* é seguida conforme a seqüência aqui apresentada. Para marcar o início dos dados a serem analisados é inserido no arquivo @data, abaixo desta

[12] O *software* foi batizado dessa maneira pelas iniciais de *Waikato Enviroment Knowledge Analysis*

tag, todos os dados serão analisados. A presença de % no início da linha comenta toda a linha e então não será considerada.

A ferramenta permite ao usuário alterar os parâmetros de entrada dos algoritmos através de uma interface gráfica que facilita a sua utilização. A ferramenta é formada por pacotes, específicos para cada aplicação de técnica de mineração de dados. Os pacotes são: *attribute selection*, *classifiers*, *clustering*, *association rules*, *filters* e *estimators*. O primeiro pacote *weka.atributeSelection* que seleciona os atributos em uma base de dados ou arquivo. Os pacotes *weka.classifiers*, *weka.cluster*, *weka.association* possuem as implementações dos algoritmos de cada técnica, respectivamente. O pacote *weka.estimators* possue subclasses que servem para "computar os diferentes tipos de distribuição de probabilidade" (OLIVEIRA et al., 2002, p. 4). O pacote *weka.filters* permite selecionar um conjunto de atributos ou instância de dados (OLIVEIRA et al., 2002).

O Weka vem sendo utilizado em vários trabalhos de DM, torna fácil sua utilização pela portabilidade e facilidade de implementação. Essa ferramenta foi escolhida por possuir os principais algoritmos e técnicas que serão estudadas e já ter sido utilizada com sucesso por outros trabalhos. Como exemplos de utilização do Weka, podem ser citados trabalhos como (OLIVEIRA et al., 2002), (PINHEIRO, 2006) e (LOPÉZ; HERRERO, 2006).

O algoritmo de associação implementado pelo Weka está localizado no pacote *weka.association* em duas classes *ItemSet* e *Apriori,* que são responsáveis pelo algoritmo *Apriori* que é um dos mais conhecidos para esse tipo de técnica. Entre os algoritmos de classificação presentes no Weka estão incluídos os algoritmos *weka.classifiers.J48.J48* e o *weka.classifiers.J48.PART*. Já o pacote *weka.cluster* implementa, por exemplo, dois algoritmos de aprendizagem não supervisionada: *Cobweb* e *EM*. O Weka também possui algoritmos para regressão que podem ser selecionados dentre os algoritmos de classificação ou através do pacote *weka.classifiers.functions*. A seguir, serão apresentados alguns algoritmos presentes no Weka.

3.1 Algoritmo de regra de associação

O algoritmo Apriori se encarrega de fazer a busca por regras de associação entre os dados contidos dentro da base. É importante ressaltar que, na maioria das vezes, é preciso que se faça um tratamento dos dados antes de serem processados pelo algoritmo. A fim de recuperar mais precisas e melhores informações, a base de dados a ser analisada deve ser possuir somente atributos considerados importantes. Ao retirar os dados que não são

relevantes ou que não devem fazer parte da análise, o algoritmo poderá realizar sua função de modo mais eficiente.

Considerando que este estudo trata de dados que estão normalizados dentro de um SGBD, estes estarão agrupados conforme a necessidade do *software* criado para realizar as vendas com os *prospects*. Junto com informações importantes como idade, residência, valor de renda encontram-se chaves primárias de tabelas, campos de texto com observações e outros dados que são de difícil agrupamento e talvez códigos que são exclusivamente de controle do *software* ou da própria empresa contratante que envia os cadastros de *prospects*. Para retirar estes dados desnecessários em um primeiro momento, pode ser feito um tratamento e recuperar somente aqueles que são inicialmente classificados importantes pelo usuário. Com a seqüência de experimentos sendo realizados é que se atinge a maturidade em relação aos dados que são realmente relevantes.

O algoritmo *Apriori*, através da análise dos dados busca recuperar conjuntos de itens que acabam sendo freqüentes e a estes conjuntos é dado o nome de *itemsets* freqüentes (L_K). Seu objetivo é encontrar todos estes conjuntos presentes na base analisada. Para aprimorar seus resultados, estes conjuntos são recuperados em uma quantidade mínima estipulada. Isso garante definir o nível de critério da associação

O algoritmo é composto por uma estrutura principal e mais duas estruturas secundárias, que são funções utilizadas pela parte principal. Uma delas realizando a geração de candidatos e eliminando os que não são freqüentes (Apriori-gen) e a outra gerando as regras de associação.

O algoritmo trabalha através de dois passos, um deles é a geração onde são criadas as combinações encontradas no arquivo e o outro passo trata de fazer o corte das combinações que não aparecem na freqüência desejada e pré-estipulada (Suporte e Confiança). Na Figura 3.2, é apresentado o algoritmo Apriori:

```
Function Apriori(Banco de Transações D, Smin) Grandes Conjuntos de Itemset (L)

L1 = {grande 1-itemsets};
k = 2;

Enquanto Lk-1 <> 0 faça
        Inicio
                Ck = apriori-gen(Lk-1; {gera os novos candidatos}
                Para toda  transação t Є D faça

                        Inicio

                                Ct = subconj(Ck, t); {candidatos contidos em t}

                                Para todo candidato c Є Ct faça

                                        contador(c) = contador(c) + 1

                        fim;

                Lk = {c Є Ck | contador(c) ≥ Smin};

                k = k + 1

        fim;

Retorna (L = Uk Lk).
```

Figura 3.2 – Algoritmo Apriori
Fonte: Adaptado de MONGIOVI , 1998

O algoritmo Apriori realiza a contagem dos grandes grupos e em seguida (k), são gerados candidatos (C_k) dentro de cada grande grupo (L_{k-1}) existente no (k-1) passo do algoritmo e verificado o suporte dos candidatos (AGRAWAL; SRIKANT, 1994).

Para que fique mais claro, é possível fazer uma simulação do funcionamento do *Apriori*. Para isso vamos considerar um pequeno exemplo com dados sobre algumas vendas[13] e seus atributos.

Na tabela 3.1, é apresentado um exemplo de dados para aplicar o algoritmo, que serão os *itemsets* (L), a coluna venda identifica o identificador de cada registro e o restante dos dados representa a presença da característica (1) ou então a ausência da mesma (0):

[13] Vendas de títulos de capitalização de uma instituição bancária.

Quadro 3.1 – Quadro de Vendas de Capitalização

Venda	Masculino	Renda > 1000	Reside em SP
1	1	1	1
2	0	1	1
3	1	1	1
4	1	1	1
5	1	0	1
6	1	0	1
7	1	1	1
8	1	1	1
9	1	1	0
10	1	1	0

No primeiro passo do algoritmo principal, este deve determinar os *itemsets* freqüentes através da contagem das ocorrências e em seguida nos passos (k), realizar outras operações. Pode ser definido como suporte mínimo 0,6. Dessa forma, já é possível separar os grandes conjuntos que são:

Tabela 3.1 – Grande grupo de um elemento

Atributo	Suporte
Masculino	0,9
Renda	0,8
SP	0,8

Tabela 3.2 – Grande grupo de dois elementos

Atributo	Suporte
Masculino, Renda	0,7
Masculino, SP	0,7
Renda, SP	0,6

Para calcular o suporte de cada grupo, é aplicada a fórmula do suporte. Para ilustrar, o suporte do atributo “Masculino”, do grupo de um elemento, é feito através do total de registros “Masculino” (9 registros) dividido pelos 10 registros que são o total da amostra, obtendo o suporte 0,9. Para o grupo de dois registros é feito o cálculo, quando os dois atributos são 1, divido pelo total: 7 (Masculino e Renda) / 10 (Total) = 0,7.

O conjunto abaixo acabou sendo cortado pelo fato do suporte não ter atingido o valor mínimo estipulado, que no caso foi 0,6.

Tabela 3.3 – Grupo que não foi selecionado pela simulação

Atributo	Suporte
Masculino, Renda, SP	0,5

Nestes próximos passos chamados de K, a primeira ação a ser tomada é gerar os chamados *itemsets* candidatos (C_k), conforme tabelas 3.1 e 3.2 e 3.3, que são *itemsets* encontrados em (k-1) e são definidos como possíveis *itemsets* freqüentes. Nesse ponto a função Apriori_gen é utilizada para gerar um conjunto de todos *itemsets* freqüentes, levando-se em conta de que se o suporte mínimo foi alcançado, os conjuntos abaixo deste também alcançarão este mínimo.

Os *itemsets* são comparados e retirados os *sub-itemsets* (c_k) que pertençam a C_k e que não pertençam (k-1) ao *itemset* L_k-1. Após isso, o algoritmo deve fazer uma nova busca entre os dados levando em conta o nível de suporte encontrado em cada candidato definido.

A função Apriori-gen tem a função de unir os elementos de L_{k-1} a cada 2 e reter apenas aqueles em que todos os seus subconjuntos de tamanho k-1 pertençam a L_{k-1}. Na figura 3.3, pode-se verificar a função Apriori-gen:

```
Apriori-gen(Lk-1)

    {Junção}
    Insert into Ck
    From  Lk-1p, Lk-1q {elementos p e q de Lk-1}
    Select p.item1 = q.item1, ...p.itemk-2 = q.itemk-2
           p.itemk-1 < q.itemk-1;

    {Poda}
    Eliminar todo c que pertence a Ck tal que algum (k-1)
    subconjunto de c não pertence a Lk-1
           Ck = {c ∈ Ck | ∀ s ⊂ c, s ∈ Lk-1}

    Para  todo c ∈ Ck faça
           Para todo s ⊂ c e |s| = k-1 faça
                  Se s não pertence a Lk-1 então elimina c de Ck
```

Figura 3.3 – Função Apriori-gen
Fonte: Adaptado de MONGIOVI, 1998

Para ilustrar o Apriori-gen, será definido como nível de confiança 0,8. Aplicando a fórmula que define a confiança citada anteriormente resultará em:

Tabela 3.4 – Regras com sua classificação de confiança

Regra	Fator de Confiança	Além do mínimo?
{Masculino} → Renda	0,77	NÃO
{Renda} → Masculino	0,87	SIM
{Masculino} → SP	0,77	NÃO
{SP} → Masculino	0,87	SIM
{Renda} → SP	0,75	NÃO
{SP} → Renda	0,75	NÃO
{Masculino, Renda} → SP	0,62	NÃO
{Masculino, SP} → Renda	0,62	NÃO
{Renda, SP} → Masculino	0,55	NÃO

Por fim, a geração das regras é feita ao percorrer os *itemsets* freqüentes e descobrir os *subsets* que não estão vazios quando o suporte do *itemset* freqüente dividido pelo suporte do *subset* atender ao mínimo de confiança definido no algoritmo.

