WILLIAM SÉRGIO AZEVÊDO GUIMARÃES

DATA MINING APLICADO AO SERVIÇO PÚBLICO, EXTRAÇÃO DE CONHECIMENTO DAS AÇÕES DO MINISTÉRIO PÚBLICO BRASILEIRO

Florianópolis – SC

2000

UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM
CIÊNCIAS DA COMPUTAÇÃO

WILLIAM SÉRGIO AZEVÊDO GUIMARÃES

DATA MINING APLICADO AO SERVIÇO PÚBLICO, EXTRAÇÃO DE CONHECIMENTO DAS AÇÕES DO MINISTÉRIO PÚBLICO BRASILEIRO

Dissertação submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação.

Prof. Rogério Cid Bastos, Dr

Florianópolis, novembro 2000

DATA MINING APLICADO AO SERVIÇO PÚBLICO, EXTRAÇÃO DE CONHECIMENTO DAS AÇÕES DO MINISTÉRIO PÚBLICO BRASILEIRO

WILLIAM SÉRGIO AZEVÊDO GUIMARÃES

Esta Dissertação foi julgada adequada para obtenção do grau de Mestre em Ciência da Computação Área de Concentração Sistemas de Conhecimento e aprovada em sua forma final pelo programa de Pós-Graduação em Ciência da Computação.

_____

Prof. Fernando Álvaro Ostuni Gauthier, Dr
Coordenador

Banca Examinadora

_____

Prof. Rogério Cid Bastos, Dr
Orientador

_____

Prof. Milton Luiz Horn Vieira, Dr

_____

Prof. Raul Sidnei Wazlawick, Dr

Aqueles que não se recordam do passado estão condenados a repeti-lo.

George Santayana

Ofereço a minha esposa Ana Paula que com paciência, compreensão e carinho soube resistir às adversidades e a meus três filhos Yuri Henrique, Bárbara Helena e Ana Luiza que são minha fonte inesgotável de energia.

Agradeço a Deus e a meus pais, pois a cada dia aprendo a ama-los mais e a compreender melhor o sentido da vida.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Descoberta de conhecimento como resultado da mineração de grandes bases de dados está se tornando cada vez mais freqüente. A busca por informação que agregue valor a produtos e serviços se tornou estratégica para a maioria dos seguimentos empresariais. Porém, instituições públicas permanentes, que influenciam milhares de pessoas em seus padrões de comportamento e valores têm deixado seus dados gerados por sistemas de informação arquivados.

Neste trabalho é proposta a adoção de ferramentas *Data Mining* para serem inseridas como ferramentas estratégicas de instituições públicas, permitindo a análise de padrões que possam ser levados em consideração na adoção de medidas preventivas e corretivas que causem grande impacto na sociedade.

Como exemplo de aplicação, estuda-se ferramentas baseadas em algoritmos de árvore de decisão e indução de regras, aplicadas ao Controle de Inquéritos Policiais (CIPO), sistema transacional do Ministério Público de Rondônia.

ABSTRACT

The discovery of knowledge as a result of the major database mining is becoming more and more frequent. The search for information that aggregates values to products and services has become strategic for most of the business trend. However, permanent public institutions that have influenced millions of people in their behavior standard and values have left their data generated by information systems in files.

In this paper, the Data Mining tools adoption to be inserted as strategic tools of public institutions, allowing the analysis of standards that can be taken into consideration in the adoption of preventive and corrective measures that cause great impact in the society.

As an example of application, algorithms decision tree and rule induction based tools, applied to CIPO – Investigation Control, Transactional System of Public Ministry of Rondônia.

CAPITULO I

1.1 Introdução

Data Mining ajuda usuários finais a extrair informações estratégicas de seus negócios que estão residentes em grandes bancos de dados. No início do século XX, certamente um proprietário de uma grande loja não necessitaria de *Data Mining*, considerando que ele tivesse apenas algumas centenas de clientes. Provavelmente os clientes seriam chamados pelo nome e certamente seus hábitos de compras seriam todos conhecidos.

Uma grande loja atualmente possui algumas centenas de milhares de clientes que estão distribuídos ao redor do mundo, com hábitos e conceitos que se modificam a todo instante. Suas informações estão sendo geridas de forma consistente por aplicações cada vez mais seguras e armazenadas em computadores com grande capacidade. Para se manter competitiva, uma grande organização necessita de ferramentas que explorem os dados armazenados objetivando descobrir novos padrões ou auxiliando na previsão de comportamentos.

Ferramentas *Data Mining* estão se popularizando e estão sendo empregadas em uma grande diversidade de negócios, não só na indústria mas também nas organizações governamentais que possuem um papel decisivo na sociedade. Este trabalho apresenta o estudo de ferramentas *Data Mining* que implementam algoritmos de aprendizagem supervisionada e não supervisionada aplicadas à aproximações lógicas, permitindo a investigação em tipos de dados numéricos e não numéricos. Para efeito de aplicação prática são utilizadas algumas soluções de mercado aplicadas a um banco de dados do Ministério Público do Estado de Rondônia, que armazena informações sobre infrações criminais, dentre outras.

1.2 Importância do trabalho

Institutos como IDC e Gartner Group afirmam que o momento no mundo da informática é do *Data Warehouse* e *Data Mining*, como conseqüência da necessidade da obtenção de conhecimento em grandes volumes de informações que são produzidas e armazenadas.