As primeiras 6 regras são geradas a partir do grande grupo de um elemento (Tabela 3.1). Para calcular a confiança de cada regra é aplicada a fórmula quando o atributo possui o valor igual a 1. Por exemplo, a regra {Masculino} → Renda é calculada, dividindo a quantidade de registros "Masculino" e "Renda" (7 registros) pela quantidade de registros "Masculino" (9 registros), resultando na confiança 0,77. Da mesma maneira são geradas as outras regras (X → Y). As últimas 3 regras são realizadas com o grupo de dois elementos, onde somente são somados os registros de X e Y quando, no caso de {Masculino, SP} → Renda, os registros possuem "Masculino", "SP" e "Renda" igual a 1 (5 registros) dividido pelos registros "Renda" (8 registros), ou seja, confiança 0,62.

Como foi definido nível de confiança 0,8, teremos somente duas regras que atingirão o mínimo de confiança que serão {Renda} → Masculino e {SP} → Masculino, ou seja, das vendas que já foram realizadas, se o cliente se possui renda maior de R$1000, então é do sexo masculino e se é morador de São Paulo, então seu sexo também é masculino.

Se ao analisar os resultados, for definido que estes são aplicáveis e trazem informações importantes, os contatos para novas vendas poderão ser direcionados para os clientes com renda maior de R$1.000 e do sexo masculino ou ainda homens e residentes em São Paulo. Esse exemplo possui poucos registros, mas serve didaticamente para compreender a aplicação da técnica de regras de associação.

3.2 Algoritmos de regra de classificação

A ferramenta Weka possui dois algoritmos de Classificação da família J48, que são o J48.J48 e o J48.PART. O primeiro destes é uma versão desenvolvida na linguagem Java para o Weka do algoritmo C4.5 release 8 de (QUINLAN, 1993) que é a última versão do algoritmo de geração de árvores antes do C5.0 e geram árvores de decisão c4.5 com ou sem poda.

O C4.5 (QUINLAN, 1993) é um algoritmo melhorado em relação do Iterative Dichotomizer 3 (ID3) (QUINLAN, 1990) que consiste em realizar indução de árvores de cima para baixo recursivamente, para tentar escolher sempre o melhor nó de decisão da árvore, que entre as melhoras inclui o combate aos métodos de *overfliting*[14], podando a árvore.

O algoritmo utiliza um tipo de pós-poda, onde um ramo da árvore é podado e transformado em folha. Esse corte é feito de forma estatística, levando em consideração os erros em um nó e seus descendentes, dessa maneira só haverá a poda se o desempenho de toda a árvore não sofrer grande impacto. Mais duas características podem ser destacadas, a validação cruzada de um ou mais grupos que serve para melhorar a estimativa de erro e a outra de gerar regras de decisão a partir de árvores e compará-las entre si (QUINLAN, 1993).

A validação cruzada é a operação onde os "dados de treinamento são misturados e reamostrados para a classificação com a árvore criada" (TRADUÇÃO NOSSA) (SANTOS, 2005, p. 6). Esta experiência é repetida conforme o número de dobras (*folds*) definidas.

O J48.J48, que é a versão Java do C4.5 no *software* Weka, "constrói um modelo de árvore de decisão baseado em um conjunto de dados de treinamento e usa esse modelo para classificar outras instâncias num conjunto de teste" (TRADUÇÃO NOSSA) (OLIVEIRA et al., 2002, p. 4-5).

Este algoritmo possui como parâmetros informações importantes para definir o grau de qualidade que as árvores serão geradas. Como exemplos dos parâmetros podem ser citados a quantidade mínima de instâncias por cada folha e também um parâmetro para identificar se a árvore poderá ser binária.

O outro algoritmo (J48.PART) constrói regras de produção a partir da árvore de decisão e está é a diferença em relação ao J48.J48. Para a criação dessas regras o algoritmo induz regras inicialmente de uma árvore montada e depois segue refinando estas. Para cada

---

14 Overfliting é quando a taxa de acertos no conjunto de treinamento é alta, mas alcança níveis muito baixos nos teste. (MATINHAGO, 2005)

uma dessas regras “é estimada a cobertura das instancias da base. Isso ocorre repetidamente até que todas as instâncias sejam cobertas” (TRADUÇÃO NOSSA) (OLIVEIRA et al., 2002, p. 5).

3.2.1 Algoritmo ID3 (*Iterative Dichotomizer* 3)

O algoritmo ID3 é um algoritmo indutivo, seu conhecimento preliminar do conjunto de treinamento gera informação, que depois vai ser validada. Esse conhecimento conforme já discutido, é gerado no formato de árvores de decisão. O algoritmo ID3 é apresentado na figura 3.4:

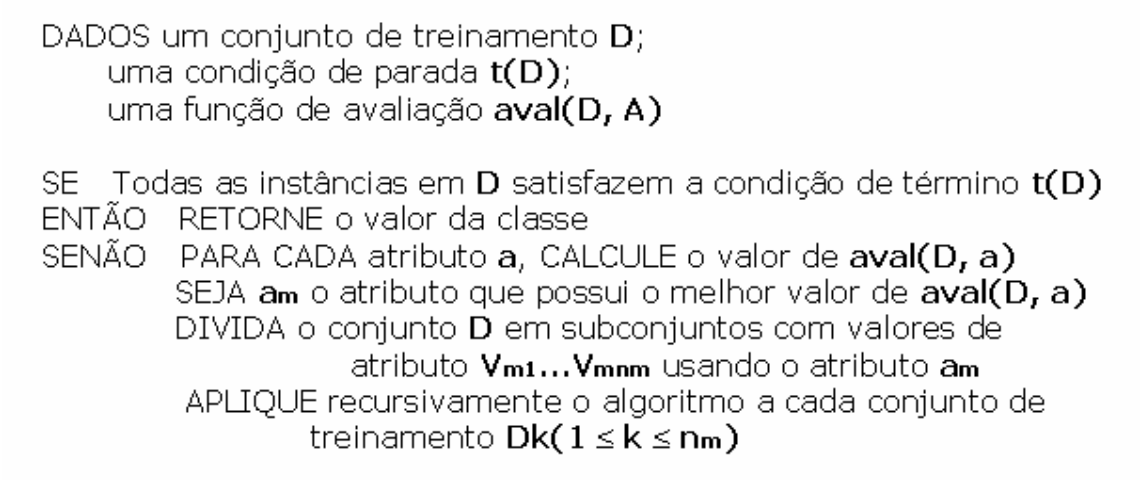

Figura 3.4 – Algoritmo ID3
Fonte: Adaptado de MONGIOVI, 1998

A função de avaliação das regras que foram induzidas é a entropia, quanto menor for este valor mais informativo será o atributo, isso significa também que menor será a árvore gerada. A escolha do melhor nó, já comentada anteriormente é feita através de uma função de avaliação, que utiliza estatística.

No quadro 3.2, pode-se observar um conjunto **S** de exemplos:

Quadro 3.2 – Conjunto S, que representa o conjunto de treinamento

	REGIÃO	MASCULINO	IDADE	GOLD	CLASSE
Ex1	Sul	Sim	> 30	Sim	Venda
Ex2	Sudeste	Sim	> 30	Não	Venda
Ex3	Norte	Não	> 30	Sim	Rejeição
Ex4	Sul	Não	> 30	Não	Venda
Ex5	Sudeste	Não	< 31	Sim	Rejeição
Ex6	Sul	Não	> 30	Não	Rejeição
Ex7	Norte	Sim	< 31	Não	Venda
Ex8	Sudeste	Não	< 31	Sim	Rejeição

E um conjunto de n classes C={C_1, C_2, C_3...C_n), no caso {Venda, Rejeição}, sendo que a probabilidade (p_i) da classe C_i em S, a entropia deste conjunto de exemplos S é representada na figura 3.5:

$$Entropia = -\sum_{i=1}^{c} p_i \log_2 p_i$$

Figura 3.5 – Função da Entropia
Fonte: Adaptado de MONGIOVI, 1998

Ou seja, a função de avaliação de um atributo **a** é a média ponderada dos grupos de exemplos segundo **a**, onde:

c – número de classe no conjunto de treinamento

p_i – é o número de exemplos em que o atributo **a** possui o valor v_i, dividido pelo número de exemplos no conjunto de treinamento com uma classe. Em outras palavras, é a probabilidade de se ter um número de exemplos (> 30 anos e venda) dividido pelo número de vendas.

Executando um exemplo, sendo que a coluna *Gold* identifica os clientes que possuem cartão *Gold* entre a base de treinamento (note que é feito para cada classe, venda e rejeição), teremos:

Entropia(Gold = Sim) = - 1/4.log2(1/4) – 3/4.log2(3/4) = 0,81

Entropia(Gold = Não) = - 3/4.log2(3/4) – 1/4.log2(1/4) = 0,81

Entropia(Gold) = (4/8).0,81 + (4/8).0,81 = 0,81

Atributo Região:

Entropia(Região = Sul) = - 1/3.log2(1/3) – 2/3.log2(2/3) = 0,881

Entropia(Região = Norte) = - 1/2.log2(1/2) – 1/2.log2(1/2) = 1

Entropia(Região = Sudeste) = - 1/3.log2(1/3) – 2/3.log2(2/3) = 0,881

Entropia(Região) = (3/8).0,881 + (2/8).1 + (3/8).0,881 = 0,785

Atributo Idade:

Entropia(Idade = >30) = - 3/5.log2(3/5) – 2/5.log2(2/5) = 0,937

Entropia(Idade = <31) = - 1/3.log2(1/3) – 2/3.log2(2/3) = 0,881

Entropia(Idade) = (5/8).0,937 + (3/8).0,881 = 0,916

Para o atributo Masculino, que foi selecionado como melhor:

Entropia(Masculino = Sim) = - 3/4.log2(3/4) – 0/4.log2(0/4) = 0,33

Entropia(Masculino = Não) = - 1/4.log2(1/4) – 4/4.log2(4/4) = 0,50

Entropia(Masculino) = (3/8).0,33 + (5/8).0,50 = 0,43

Aplicando o algoritmo chega-se até a árvore da figura 3.6:

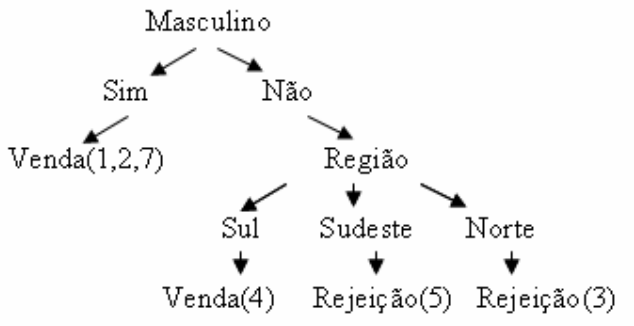

Figura 3.6 – Árvore gerada através do algoritmo ID3
Fonte: FIGURA NOSSA

Para entender a aplicação, vamos seguir os passos. O atributo escolhido “Masculino” será utilizado para criar as classes D_k de D, onde D(sim) = {1, 2, 7} e como todos são “Venda” a execução é interrompida já que a regra diz que todos os masculinos refletem em venda (Masculino = Sim: Venda). O próximo passo do algoritmo é buscar por outro atributo para seguir após “Masculino = Não”. A escolha foi “Região” e o algoritmo recursivamente segue testando e escolhendo os melhores nós com a menor entropia.

Ao utilizar classificação podem ocorrer situações onde alguns registros não podem ser classificados pelo fato de não pertencerem à árvore. Para que fique mais claro, ao observar os registros 4 e 6 da tabela 3.6, pode-se notar que ambos possuem os atributos Região = SUL e Masculino = NÃO, porém, o primeiro é uma venda e o outro é uma rejeição.

Para o exemplo ilustrado a escolha foi venda e o registro 6 passa a fazer parte dos registros que não puderam ser classificados. Esse registro possui os mesmo atributos do registro 4, porém, o próximo atributo escolhido é a região e por este motivo, como o registro não é da região sul como as outras vendas não pôde ser classificado. Ao gerar a árvore o algoritmo chegou até a regra que a região sul identifica que o registro é um venda e não uma rejeição.

Os registros com erro são aqueles em que o registro é classificado, porém suas características não o identificam como tal. É o caso do registro 4 que embora seja venda que não é do sexo masculino, porém faz parte da região das outras vendas (SUL). Dessa maneira ficou classificado, mas classificado como erro.

3.2.2 Algoritmo C4.5

Em 1993, QUINLAN (1993) apresentou uma inovação do Algoritmo ID3, utilizando poda da árvore. Essa poda tem o objetivo de remover ramos da árvore, que é chamado de sobreajustamento, que pode significar que a árvore ficou mais complexa do que deveria ser.

No C4.5 é utilizado a abordagem de “dividir para conquistar” (TRADUÇÃO NOSSA) (MITCHEL, apud MARTINHAGO, 2005, p. 48). Nesse tipo de abordagem o problema original é divido em partes menores semelhantes ao original, recursivamente vão sendo resolvidos e suas soluções formarão uma combinação para o problema inicial (CORMEN et al., 2002). Como passos do algoritmo, conforme MARTINHAGO (2005) devem-se:

- Escolher um atributo;

- Adicionar um ramo para cada atributo;

- Passar os exemplos para as folhas (levando em conta o atributo escolhido);

- Para cada nó folha (se forem da mesma classe) associar a classe ao nó folha, se não, repetir os passos anteriores.

O algoritmo, de forma recursiva, vai dividindo o conjunto de treinamento até que resulte apenas uma classe em cada subconjunto. Ao invés de gerar um ramo que possui mais de uma possibilidade, é feita a poda para diminuir esse ramo em apenas uma folha. Esse é o princípio da poda da árvore, transformar ramos em folhas.

O C4.5 realiza esses cortes baseado em métodos estatísticos com base nos erros do nós e seus descendentes. Para identificar a raiz e seus descendentes são realizados os cálculos da entropia e do ganho de informação. A entropia, já utilizada no ID3, mede o quanto maior é a capacidade de previsão. Na figura 3,7 é apresentado:

-Árvore sem poda gerada pelo algoritmo ID3;

-Árvore com poda gerada pelo algoritmo C4.5, onde o mesmo decidiu por cortar o ramo do atributo “Masculino” com valor igual a não e transformá-lo em uma folha.

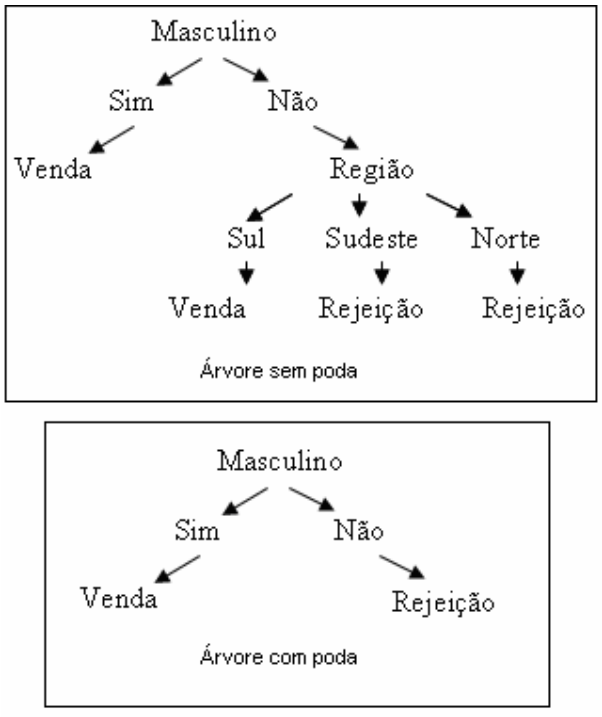

Figura 3.7 – Árvore sem poda e árvore com poda
Fonte: FIGURA NOSSA

O ganho de informação vai medir a redução da entropia nas várias partições dos exemplos de treinamento, de acordo com o valor de um atributo (QUINLAN, 1996) e, dessa maneira, consiga gerar árvores com profundidade e menos nós. A figura 3.8, representa a expressão do ganho de informação.

$$Ganho(S, A) = Entropia(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v)$$

Figura 3.8 – Expressão do Ganho de Informação
Fonte: Adaptado de LÓPEZ; HERRERO, 2004

Nessa expressão, A é o atributo e o subconjunto de S é aquele em que o atributo A tem valor v. Dessa forma, Sv são os casos em cada classe. O ganho de informação mede a eficácia de um atributo nos dados de treinamento, a escolha do atributo que mais reduz a entropia faz com que se gerem árvores com menos nós e ramificações.

O algoritmo C4.5, possui ainda a capacidade de melhorar a estimativa do erro, fazendo a validação cruzada com dois ou mais grupos, chamada de *v-fold*. Além disso, é possível trabalhar com valores contínuos ou indisponíveis (FERNÁNDEZ, 2004). Na figura 3.9, é apresentado o pseudocódigo do algoritmo C4.5.

```
Função C45
            (R: conjunto de atributos não classificadores,
            C: atributo classificador,
            S: conjunto de treinamento) retorna árvore de descisão;
Inicio
            Se S está vazio,
                  retornar um único nó com valor Falha;
            Se todos os registros de S têm o mesmo valor para o atributo classificador,
                  retornar um único no com esse valor;
            Se R está vazio,
                  retornar um único nó com o valor mais freqüente do atributo
                  classificador dos registros de S;
            Se R não está vazio,
                  D = atributo com maior valor de ganho(D,S) dos atributos de R;
                  Sejam {dj | j=1,2,..., m} os valores do atributo D;
                  Sejam {dj | j=1,2,..., m} os subconjuntos de S correspondentes aos
valores de dj respectivamente;
                  Devolver uma árvore com a raiz nomeada como D, com arcos
nomeados (d1, d2,..., dm) que vão respectivamente às árvores
                  C4.5(R-{D}, C, S1), C4.5(R-{D}, C, S2), C4.5(R-{D}, C, Sm);
Fim
```

Figura 3.9 – Pseudocódigo algoritmo C4.5
Fonte: Baseado em FERNÁNDEZ, 2004, p. 9

O algoritmo parte de um conjunto de treinamento (S) e um atributo classificador (C). Para ilustrar o exemplo, vamos imaginar que o atributo (C) escolhido no conjunto de treinamento da Tabela 3.6 seja MASCULINO. Se o conjunto de treinamento estiver nulo, não é possível continuar a execução do algoritmo e nesse caso, será retornada uma falha. Se todos os registros possuírem o atributo MASCULINO = SIM, nesse caso não há o que fazer senão retornar um nó MASCULINO = SIM, que foi o caso da figura 3.6, onde todos os atributos deste tipo eram VENDA.

Se não existem atributos que não sejam o classificador, então deve ser retornado um nó com o valor do classificador que mais aparece nos registros. Caso contrário deve ser aplicada a fórmula e gerado o ganho de informação dos atributos não classificadores, o maior valor é retornado como a raiz da árvore, da mesma maneira da escolha do atributo REGIÃO (figura 3.6) e com os arcos SUL, SUDESTE e NORTE, que são seus subconjuntos.