Um projeto de *Data Warehouse* tem por finalidade organizar os dados operacionais de uma empresa, em um local em que eles não sejam alterados e estejam modelados e dimensionados de acordo com as necessidades e estratégias do negócio em questão. *Data Mining* faz uso destes dados para permitir análises e proporcionar descoberta de conhecimento, composto por um conjunto de técnicas que, de maneira automática, realiza a exploração de um grande volume de dados, a procura de padrões, tendências e relacionamento entre os dados.

Segundo a AMR Research e o Gatner Group (www.datawarehouse.inf.br), no ano 2000 as vendas de soluções para gerenciamento de relações entre clientes movimentará US$ 5,4 Bilhões no mundo, sendo que US$ 108 milhões na América Latina. Já as ferramentas de *Business intelligence* movimentarão em torno de US$ 36 bilhões no mundo e perto de US$ 720 milhões na América Latina.

Com o crescimento da *internet* e do *e-comerce, Data Mining* deverá ser uma solução viável à empresas que pretendem manter-se competitivas. “*Data Mining* é muito popular hoje, e muitas pessoas estão interessadas em usar estas ferramentas poderosas”. [PYLE99]

Arno Penzias, Prêmio Nobel de física em 1978, autor de livros de sucesso em tecnologia da informação considera *Data Mining* uma aplicação chave e indispensável para as corporações nos próximos anos (ComputerWord, janeiro 1999).

Trabalhos envolvendo exploração de conhecimento tem sido objeto de estudos[PYLE99] ; [DODG98] ;[INMO97] e [KIMB96].

1.3 Objetivos

A Constituição Federal de 1988, em seu artigo 127, define o Ministério Público (MP) como uma instituição permanente, essencial à função jurisdicional, incumbindo-lhe a defesa da ordem jurídica, do regime democrático e dos interesses sociais e individuais indisponíveis. Em resumo, o Ministério Público é o advogado da sociedade, dos interesses sociais e individuais indisponíveis (como o de órfãos e interditos).

Cabe ao MP, exigir dos poderes públicos e dos serviços de relevância pública respeito aos direitos contidos na Constituição, promovendo as medidas necessárias a sua garantia. O MP, portanto, é a instituição que a Constituição Federal atribui a defesa da sociedade.

Como objetivo geral busca-se agregar o uso de ferramentas *Data Mining* de aprendizagem supervisionada e não supervisionada para serem aplicadas nas bases de dados transacionais do Ministério Público do Estado de Rondônia.

Tem-se como objetivos específicos:

- Estudar e revisar os conceitos básicos envolvidos.
- Estudar o processo *Data Mining* com implementação de algoritmos para Árvore de Decisão e para Indução de regras.
- Avaliar ferramentas de *Data Mining* baseadas em algoritmos de Árvore de Decisão e Regras de indução quando aplicadas as Sistema de Controle de Inquéritos Policias do Ministério Público de Rondônia.
- Demonstrar a viabilidade do uso de ferramentas *Data Mining* em instituições Públicas.

1.4 Limitações do trabalho

Quanto a necessidade de informações institucionais, encontra-se limitações a questões relativas a:

- Adoção de ferramentas *Data Mining shareware*;
- Por segredos de justiça é vedada a divulgação de nome de infratores e réus. A Base de dados do Controle de Inquéritos Policiais - CIPO ficou restrita a dados que não violassem esta regra.

Com relação as necessidade de realizar estudos para alcançar os objetivos propostos, encontra-se limitações no escopo do CIPO :

- Área CRIMINAL;
- Infrações relacionadas ao Código Penal e a Lei de Tóxico;
- Informações extraídas do Banco de dados no período de 1995 até 1999 ;
- Em infrações ocorridas no município de Porto Velho;

Quanto as aplicações *Data Mining*, são abordadas técnicas que empregam algoritmos para exploração e descoberta aplicada a aproximações lógicas, relativos a ferramentas que implementam a Árvore de Decisão e a Indução de Regras.

*A aplicação de Data Mining n*a WEB não é coberta por este trabalho, haja vista que o enfoque da pesquisa está relacionado a extração de conhecimento de bases de sistemas transacionais não disponíveis para acesso via *internet.*

1.5 Estrutura do trabalho

Este trabalho está estruturado em 7 capítulos. No primeiro é apresentado e delimitado o problema a ser estudado, identificando a sua importância.

No segundo capítulo é feito um estudo sobre a tecnologia *Data Mining* como processo de descoberta e predição de informações, dando ênfase aos algoritmos de representação lógica.

No terceiro capítulo é feito um estudo das árvores de decisão como ferramenta de descoberta e predição de informações, sendo apresentada os principais algoritmos implementados e suas respectivas funcionalidades.

No quarto capítulo é feito um estudo de indução de regras como ferramenta de descoberta e predição de informações, sendo apresentada a funcionalidade dos algoritmos e ainda uma comparação entre estas regras e a árvore de decisão.

O quinto capítulo apresenta *Data Mining* como processo de descoberta de conhecimento que pode ser aplicado nos dados dos sistemas transacionais de instituições públicas, levando em consideração as diversas fases do *KDD* (*Knowledge Discovery in Database*) e ainda avaliando a implementação de *Data Warehouse* e *Data Marts* como suporte ao processo.

No sexto capítulo é realizada a validação do processo de descoberta de conhecimento através do teste de ferramentas que implementam algoritmos de árvore de decisão e indução de regras aplicados aos dados do CIPO, considerando o modelo apresentado no capítulo anterior.