3.2.3 Algoritmo J48.J48

Para que o usuário possa utilizar e gerar as árvores de decisão no *software* Weka, é necessário que utilize o algoritmo J48. O Weka possui uma implementação da última versão pública da família do C4.5, que no caso é a *release* 8. Após esta, foi lançada a versão comercial do C5.0. Essa classe gera as árvores de C4 com ou sem poda.

Os algoritmos dentro do Weka, são organizados em pacotes e no caso do J48, estão localizados em *weka.classifiers.j48.J48*.

O Weka permite que estes algoritmos sejam executados em linha de comando e dessa forma, para executar o algoritmo J48 deve-se utilizar a seguinte linha de comando:

java weka.classfiers.j48.J48 –t arquivo.arff

Essa linha fará a invocação da *Java Virtual Machine* (JVM) e segundo WEKA (2007), através do parâmetro –t é indicado qual conjunto de treinamento será utilizado. A classe J48 na verdade, não possui as rotinas para a geração de árvores, mas inclui as instâncias de outras classes. O pacote j48 é o que possui as classes que executam o J4.8. Na figura 3.10, pode-se observar os parâmetros do algoritmo no Weka:

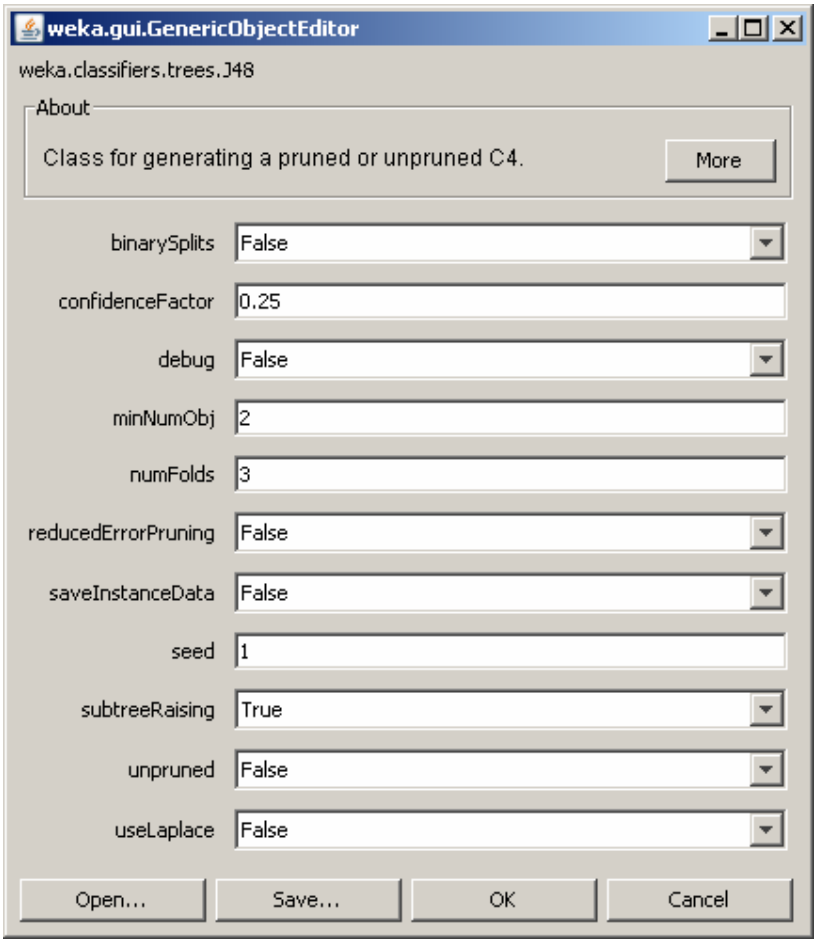

Figura 3.10 – Tela de configuração do J48 no Weka
Fonte: WEKA, 2007

Os parâmetros da figura 3.10 representam:

binarySplits – informa ao algoritmo se é possível a geração de árvores binárias;

confidenceFactor – no caso da utilização de poda, quanto menor o valor de confiança, maior será a poda;

debug – essa opção serve exclusivamente para gerar informações sobre a execução do algoritmo que não são mostradas enquanto estiver marcada como falso;

numMinObj – representa o número mínimo de instâncias por cada folha;

numFolds – está diretamente ligado com a redução de erros na poda. Uma dobra (ou *Fold*) é utilizada para a poda e o restante serve para crescer a árvore (WEKA, 2007). Cada dobra representa a execução do algoritmo. A cada execução o resultado vai sendo refinado, testado e comprovada sua qualidade em qualquer árvore;

reducedErrorPruning – ativa ou não a redução de erros na poda. É realizada a poda através de um conjunto de validação;

saveInstanceData – salva os dados de treinamento para visualização;

seed – esse parâmetro é utilizado, quando ativo, para a redução de erros na geração das árvores, testando-a através de dados gerados aleatoriamente. Cada semente utilizada realiza a mistura dos dados para serem processados. Como a mistura é aleatória, árvores diferentes são testadas, aprimorando e reduzindo erros de classificação;

subTreeRaising – utilizar uma operação de substituir um nó interno da árvore por um dos nós que estão abaixo (a escolha do nó abaixo é feita através da taxa de erros) e logo após classificar novamente a árvores. Esta é uma ação utilizada na poda;

unpruned – não utilizar poda;

useLaplace – conta o número de folhas excluídas baseadas na transformada de Laplace (WEKA, 2007), "que é um método simples para transformar um Problema com Valores Iniciais (PVI), em uma equação algébrica, de modo a obter uma solução deste PVI de uma forma indireta, sem o cálculo de integrais e derivadas para obter a solução geral da Equação Diferencial" (TRADUÇÃO NOSSA) (SODRÉ, 2003, p. 1).

Além dessas configurações, é possível, através do modo gráfico, marcar e escolher um conjunto de teste. Também se pode ativar a validação cruzada e o percentual de separação, que determina o tamanho do conjunto de registros que serão separados para a geração dos sucessores e criação de novos classificadores (WEKA, 2007), conforme figura 3.11:

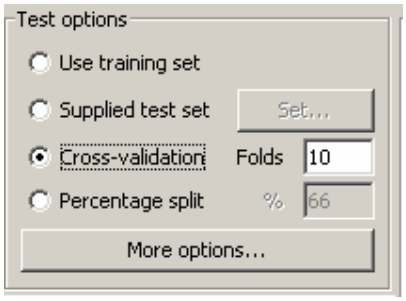

Figura 3.11 – Tela de opções de teste no Weka
Fonte: WEKA, 2007

O Weka permite, para a classificação, observar os erros, visualizar a árvore, entre outras opções utilizando o clique do botão direito na lista de resultados, conforme figura 3.12:

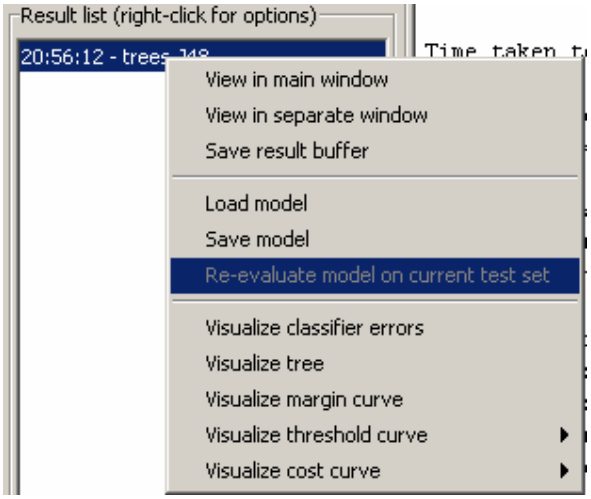

Figura 3.12 – Outras opções de visualização de classificação no Weka
Fonte: WEKA, 2007

Na opção de visualização de árvores, a ferramenta gera a árvore para ser analisada, conforme figura 3.13:

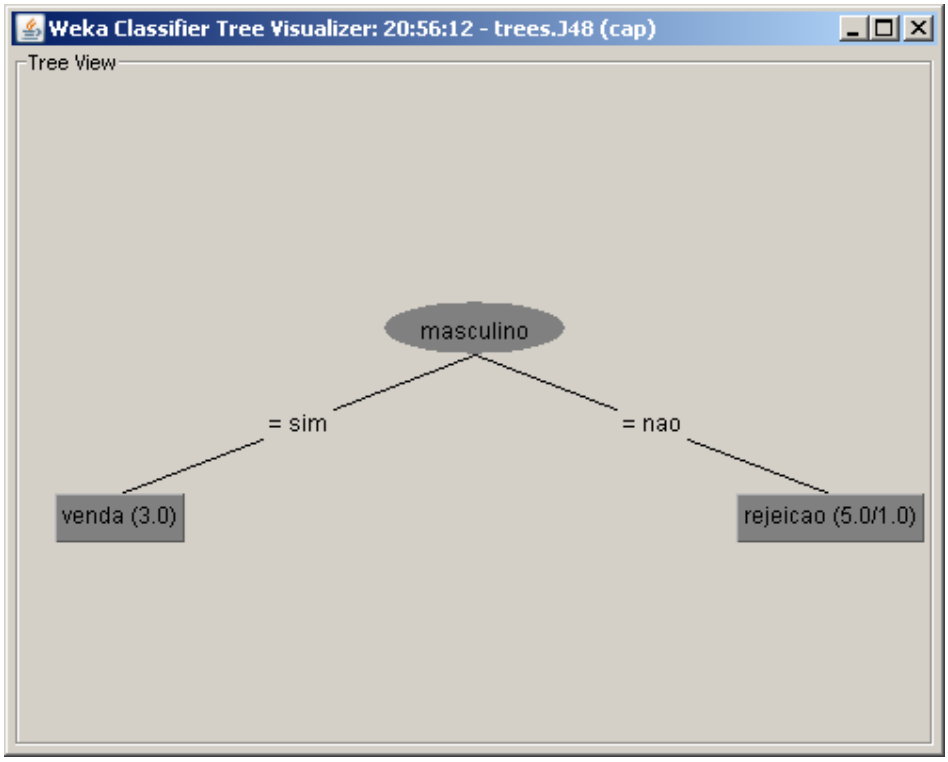

Figura 3.13 – Visualização de árvore
Fonte: WEKA, 2007

Na figura 3.14, pode-se observar o gráfico dos erros de classificação. Nesse exemplo houve um erro que está identificado por um quadrado no gráfico.

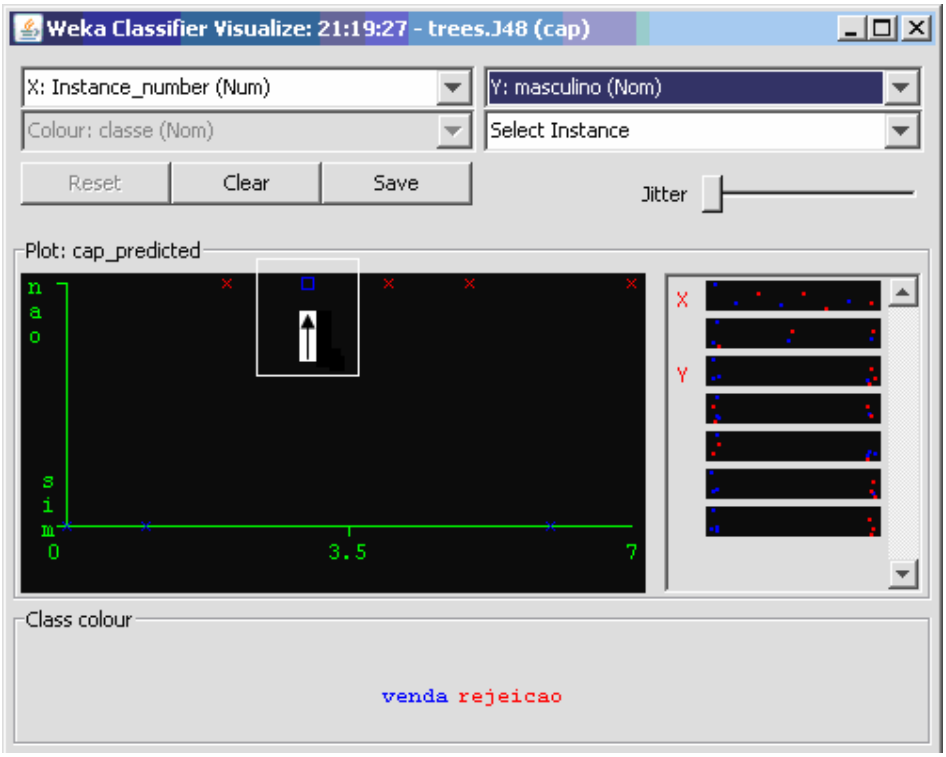

Figura 3.14 – Visualização de erros de classificação
Fonte: WEKA, 2007

Nessa situação, esse registro embora seja de “rejeição” está dentro das características que identificam uma venda. Nesse exemplo, o algoritmo, utilizando poda chegou à conclusão que sempre que MASCULINO = ‘SIM’ então VENDA e no gráfico pode-se visualizar em destaque um caso de VENDA onde MASCULINO = ‘NÃO’, por este motivo o Weka destacou no gráfico essa informação para que o usuário analise.

3.2.4 Algoritmo J48.PART

Conforme já foi introduzido, o algoritmo PART constrói regras de produção a partir da árvore de decisão. Para a geração da lista de decisão, o algoritmo parte de uma árvore já montada e realiza então a indução de regras, que após vão sendo comprovadas ou alteradas. Este algoritmo também atua segundo a abordagem “dividir para conquistar”. Segundo WEKA (2007), a cada iteração é criada uma árvore de forma parcial e transformando a melhor folha (maior ganho de informação) em uma regra.

Para utilizar o algoritmo via linha de comando, J48 deve-se utilizar a seguinte linha:

java weka.classfiers.j48.PART arquivo.arff

Até que todas as instâncias (registros de base de treinamento) possuam a estimativa de quanto representam em relação aos outros registros da base, o PART executa repetidamente a fim de refinar e selecionar as regras com cobertura de folhas mais alta, sempre em relação à quantidade de registros da base. Para ilustrar a execução do PART, será utilizado o mesmo exemplo Tabela 3.6.

Na execução do algoritmo J48, já tratado anteriormente, pode-se observar na figura 3.13 que o J48 gerou uma árvore com a raiz "MASCULINO" e que se este for 'Sim' é uma venda e se for 'Não' caracteriza uma rejeição.

Na figura 3.16, pode-se observar a lista de regras de decisão geradas a partir do exemplo. A lista de decisão do algoritmo PART gerou duas regras:

```
masculino = nao: rejeicao (5.0/1.0)

: venda (3.0)

Number of Rules  :      2
```

Figura 3.16 – Lista de regras geradas pelo J48.PART
Fonte: WEKA, 2007

Nesse exemplo se chegou até duas regras que compõem a lista de decisão. A primeira regra informa que quando o SEXO for 'Masculino' esse cliente será uma rejeição, pode-se ainda observar que dos 5 registros classificados como rejeição, apenas 1 deles não era do sexo masculino. A segunda regra informa ao usuário que em qualquer outro caso seria venda, ou seja, o resultado diz que o atributo sexo é fundamental para classificar esse conjunto de registros.

A característica do PART, que é gerar regras a partir de árvores e transformar o melhor nó em uma regra, agora pode ser observada. Nesse exemplo, o algoritmo J48.PART partiu da árvore gerada como no J48 (até então, ambos os algoritmos são iguais) e selecionou a folha VENDA e a transformou em uma regra, em que todo o registro que não se enquadrar na primeira regra deve ser classificado como venda.

Da mesma maneira do J48.J48, este algoritmo pode ser utilizado através da tela *Weka Explorer*, porém o PART possui menos parâmetros que podem ser vistos na figura 3.17. A utilização dos parâmetros é a mesma do outro algoritmo.

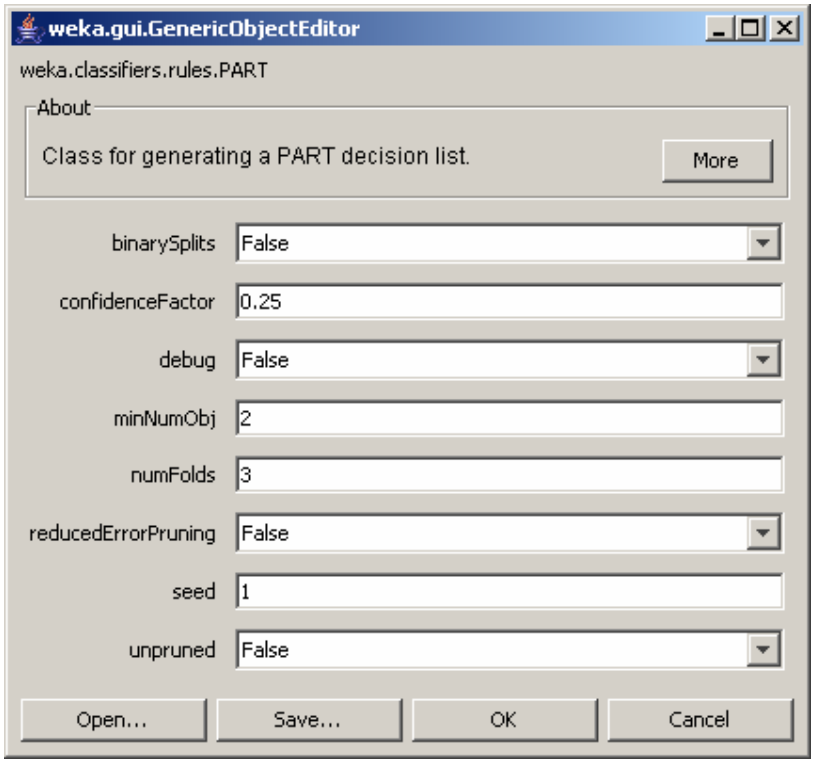

Figura 3.17 – Parâmetros do J48.PART
Fonte: WEKA, 2007

Juntamente com o J48 e J48.PART podem ser utilizados métodos para aumentar desempenho, como os métodos *Bagging* e *Boosting*.

3.2.5 Método *Bagging* e *Boosting*

O Weka possui métodos de meta aprendizagem. Estes podem ser utilizados na tarefa de construir conjuntos de classificadores (OLIVEIRA et al., 2002). Todos os métodos de meta aprendizagem podem ser selecionados no Weka através do pacote "*classifiers*" no item "meta", disponíveis na ferramenta junto com os algoritmos de classificação. Essas classes têm a função de aumentar o desempenho e capacidade da geração de regras, por este motivo são incorporados a outros algoritmos de aprendizagem (WEKA, 2007). Para este estudo, foram destacados apenas dois métodos: *Bagging* e *Boosting*, utilizados também em outros trabalhos como OLIVEIRA et al. (2002), por exemplo.

Bagging é o nome dado a um procedimento que constrói classificadores partindo de conjuntos de amostras que são processadas de forma independente e sucessiva, onde se pode definir o número de interações, sendo que o padrão são 10 *Bagging interations*. Em outras palavras, o *Bagging* se diferencia da utilização d o número de sementes do J48 pelo fato que o

conjunto de reamostragem é gerado com número de sementes definidas de maneira aleatória, onde as ocorrências de repetição são substituídas (OLIVEIRA et al., 2002) (WEKA, 2007).

Para utilizar *Bagging* no algoritmo J48, por exemplo, pode-se utilizar o Weka selecionando o *Bagging* dentre os métodos "meta" ou então através de linha de comando no sistema operacional, da seguinte maneira:

java weka.classifiers.bagging –W jaws.classifiers.j48.J48...-- -U

Dessa maneira, todas as opções do Bagging e J48 podem ser utilizadas, sendo separadas por "--", "-W" do Bagging e "–U" do J48, sem que exista conflito entre as opções de cada um.