Por último, o sétimo capítulo apresenta as conclusões e recomendações deste trabalho.

CAPÍTULO II

DATA MINING, TECNOLOGIAS E PROCESSOS

Nos últimos anos observou-se uma onda crescente no nível de interesse por *Data Mining*. Usuários empresariais querem tirar proveito desta tecnologia como forma de estratégia competitiva. O interesse crescente por *Data Mining* também resultou na introdução de uma grande quantidade de produtos comerciais, cada um descrevendo um conjunto de termos que soam semelhantes, mas na realidade recorrem a funcionalidades muito diferentes.

Os gerentes encarregados de selecionar um sistema de apoio a decisão (SAD), face a um desafio que responda às necessidades dos usuários empresariais e que normalmente são urgentes, buscam nos conceitos subjacentes do *Data Mining* respostas mais complexas do que as consultas e os relatórios que retratam a base de dados existente.

Data Mining é "um processo de apoio a decisão no qual procura-se padrões de informações nos dados"[PYLE99]. Esta procura pode ser feita pelo usuário, ou pode ser facilitada por um programa inteligente que busque padrões. Uma vez encontrada a informação, ela precisa ser apresentada numa forma satisfatória, com gráficos, relatórios, etc.

2.1 Processos Data Mining

Há dois tipos de análises estatísticas que podem ser empregadas pelo processo *Data Mining*: análise confirmatória e análise exploratória. Em análise confirmatória, a pessoa tem uma hipótese e a confirma ou a refuta, porém, o ponto de maior dificuldade encontrado na análise confirmatória é a escassez de hipóteses por parte do analista. Em análise exploratória, procura-se hipóteses satisfatórias para confirmar ou refutar, e o sistemas tem a iniciativa da análise de dados, não o usuário.

O conceito da "iniciativa" também aplica-se a espaços multidimensionais. Em um Sistema *OLAP (on line analitical process)*, o usuário pode pensar em

uma hipótese e gerar um gráfico. Mas em *Data Mining*, o sistema pensa por si só nas perguntas [PARS96]. *Data Mining* é um processo automatizado de análise de dados no qual o sistema tem a iniciativa de gerar padrões por si só.

Atividades *Data Mining*: descoberta, modelagem preditiva e análise forense, podem ser vistas na Figura 2.1 [PARS97].

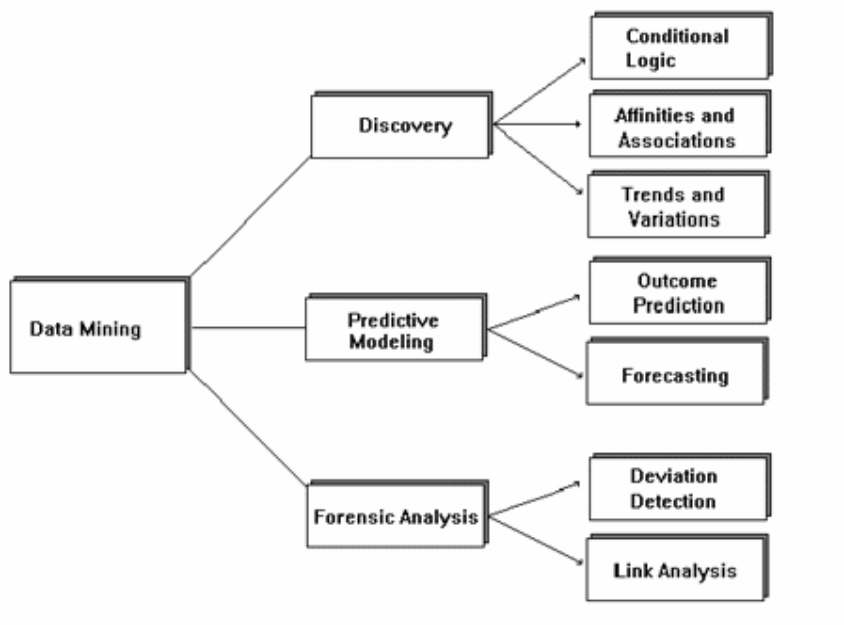


Figura 2.1 – Atividades *Data Mining*

Descoberta é o processo de olhar em um banco de dados para achar padrões escondidos sem uma idéia predeterminada ou hipótese sobre o que os padrões podem ser. Em outras palavras, o programa toma a iniciativa achando os padrões interessantes, sem que o usuário tenha que pensar primeiro nas perguntas pertinentes. Em bancos de dados grandes, há tantos padrões que o usuário nunca poderia pensar praticamente nas perguntas certas. O objetivo principal é o enriquecimento dos padrões que podem ser expressos e podem ser descobertos e a qualidade da informação obtida, determinando o poder e utilidade da técnica de descoberta.

Levando em consideração um banco de dados de uma grande instituição bancária, o usuário pode tomar a iniciativa para fazer uma pergunta, como: " Qual é a idade comum dos correntistas ?" O sistema pode apresentar 47 anos como a idade comum. O usuário pode pedir então para o sistema procurar por

si só algo interessante sobre "idade". O sistema agirá então como um analista humano. Olhará para algumas características de dados, distribuições, etc. e tentará achar um pouco de densidades nos dados que poderiam estar longe do usual. Neste caso o sistema pode apresentar a regra: " SE Profissão = Atleta ENTÃO Idade <30, com um percentual de 71% de exatidão". Esta regra significa que se forem escolhidos 100 atletas do banco de dados, é provável que 71 deles estejam com idade inferior a 30 anos. Considerando idade < 60 anos, há um percentual de 97% de exatidão". Esta regra significa que se forem escolhidos 100 atletas do banco de dados, é provável que 97 deles estejam com idade inferior a 60 anos.

Em modelagem preditiva, padrões descobertos no banco de dados são usados para predizer o futuro. Permite ao usuário submeter registros com poucos valores em campos desconhecidos, e o sistema induzirá os valores desconhecidos baseado em padrões prévios descobertos no banco de dados. Enquanto a descoberta acha padrões em dados, modelagem preditiva aplica os padrões para induzir valores para novos itens de dados.

Usando o exemplo acima, sabe-se que os atletas normalmente possuem idade inferior a 30 anos; pode-se então induzir a idade de alguém se for conhecido que ele é um atleta. Por exemplo, se é apresentado um registro de “João da Silva” cuja profissão é “Atleta” aplicando as regras acima, pode-se ter uma segurança de 70% que ele possui menos de 30, e quase que com certeza absoluta pode-se presumir que ele possui menos que 60 anos. Descoberta ajuda a encontrar "conhecimento geral", mas predição exige um pouco de indução. Neste caso a predição é " transparente " (sabe-se por que acertou-se a idade como abaixo de 30 anos). Em alguns sistemas a idade é induzida, mas a razão para a suposição não é provida, tornando o sistema " opaco ".

Análise forense é o processo de aplicar os padrões para localizar elementos de dados incomuns. Para descobrir o incomum, primeiro defini-se o que é normal, então procura-se valores que divergem do habitual dentro de um determinado limite. Usando o exemplo acima, nota-se que 97% dos atletas possuem idade inferior a 60 anos, pode-se desejar saber sobre os 3% que

possuem mais de 60 anos e se são atletas saudáveis ou praticam esporte onde a idade é menos importante (por exemplo, golfe) ou o banco de dados pode conter erros, etc. Nota-se que descoberta ajuda a encontrar "conhecimento habitual", análise forense procura encontrar casos incomuns e específicos.

Há vários tipos de descoberta de padrões, como regras *If / Then*, associações, etc. Enquanto as regras discutidas acima possuem uma natureza *If / Then*, regras de associação recorrem a agrupamentos de itens (por exemplo, quando alguém comprar um produto em uma loja, ele pode comprar outros produtos ao mesmo tempo, este processo é normalmente definido como análise de cesta de mercado). O poder de um sistema de descoberta é medido pelos tipos e generalidade dos padrões que pode encontrar e expressar em um idioma satisfatório.

2.2 Usuários e Atividades Data Mining

É necessário distinguir os processos de *Data Mining* das atividades *Data Mining*, e os usuários que os executam.

Atividades *Data Mining* normalmente são executados por três classes diferentes de usuários: os executivos, usuários finais e analistas.

Os executivos precisam de perspicácias no nível macro e gastam menos tempo com computadores que os outros grupos, o espaço de atenção deles normalmente é inferior a 30 minutos. Eles podem querer informação além do que está disponível no sistema de informação executivo (EIS). Executivos normalmente são ajudados pelos usuários finais e analistas.

Os usuários finais sabem usar uma planilha eletrônica, mas eles não programam, eles podem passar várias horas ou um dia com computadores. Exemplos de usuários finais são as pessoas de vendas, investigadores de mercado, cientistas, engenheiros, médicos, etc, às vezes, os gerentes assumem o papel de executivo.

Analistas sabem como interpretar dados, mas não são programadores. Eles podem ser os analistas financeiros, estatísticos, consultores, ou desenhistas de banco de dados. Analistas normalmente conhecem algumas

estatísticas e SQL[PARS97]. Usuários normalmente executam três tipos de atividades de *Data Mining* dentro de um ambiente corporativo: *Data Mining* episódico, estratégico e contínuo.

Em *Data Mining* episódico, olha-se para dados de um episódio específico como uma determinada campanha de marketing. Pode-se tentar entender estes dados fixados, ou usar isto para predição em novas campanhas de marketing. Mineração episódica normalmente é executada por analistas.

Em *Data Mining* estratégico, olha-se para conjuntos maiores de dados corporativos com a intenção de ganhar uma compreensão global de medidas específicas como, por exemplo, rentabilidade. Consequentemente, um exercício de mineração estratégico pode procurar responder perguntas como: "de onde nossos lucros vêm ?" ou " como se comporta o nosso segmento de clientes e de produtos padrões ?".

Em *Data Mining* contínuo, tenta-se entender como o mundo mudou dentro de um determinado período de tempo e tenta-se ganhar uma compreensão dos fatores que influenciaram as mudanças. Por exemplo, pode-se perguntar: "Como os padrões de vendas mudaram este mês ?" ou " Quais foram as fontes variáveis de atrito de cliente no trimestre passado ?" Obviamente Mineração contínua é uma atividade em andamento e normalmente leva a lugares aonde a mineração estratégica foi executada para permitir uma primeira compreensão dos assuntos.

Mineração contínua e estratégica é dirigida freqüentemente para os executivos e gerentes, embora os analistas possam ajudar. Diferentes tecnologias são assinaladas para cada um destes tipos de atividade *Data Mining*.