No método *Boosting*, cada instância gerada recebe um peso, sendo que inicialmente todas possuem o mesmo. Ao ser feita a primeira indução, os pesos daquelas que foram definidas como erros são alterados. Essa comparação e definição de erros são realizadas com base nos classificadores que já foram construídos (OLIVEIRA et al., 2002). Dessa maneira, aqueles conjuntos de amostras, que acabam sendo identificados com erros, são separados e como resultado se obtém os conjuntos de amostras que conseguiram ter o melhor aproveitamento nas interações.

3.3 Algoritmos de *Clustering*

O Weka possui algoritmos de *Clustering* para procurar grupos de instâncias que sejam similares dentro da base de dados, embora tenha seu foco direcionado à classificação. Dos geradores de *cluster*, presentes na ferramenta podem ser citados: k-Means, EM, Cobweb, X-means, FarthestFirst.

Para está técnica, pode-se destacar o algoritmo SimpleKMeans que assim como os outros algoritmos presentes no Weka, pode ser aplicado sobre um conjunto de treinamento, que serve para que seja avaliado como o algoritmo se comporta nessa amostra da base de dados. Os algoritmos de *Clustering* funcionam "modelando a distribuição de instâncias probabilisticamente" (TRADUÇÃO NOSSA) (WEKA, 2007, p. 34) e é por este motivo que pode ser interessante verificar os resultados na base de teste. O SimpleKMeans executa 10 vezes (opção definida por padrão e que pode ser configurada) o algoritmo K-Means para trazer um resultado que pode ser considerado estatisticamente bom para o usuário (PINHEIRO, 2006). Na figura 3.18, é apresentado o K-Means:

```
1.  Eleger k exemplos que atuam como sementes (k número de clusters);
2.  Para cada exemplo, acrescentar exemplo à classe mais similar;
3.  Calcular o centroide de cada  classe, que passam a ser as novas sementes;
4.  Se não se chega a um critério de convergência (por exemplo, duas
interações não mudam as classificações dos exemplos), voltar a 2.
```

Figura 3.18 – Pseudocódigo algoritmo K-Means
Fonte: Adaptado de LÓPEZ; HERRERO, 2004, p.45

O pacote que contém o algoritmo é *weka.clusterers* e a ferramenta, para os algoritmos que fazem a modelagem de instâncias conforme citado, mostra como as muitas instâncias são atribuídas a cada um dos *clusters*, conforme WEKA (2007). Utilizando os mesmo dados do quadro 3.2 e aplicando o *clustering* com percentual de divisão de 66%, o que significa que esse percentual da base de dados é utilizado para treinamento e o restante para teste. Como opções, número de clusters igual a 2 e 10 *seeds*.

O resultado desse teste é apresentado na figura 3.19:

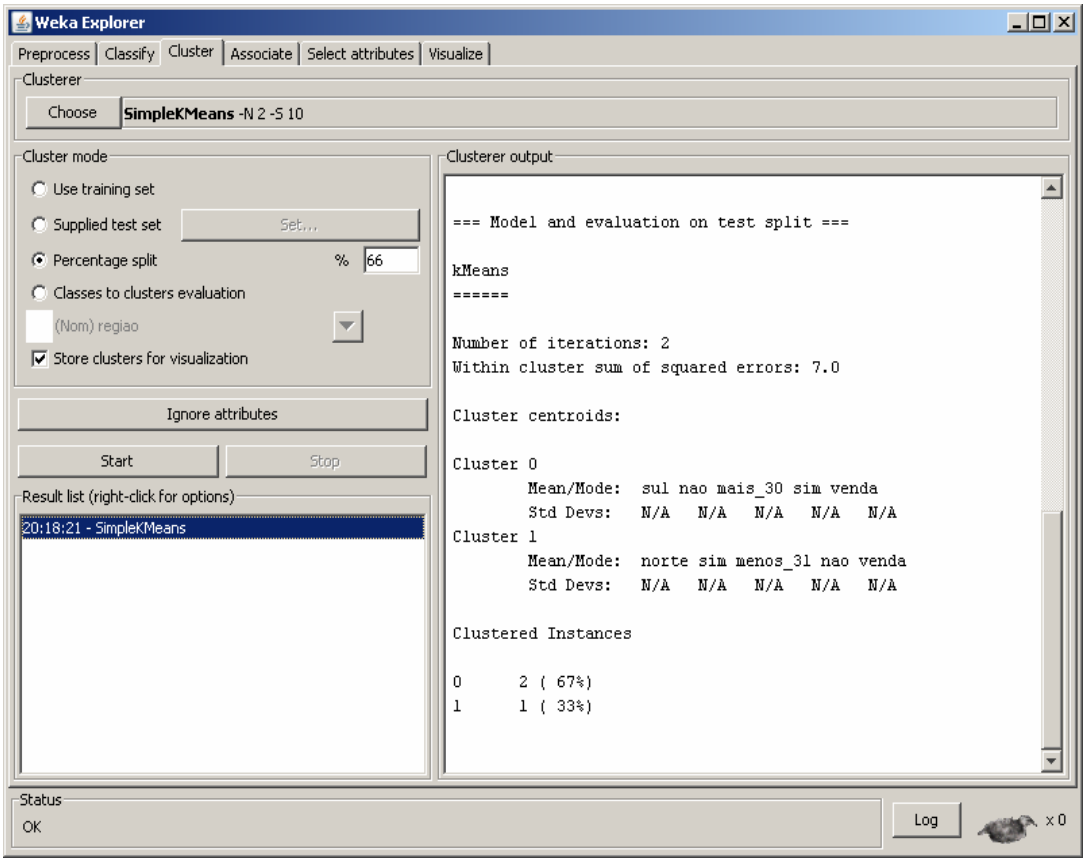

Figura 3.19 – Saída algoritmo SimplekMeans, para o conjunto testado
Fonte: WEKA, 2007

Nessa figura 3.19, pode-se observar que o algoritmo realizou duas interações, onde a soma dos erros enquadrados totalizou 7.0. São 8 instâncias com 5 atributos:

- REGIAO {sul, sudeste, norte};

- MASCULINO {sim, nao};

- IDADE {mais_30, menos_31};

- GOLD {sim, nao};

- CLASSE {venda, rejeicao};

Dos oito registros, foram encontrados dois *clusters*:

-{sul, não, mais_30, sim, venda}: 2 registros;

-{norte, sim, menos_31, não, venda}: 1 registro;

A partir desse resultado, o usuário pode avaliar o quanto bem o modelo representa os dados. Em um modelo maior, e com a configuração do número de *clusters*, podem-se encontrar as diferentes categorias presentes na base de dados.

4 ESTUDO DE CASO

Um estudo de caso é uma técnica utilizada para realizar algum tipo de estudo, envolvendo uma pesquisa para conhecer detalhes sobre algo real. A definição para SIMON, apud FELBER (2005, p. 57) especifica que é uma técnica "[...] onde se faz uma pesquisa sobre um caso particular, para tirar conclusões sobre princípios gerais daquele caso específico".

Busca-se estudar o fenômeno da utilização de KDD na área de *telemarketing* e conhecer todo o contexto. É utilizado porque não fica evidente e ainda não se conhece a relação entre contexto (*Call Center*) e fenômeno (KDD). Para este tipo de situação que se torna possível utilizar um estudo de caso (YIN, apud FELBER, 2005).

Para utilizar técnicas de descoberta de conhecimento em um *Call Center* ativo, serão utilizados dados reais de uma empresa deste ramo, que atualmente não possui nenhuma ferramenta de garimpagem. Como os dados são confidenciais, estes não deverão conter nomes de clientes, documentos ou qualquer outro dado pessoal. No momento de realizar a validação e comparação de resultados, a fim de proteger a empresa, o *Call Center* não será identificado. Entretanto, os testes serão realizados de maneira que poderão ser aplicados a qualquer outra empresa de *Call Center*, desde que funcionem de modo similar.

Foi escolhida para este estudo, a venda de títulos de capitalização, e tal escolha se deve ao fato de existir registros de atividades em um período maior de um ano na comercialização deste produto e uma grande quantidade de atributos de cliente, presentes nas bases que são enviadas para a empresa de *telemarketing*. Essa venda de títulos é realizada para clientes que possuem cartão de crédito da empresa contratante do serviço.

4.1 Modelagem dos dados

Na lista de *prospects* enviada pela empresa contratante pode-se, antecipadamente, destacar alguns atributos que serão utilizados no processo da descoberta de conhecimento. Na

figura 4.1, é apresentada a tabela do banco de dados onde são armazenados os registros recebidos:

```
Table: MAILING
Column Name                    Data Type
------------------------------ ------------------
NUM_MAILING                    NUMBER (6)
SEQ_MAILING                    NUMBER (6)
NOME_CLIENTE                   VARCHAR2 (50)
CPF                            VARCHAR2 (16)
DDD1                           NUMBER (4)
FONE1                          NUMBER (8)
CLASSIFICACAO                  VARCHAR2 (20)
DATA_NASCIMENTO                DATE (7)
DATA_ABERT_CONTA               DATE (7)
DIA_VENC_FATURA                NUMBER (2)
VALOR_LIMITE                   NUMBER (9,2)
COD_REGIAO_MAILING             NUMBER (2)
COD_CLASSIFICACAO              NUMBER (3)
COD_FAIXA_ETARIA               NUMBER (3)
COD_TEMPO_CONTA                NUMBER (3)
COD_VALOR_LIMITE               NUMBER (3)
```

Figura 4.1 – Resultado do comando DESC de uma tabela de *Mailing*
Fonte: FIGURA NOSSA

Os atributos pré-selecionados são: região onde o cliente mora, classificação para o contratante (cliente Gold, Internacional, etc), seu valor de limite no cartão de crédito, tempo de fidelidade ao contratante e faixa etária.