2.3 Tecnologia Data Mining

O nível mais alto de divisão das tecnologias de *Data Mining* pode estar baseado na retenção de dados; quer dizer, há necessidade de manter-se os dados depois que forem minerados? Na maioria dos casos, não. Porém, em

algumas aproximações muito dos conjuntos de dados ainda são mantidos para padrões futuros. Retenção só se aplica às tarefas de modelagem preditiva e análise forense. Descoberta de conhecimento não geram padrões.

Aproximações baseadas em retenção de dados geram problemas por causa dos grandes conjuntos de dados. Porém, em alguns casos, resultados preditivos podem ser obtidos. Apresenta-se na Figura 2.2 [PARS97], aproximações baseadas em estímulo de padrão, que recaem em três categorias: lógica, tabulação e equações. Cada folha da árvore da Figura apresenta um método distinto de implementar um sistema baseado em uma técnica (por exemplo, vários tipos de algoritmos de árvore de decisão).

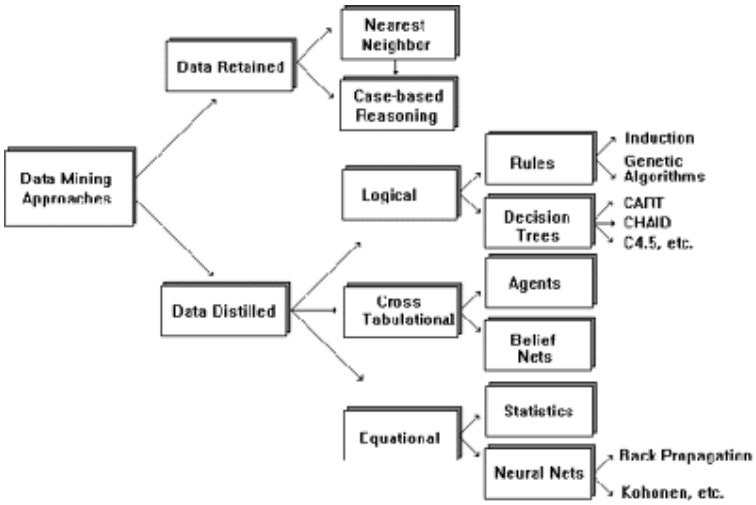


Figura 2.2 – Aproximações *Data Mining*

Nem todas as aproximações baseadas em estímulo de padrão provêem conhecimento, padrões podem não ser estimulados facilmente em um "idioma compreensível" ou formalizados de forma legível por usuários como equações muito complexas.

2.4 Retenção de dados

Um exemplo de uma aproximação baseado em retenção de dados é o método do "vizinho mais próximo" (*nearest neighbor*). Um conjunto de dados é mantido (normalmente em memória) para comparação com itens de dados novos. Quando um registro novo é apresentado para predição, a "distância" entre ele e o registro semelhante no conjunto de dados é determinada, e o mais semelhante (ou mais próximos vizinhos) é identificado.

O termo " vizinho K mais próximo " é usado para indicar que seleciona-se o topo K (por exemplo 10) dos vizinhos para um determinado cliente, como apresentado na Figura 2.3 [PARS97]. Logo, uma comparação mais próxima é executada para selecionar a maioria dos produtos a serem oferecidos para o novo cliente, baseado nos produtos usados pela lista dos K vizinhos mais próximos.

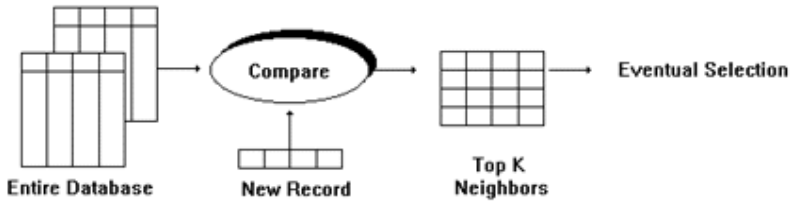


Figura 2.3 – Retenção de Dados

É bastante caro manter todos os dados, e muitas vezes somente um conjunto de "casos típicos" é retido. Pode-se selecionar um conjunto de 100 “clientes típicos" como a base para comparação. Isto é chamado freqüentemente caso baseado em raciocínio.

Um problema normalmente fatal para estas aproximações está relacionado com bancos de dados com um número grande de valores não-numéricos (por exemplo, muitos produtos de supermercado ou partes de um automóvel). Estas distâncias entre valores não-numéricos não são computadas facilmente, algumas medidas tem necessidades de aproximações, isto é

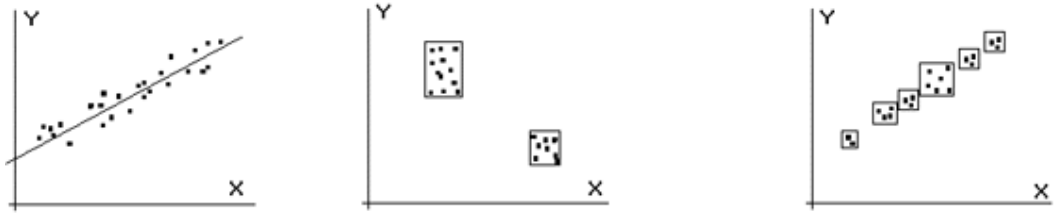
freqüentemente difícil. E se houver muitos valores não-numéricos, haverá muitos casos para serem administrados.

2.5 Estímulo de dados

Tecnologias de estímulos extraem padrões de um conjunto de dados. Naturalmente, as primeiras perguntas a serem respondidas são: Que tipos de padrões pode ser extraído ? e como eles são representados?

Padrões precisam ser expressos dentro de um formalismo e um idioma. Esta escolha dá origem a três aproximações distintas: lógica, equações, ou tabulações. Cada uma dessas aproximações traçam suas raízes históricas para uma origem matemática distinta.

O conceito do idioma usado para expressão de padrão pode ser simplificado com alguns diagramas simples, como na Figura 2.4 [PARS97]. Por exemplo, há uma distinção entre equações e lógica. Em uma equação podem ser usados os operadores de sistemas como " vantagem" e "tempo" para relacionar variáveis, por exemplo, ((A * X) + b) enquanto em um sistema lógico os operadores fundamentais são condicionais, por exemplo, SE 6 < X < 7 ENTÃO 1 < Y< 2.



a) Regressão Linear
Y = (a * X) + b

b) Representação Lógica
SE 6 <X <7 ENTÃO 1 <Y <2

c) Aproximação Universal

Figura 2.4 – Aproximações *Data Mining*

A equação na Figura 2.4a. forma a base da regressão linear e foi extensamente usada para análise estatística. É boa para representar padrões lineares. A aproximação lógica na Figura 2.4b. é melhor para lidar com *crisp box*, mas a Figura 2.4c, pode ser usada para aproximar algumas linhas.

Lógica pode lidar com dados numéricos e não-numéricos. Equações exigem que todos os dados sejam numéricos, ao contrário, tabulações só trabalham com dados não-numéricos; uma fonte fundamental de problemas. O mais importante é que equações computam distâncias de superfícies (como linhas) enquanto tabulações focalizam acontecimentos.

As Redes Neurais são técnicas equacionais opacas; interiormente elas computam superfícies dentro de um espaço numérico. Como dados são alimentados repetidamente na rede, os parâmetros são mudados de forma que a superfície fica mais íntima ao ponto de dados.

Em *Data Mining* é necessário distinguir entre "análise dirigida" e "análise livre". Em análise dirigida, também chamada aprendizagem supervisionada, há um "professor" que ensina o sistema, dizendo quando uma predição estava correta ou incorreta. Aqui os dados tem uma coluna específica em que é usada uma meta para descoberta ou predição.

Em aprendizado não supervisionado, o sistema não tem nenhum professor, mas simplesmente tenta achar agrupamentos interessantes de padrões dentro do conjunto de dados (*dataset)*.

A maioria das aplicações empresariais de *Data Mining* envolve Mineração dirigida, enquanto descoberta não supervisionada às vezes pode ser usada para segmentação de dados ou agrupando (por exemplo, achando classes de clientes para agrupamento).

2.6 Aproximações lógicas

A Lógica forma a base da maioria das linguagens. Padrões expressos em linguagens lógicas são distintos através de duas características principais: por um lado são legíveis e compreensíveis, por outro são excelentes para representar agrupamentos de elementos de dados.

O operador central em uma linguagem lógica normalmente é uma variação na declaração *If / Then* (por exemplo, " Se estiver chovendo, então está nublado"). Porém, nota-se que enquanto a forma mais comum de lógica

for lógica condicional, freqüentemente há necessidade de usar outras formas lógicas como lógica de associação com *When/Also*, (por exemplo, Quando tinta é comprada, também um pincel é comprado).

Sistemas de lógica condicionais podem ser separados em dois grupos distintos: regras e árvores de decisão. Regras condicionais podem ser implementadas por indução ou algoritmos genéticos e há várias aproximações para geração de árvores de decisão (por exemplo, CART, CHAID, C4.5).

2.6.1 Regras

Normalmente são representadas relações lógicas como regras. Os tipos mais simples de regras são condicional expressa ou relações de associação. Uma regra condicional é uma declaração da forma:

Se Condição 1
Então Condição 2

Por exemplo, em um banco de dados pode-se ter a regra: Se Profissão = Atleta Então Idade < 30 anos. Os valores dentro de campos de uma determinada tabela são comparados. Profissão é o atributo e Atleta o valor. Outro exemplo de uma expressão de atributo-valor é " Estado = Rondônia " onde Estado é o atributo e Rondônia o valor.

Regras condicionais normalmente trabalham com atributos de tabelas (campos ou colunas) e valores, como abaixo :

NOME	PROFISSÃO	IDADE
João	Atleta	27
...	...	...

Regras podem ir facilmente além de representações de atributo-valor. Elas podem ter declarações como "Enviado = Recebido ". Em lógica de atributo, comparam-se os valores de dois campos, sem nomear qualquer valor

explicitamente. Esta relação não pode ser declarada por árvores de decisão ou tabulação.

Lógica de afinidade é distinta de lógica condicional em termos de expressão da linguagem e estruturas de dados usadas. Análise de afinidade (ou análise de associação) é a procura por padrões e condições que descrevem como vários itens "se agrupam" ou "acontecem juntos" dentro de uma série de eventos ou transações.

Quando Item1
Também Item2.

Um exemplo, “QUANDO comprar tinta, TAMBÉM comprar pincel”. Um simples sistema de análise de afinidade usa uma tabela de transação como:

Transação #	Itens
123	Pintura
123	Pincel
123	Pregos
124	Pintura
124	Pincel
124	Madeira
125	....