Para se obter uma melhor organização das informações, serão utilizadas tabelas para cadastro de cada uma das características selecionadas. Dessa maneira pode-se ter controle sobre cada um dos atributos e todos os possíveis valores. Por exemplo, deverá existir uma tabela chamada FAIXAS_ETARIAS para que se cadastrem as faixas de idade que serão utilizadas. Os campos da figura 4.1 iniciados por "COD_", identificam que essa coluna é uma chave estrangeira para outra tabela.

Como os dados são tratados de maneira diferente pelo cliente, na tabela de *mailing*, deverão possuir colunas para os dados originais e colunas para realizar a inclusão dos códigos, segundo as tabelas. Por exemplo, o contratante envia apenas a data de nascimento de cada *prospect* e por isso deve ser feito o cálculo e definida sua faixa etária. Na figura 4.2, é apresentado o Modelo Entidade Relacionamento (MER) para os dados utilizados durante a aplicação de KDD:

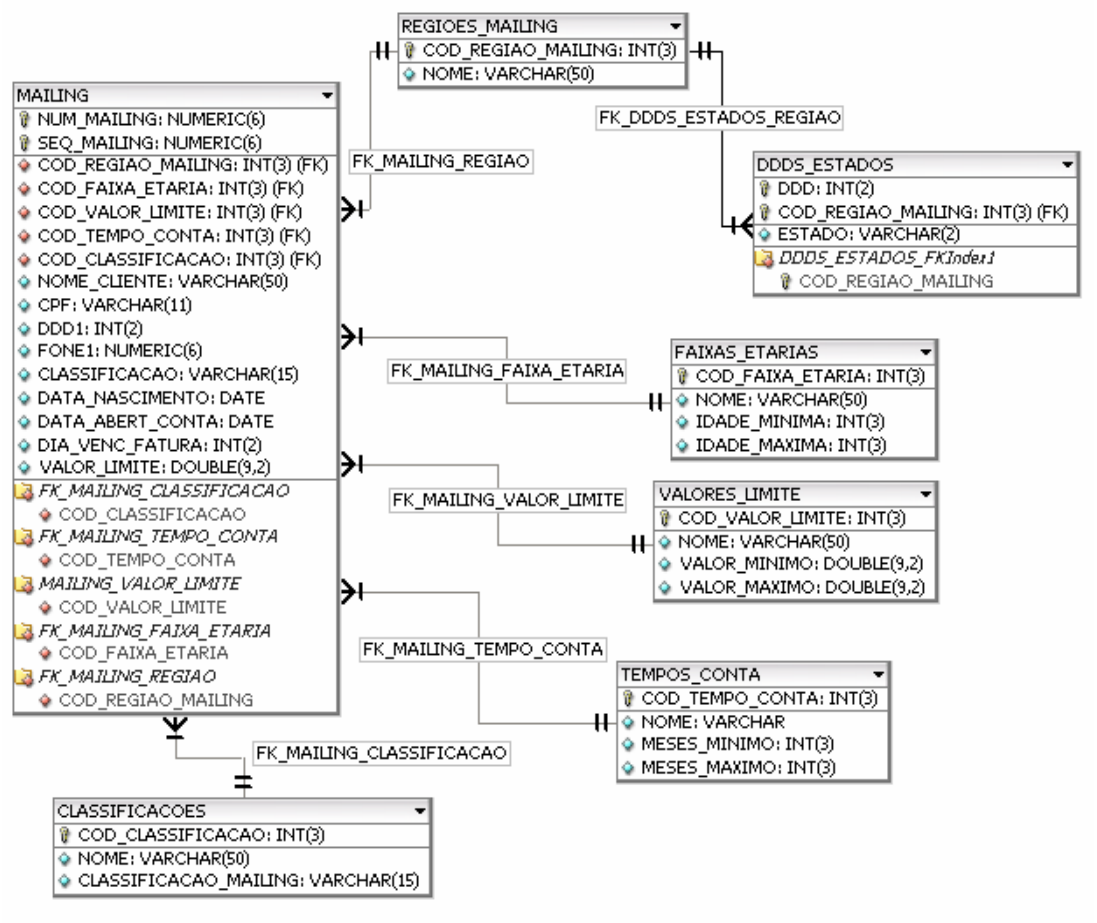

Figura 4.2 – Modelo ER dos dados de vendas de títulos de capitalização
Fonte: Gerado com DBDESIGNER 4 (DBDESIGNER, 2007)

Na figura 4.2, é apresentada a estrutura de tabelas no banco de dados utilizado para armazenar os nomes das listas de *mailing* e seus atributos. O objeto MAILING possui os dados recebidos pelo *Call Center*. Os objetos CLASSIFICACOES, TEMPOS_CONTA, VALORES_LIMITE, FAIXAS ETARIAS e REGIOES_MAILING possuem os cadastros dos tipos de cada atributo que será utilizado. Por exemplo, na tabela FAIXAS ETARIAS, existirão todas as faixas etárias desejadas e ainda uma faixa para os casos onde não seja informada. Na figura 4.3 pode-se observar os dados da tabela de faixas etárias:

COD_FAIXA_ETARIA	NOME	IDADE_MINIMA	IDADE_MAXIMA
1	NÃO INFORMADO	0	0
2	18 A 29 ANOS	1	29
3	30 A 39 ANOS	30	39
4	40 A 49 ANOS	40	49
5	50 A 59 ANOS	50	59
6	ACIMA DE 60	60	999

Figura 4.3 – Registros da tabela de faixas etárias
Fonte: FIGURA NOSSA

Por questões de desempenho de consultas ao banco de dados e relacionamento de 1 para N entre as tabelas é armazenada uma chave estrangeira na tabela de *mailing* com a respectiva faixa etária do momento da inclusão de um novo cliente. Cada faixa, tempo de conta, classificação e região poderá estar incluído em 1 ou mais clientes e o respectivo campo na tabela *mailing* não pode ser nulo.

A mesma lógica é utilizada para os outros objetos, com diferença para o tempo de conta que utiliza a data de abertura de conta, valor limite que é definido através do valor de limite, da classificação que é através do campo de texto recebido com a descrição da classificação do cliente e a região que é definia através do DDD do telefone do cliente com o auxílio da tabelas de DDD por cada estado. Ainda é possível utilizar o dia de vencimento da fatura, nesse caso com valores que são entre 1 e 30.

Em um primeiro momento, para a aplicação de KDD na venda de títulos de capitalização, será utilizado este modelo e ao longo do trabalho serão realizadas as alterações necessárias.

4.2 Técnicas de KDD a serem utilizadas

As técnicas de KDD que serão utilizadas, em um primeiro momento, serão classificações e, se houver tempo disponível, regras de associação. A classificação foi escolhida para gerar árvores de decisão, a fim de encontrar os atributos dos clientes que os fazem adquirirem os títulos ou então recusarem a proposta, montando, desta forma, perfis de compra. Além de encontrar as características determinantes para a compra, pode-se evitar que sejam contatados clientes com as características que levam à rejeição do produto.

Através das regras de associação, podem ser encontradas regras que associem as características dos clientes que adquirem os títulos, ou daqueles que não os adquirem, servindo, portanto, como conhecimento complementar ao gerado pelas regras da classificação. O algoritmo Apriori será o algoritmo utilizado no momento da geração de regras de associação.

Para encontrar os melhores resultados de classificação, serão feitos comparativos entre os resultados do J48.J48 e J48.PART. Os métodos de *Bagging* e *Boosting* serão utilizados para aprimorar e refinar as regras, assim como utilizado em (OLIVEIRA et al.,

2002). Serão realizados diversos testes com os algoritmos e alternando as opções dos mesmos para encontrar a árvore de decisão que apresente os melhores resultados.

As regras e informações produzidas serão repassadas ao setor responsável pelo gerenciamento das bases de clientes e será realizado o acompanhamento dos resultados durante a aplicação dos testes. Dessa maneira, esses responsáveis poderão verificar o conhecimento gerado e testá-lo diretamente na produção da empresa.

4.3 Validação

A validação dos resultados será realizada de duas maneiras, utilizando os registros anteriores à data do conjunto de treinamento, para testar a regras em um conjunto maior de dados e como outra forma de validação aplicar as regras geradas diretamente na empresa, para selecionar os *prospects* que serão disponibilizados para contato para equipe de *telemarketing*.

Pelo fato de existir uma grande base de dados das transações que já foram realizadas, é possível que seja feita uma verificação das regras, conferindo se os registros que não faziam parte do conjunto de treinamento também podem ser classificados dentro das mesmas regras. O conjunto de treinamento pode ser aumentado para refinar o resultado.

O setor de análise de informações da empresa onde será utilizado KDD pode dar suporte e informações referentes aos perfis de clientes que hoje em dia recebem um foco maior. Na figura de especialistas da área de *telemarketing* podem analisar as regras geradas e definir se alguma informação não é aplicável.

Esse mesmo setor possui todas as ferramentas para direcionar para os atendentes os clientes com o perfil definido na regras. No momento que os registros das listas de *mailing* entrarem em produção, já se pode começar a colher informações do desempenho que as regras estão tendo.

Para finalizar a validação será realizada uma comparação do desempenho da produção com KDD e no método utilizado hoje. Para tornar a validação mais eficiente, será realizada uma comparação do número de contatos necessários para se chegar a um total determinado de vendas, primeiramente analisando a forma utilizada atualmente para distribuir os nomes, e depois a quantidade de contatos necessários para o mesmo número de venda utilizando KDD. Para obter esses números ao utilizar mineração, somam-se todos os contatos menos àqueles que não fazem parte do perfil gerado pelo processo de garimpagem.

4.4 *Software* para utilização das técnicas de KDD

Para facilitar a utilização do Weka, será criado um *software* que fará uma interface entre o usuário, base de dados e o Weka. O *software* terá a função de possibilitar que o usuário possa selecionar os dados e os atributos que serão utilizados na mineração. Dessa maneira será transparente a criação do arquivo (.arff) utilizado pelo Weka.

A conexão com o banco de dados poderá ser configurada e mais de um banco de dados poderá ser utilizado em cada garimpagem.

O Weka que possui os algoritmos das técnicas de KDD será a ferramenta que gerará as árvores de classificação que serão utilizadas e testadas na empresa de *Call Center*.

Esses resultados serão apresentados ao usuário, que poderá salvá-los em formato de texto. Caberá ao usuário interpretar as informações descobertas e geradas pela técnica utilizada.

As definições de plataforma, linguagem de programação utilizada e metodologia serão tratadas no trabalho de conclusão II, podendo ocorrer alterações no que foi definido até o momento para o programa responsável por fazer a comunicação entre o usuário e a ferramenta de aprendizagem.

CONCLUSÃO

Nota-se que o fato do *Call Center* não possuir nenhuma técnica de mineração de dados pode significar um mau aproveitamento dos dados disponíveis. Existem informações importantes que não estão disponíveis na visualização dos dados, que dizem respeito à relação entre as características dos clientes e o resultado do contato. São vários os fatores que influenciam as vendas. A combinação de mais de uma característica do cliente pode ser fundamental para definir o perfil da pessoa que mais adquire produtos.

A venda de títulos de capitalização foi selecionada para aplicação das técnicas de KDD e pela quantidade de atributos presentes em cada registro de cliente a ser contatado, cresce a possibilidade de aumentar o grau de qualidade e diversidade das informações geradas.

Das técnicas de KDD, a classificação foi selecionada como a principal técnica a ser utiliza no problema. As árvores de classificação tornarão possível organizar os atributos e a relação entre estes para os contatos com venda ou recusa. O *software* Weka possui todas as funcionalidades necessárias para aplicar KDD e dará todo o suporte para que os dados sejam garimpados.

Além de conhecer os perfis dos compradores em potencial dos produtos, será possível unir o conhecimento adquirido pelos analistas de informações da empresa de *telemarketing* e o conhecimento descoberto com a mineração. Desse modo, a utilização de técnicas de KDD em um *Call Center* auxiliará a seleção mais eficiente dos *prospects*, aumentando as vendas e reduzindo a quantidade de contatos telefônicos com os clientes.

REFERÊNCIAS BIBLIOGRÁFICAS

AGRAWAL, Rakesh; IMIELINSKI, Thomas; SWAMI, Arun. Mining Association Rules between Sets of Items in Large Databases. In: ACM SIGMOD International Conference on Management of Data 5/93, 1993, Washington/USA. **Proceedings of SIGMOD 5/93**. Washington/USA, 1993. p. 207-216.

AGRAWAL, Rakesh; SRIKANT, Ramakrishnan. Fast Algorithms for Mining Association Rules. In: 20th International Conference on Very Large Databases (VLDB) CONFERENCE, 1994, Santiago/Chile. **Proceedings of the 20th International Conference on Large Databases**. Santiago/Chile, 1994. p. 487-498.

ATENDEBEM. São Leopoldo - RS. Atendebem Soluções de Atendimento, 2007. Website da empresa de *Call Center*. Disponível em: <http://www.atendebem.com.br>. Acesso em: 03 jun. 2007.

BRACHMAN, Ronald J.; ANAND, Tej. The Process of Knowledge Discovery in Databases. In: FAYYAD et al. **Advances in knowledge discovery and data mining**. Cambridge-Mass:AAAI/MIT Press, 1996. p. 37-57.

CABENA, Peter et al. **Discovering Data Mining from Concept to Implementation**. New Jersey-USA: Prentice Hall PTR, 1997. 193 p.

CARVALHO, Juliano V.; SAMPAIO Marcus C.; MONGIOVI, Giuseppe. Utilização de Técnicas de "Data Mining" para o Reconhecimento de Caracteres Manuscritos. In: XIV SIMPÓSIO BRASILEIRO DE BANCO DE DADOS, 1999, Florianópolis. **XIV Simpósio Brasileiro de Banco de Dados - Anais.** Florianópolis, 1999. p. 235-249.

CORMEN, Thomas H. et al. **Algoritmos:** teoria e prática. 2. ed. Rio de Janeiro: Campus, 2002. p. 21-25

DANTAS, Edmundo Brandão. **Telemarketing:** a chamada para o futuro. 2. ed. São Paulo: Atlas, 1994. 206 p.

DBDESIGNER. FabForce.net, 2007. Apresenta todas as características do projeto, documentação e o software DBDesigner. Disponível em: < http://fabforce.net/dbdesigner4/>. Acesso em: 18 mai. 2007.

DWBRASIL. DW Brasil, 2007. Website sobre banco de dados e Data Warehouse. Disponível em: <http://www.dwbrasil.com.br/html/dmining.html>. Acesso em: 28 mai. 2007.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery: An overview. In: FAYYAD et al. **Advances in knowledge discovery and data mining**. G. Cambridge-Mass:AAAI/MIT Press, 1996. p. 1-27.

FELBER, Edmilson J. W. **Proposta de Uma Ferramenta OLAP em um Data Mart Comercial:** Uma Aplicação Prática na Indústria Calçadista. Novo Hamburgo: 2005. 102 p.

Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Instituto de Ciências Exatas e Tecnológicas, Feevale, 2005.

FRAWLEY, William J.; PIATETSKY-SHAPIRO, Gregory; MATHEUS, Christopher J. Knowledge Discovery in Databases: An overview. In: AI Magazine, 1992, Menlo Park. **American Association for Artificial Intelligence**. Menlo Park, CA, USA, 1992. p. 57-70.

FREITAS, Alex A. Data Mining, In: XIII Simpósio Brasileiro de Banco de Dados. Maringá/Brasil, 1998. 104 p.

HOAGLIN, David C.; MOSTELLER, Frederick; TUKEY, John W. **Análise exploratória de dados**: técnicas robustas: um guia. Lisboa: John Wiley, 1992. 446 p.

LÓPEZ, José M. M.; HERRERO, Jesús G. **Aplicaciones Prácticas Utilizando Microsoft Excel y Weka** . Apostila Técnicas de Análises de Dados. S.l., 2004. 160 p. Disponível em: <http://galahad.plg.inf.uc3m.es/~docweb/ad/transparencias/apuntesAnalisisDatos.pdf>. Acessado em: 02 jun. 2007.

MARTINHAGO, Sergio. **Descoberta de Conhecimento sobre o processo seletivo da UFPR**. Curitiba: 2005. 114 p. Dissertação (Mestrado em Ciências) – Departamento de Matemática, UFPR, 2005.

MONGIOVI, Giuseppe. **T.E.I. Data Mining**. Notas de Aula Data Mining. Campina Grande, 1998. p. 1-102.

OLIVEIRA Fernando L. et al. Utilização de Algoritmos Simbólicos para a Identificação do Número de Caroços do Fruto Pequi, In: IV Encontro de Estudantes de Informática do Estado do Tocantins**,** 2002, Palmas. **Encontro de Estudantes de Informática do Tocantins – Encoinfo**. Palmas, 2002. p. 34-43.

PINHEIRO, Luciane C. **Método de Representação Espacial de Clustering**. Curitiba: 2006. 123 p. Dissertação de Mestrado – Departamento de Informática, UFPR, 2006.

PRODANOV, Cleber C. **Manual de Metodologia Científica**. 3. ed. Novo Hamburgo: Editora Feevale, 2006. 77 p.

QUINLAN, Ross. Induction of Decision Trees. In: SHAVLIK, Jude (ed.); DIETTERICH, Thomas (ed.). **Readings in Machine Learning**, San Mateo, CA: Morgan Kaufmann Publishers, 1990. p. 81-106.

QUINLAN, Ross. **C4.5:** Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers, 1993. p. 1-109

QUINLAN, Ross. Improved use of continuous attributes in C4.5. In: Journal Of Artificial Intelligence Research 4, 1996, p. 77-90. Disponível em: <http://www.jair.org/media/279/live-279-1538-jair.pdf>. Acessado em: 02 jun. 2007.

SANTOS, Rafael. **Weka na Munheca:** Um guia para uso do Weka em scripts e integração com aplicações em Java. Apostila Princípios e Aplicações de Mineração de Dados. S.l., 2005. 20 p. Disponível em: <http://www.lac.inpe.br/~rafael.santos/CAP/cap359/2005/weka.pdf>. Acessado em: 02 jun. 2007.

Savasere, Ashok; Omiecinsky, Edward; Navathe, Shamkant. An Efficient Algorithm for Mining Association Rules in Large Databases. In: 21st International Conference on Very Large Databases (VLDB) Conference, 1995, Zurich/Suiça. **VLDB Journal**. Zurich/Suíça, 1995. p. 432-444.

SODRÉ, Ulysses. **Transformadas de Laplace**. Notas de Aula Computação, Engenharia Elétrica e Engenharia Civil. S.l., 2003. 39 p.

SRIKANT, Ramakrishnan; AGRAWAL, Rakesh. Mining Quantitative Association Rules in Large Relational Tables. In: ACM SIGMOD International Conference on Management of Data 6/96, 1996, Montreal-Canadá. **Procedings**. Montreal, 1994. p. 1-12.

TUKEY, John W. **Exploratory Data Analysis**. Reading: Addison-Wesley, 1977. p. 95-167

WEKA 3: Data Mining Software in Java. Nova Zelândia. Universidade de Waikato, 2007. Apresenta todas as características do projeto, documentação e o software Weka. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/index.html>. Acesso em: 28 mar. 2007.

WIEDERHOLD, Gio. On the barriers and Future of knowledge discovery. In: FAYYAD et al. **ADVANCES in knowledge discovery and data mining**. Cambridge, Mass:AAAI/MIT Press, 1996. p. VII-XI.

WITTEN, Ian H.; FRANK, Eibe. **Data Mining:** practical machine learning tools and techniques. 2. ed. San Francisco: Morgan Kaufmann, 2005. 525 p.