O campo Transação # é usado para agrupar itens, enquanto o campo Itens# inclui as entidades que se agrupam. A afinidade para Transações 123 e 124 são o par (Pintura, Pincel).

Isto é uma estrutura de dados distinta da regra de lógica condicional acima. Uma afinidade dimensional tem a forma:

Confiança = 95%
SE
    Dia = sábado
QUANDO
    item = Pincel
TAMBÉM
    item = Tinta

São combinadas condições lógicas e associações. Esta forma de estrutura híbrida apresenta o poder da lógica transparente.

Regras têm a vantagem de poder lidar com dados numéricos e não-numéricos de uma maneira uniforme. Ao lidar com dados numéricos, algumas aproximações têm que quebrar campos numéricos em "códigos" ou valores específicos. Isto pode remover todas as considerações numéricas efetivamente dos códigos, resultando assim na perda de padrões. Por exemplo, o campo Idade pode precisar ser quebrado em 3 conjuntos (1-30), (31-60), (61-100), correspondendo respectivamente, jovem, de meia-idade e velho. Claro que, os dados podem segurar padrões que sobrepõem quaisquer destes conjuntos (por exemplo, o conjunto (27-34) pode ser muito significativo para alguns padrões e qualquer aproximação baseado em código-assinalado os perderá.

Regras também podem trabalhar bem com dados multidimensionais e OLAP porque eles podem lidar com conjuntos de dados numéricos e os formatos lógicos deles permitem fundir os padrões ao longo de dimensões múltiplas [PARS96].

Regras se parecem às vezes com árvores de decisão, mas apesar da semelhança superficial elas são técnicas distintas e diferentes. Isto é fácil de ser visto quando considera-se o fato que árvores de decisão não expressam associações, ou atributos baseados em padrões como "Enviado = Recebido", onde são comparados os valores de dois campos, sem nomear qualquer valor explicitamente.

2.6.2 Indução de regra

Indução de regra é o processo de olhar para um conjunto de dados e padrões geradores. Explorando os dados fixados automaticamente, como apresenta-se na Figura 2.5 [PARS97], o sistema de indução forma hipóteses que conduzem a padrões.

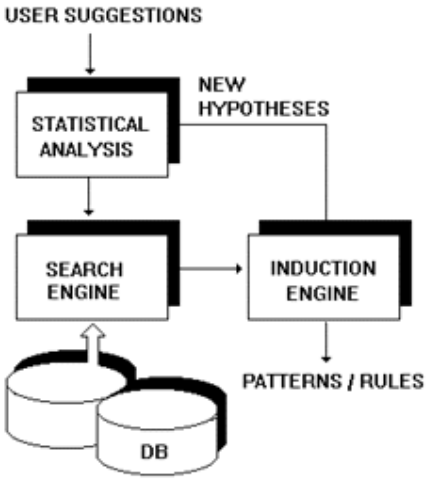


Figura 2.5 – Indução de Regras

O processo é na essência semelhante ao que um analista humano faria em análise exploratória. Por exemplo, escolhido um banco de dados de informação demográfica, o sistema de indução pode olhar primeiro como são distribuídas as idades, e pode notar uma variação interessante para essas pessoas cuja profissão é listada como atleta profissional. Esta hipótese é achada para ser pertinente, então o sistema imprimirá uma regra como:

SE Profissão = Atleta
ENTÃO Idade <30

Esta regra pode ter uma "exatidão" de 70% fixada para isto. Porém, este padrão pode não ser assegurado para as idades de banqueiros ou professores no mesmo banco de dados.

Regras fuzzy são diferentes de regras inexatas. Regras inexatas têm freqüentemente um "fator de confiança fixo" presos a elas, e representam sua

validez. Porém, a confiança em umas regras *fuzzy* pode variar em termos dos valores numéricos no corpo da regra; por exemplo a confiança pode ser proporcional à idade de uma pessoa e como varia a idade assim também varia a confiança. Deste modo regras *fuzzy* podem produzir expressões muito mais compactas de conhecimento e podem conduzir a comportamento estável.

Indução de regra pode descobrir regras muito gerais com dados numéricos e não-numéricos. Regras podem combinar sentenças condicionais e declarações de afinidade em padrões híbridos.

2.6.3 Algoritmos Genéticos

Algoritmos genéticos também geram regras de conjuntos de dados, mas não segue o protocolo de exploração orientada da indução de regra. Ao invés, eles confiam na idéia de "mutação " para fazer mudanças em padrões até uma forma satisfatória de padrão emerge por procriação seletiva, como mostrado na Figura 2.6 [PARS97].

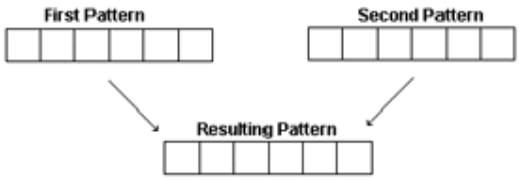


Figura 2.6 – Algoritmos Genéticos

Isto é diferente de indução de regra no foco principal. Algoritmos genéticos são resultantes das combinações de padrões de regras que vêem sendo descobertas há muito tempo, enquanto que na indução de regra a atividade principal é o conjunto de dados(*dataset*).

Algoritmos genéticos não são só para geração de regra e pode ser aplicado a uma variedade de outras tarefas para as quais regras não se aplicam, como a descoberta de padrões em texto, planejamento e controle, otimização de sistema, etc.

2.6.4 Árvores de decisão

Árvore de Decisão expressa uma forma simples de lógica condicional. Um sistema de árvore de decisão simplesmente particiona uma tabela em tabelas menores selecionando subconjuntos baseados em valores por um determinado atributo. Baseados em como a tabela é dividida, alguns algoritmos de árvore de decisão diferentes são o CART, CHAID e C4.5.

Uma árvore de decisão desta tabela é mostrada na Figura 2.7 [PARS97].

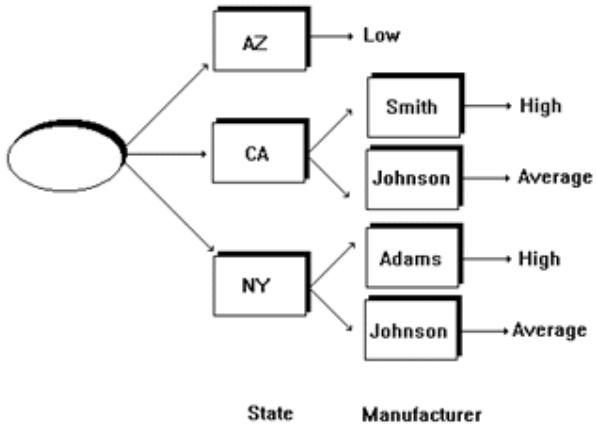


Figura 2.7 – Árvore de Decisão

As técnicas fundamentais usadas para *Data Mining* podem ser classificadas em grupos distintos, por vantagens e áreas de atuação comercial. As técnicas modernas confiam em destilação de padrão, em lugar de retenção de dados. Destilação de padrão pode ser classificada em métodos: lógico, equacional e tabulação. É provável que aproximações híbridas tenham sucesso, enquanto fundindo lógica e equações com análise multidimensional.

CAPÍTULO III

ÁRVORE DE DECISÃO

Árvore de decisão é um modelo preditivo, que como o próprio nome diz pode ser visualizada como uma árvore. Especificamente cada galho da árvore é uma pergunta de classificação e as folhas da árvore são partições de conjuntos de dados com suas classificações. Por exemplo : Numa industria de telefones celulares classificando-se os clientes que não renovam seus contratos telefônicos, uma árvore de decisão pode parecer como o que é mostrado na figura 3.1

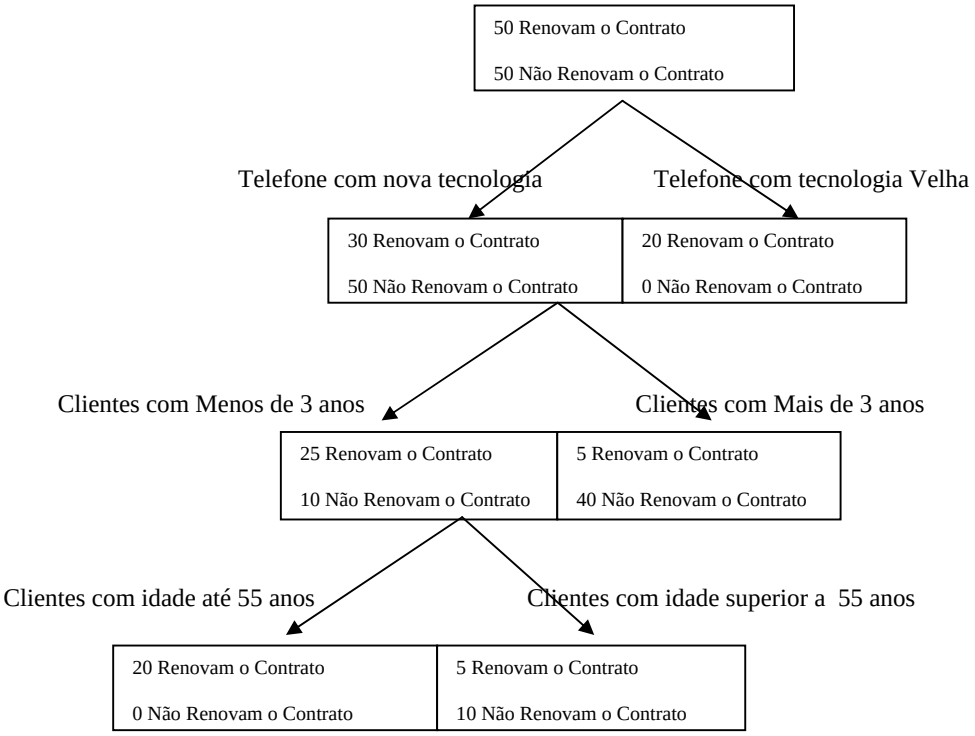


Figura 3.1 - Árvore de Descisão - exemplo

De uma perspectiva de negócios as árvores de decisões podem ser visualizadas como a criação de um segmento de conjuntos de dados originais; cada segmento seria uma das folhas da árvore. Segmentação de clientes, produtos e regiões de vendas são algo que os gerentes de marketing tem feito por muitos anos; no passado essa segmentação foi feita para conseguir uma visualização de alto nível de um grande montante de dados - sem nenhuma razão particular para criar a segmentação exceto em que os registros dentro de cada segmentação fossem similares a qualquer outro.

Neste caso a segmentação é dada para uma razão de predição de umas poucas informações importantes. Os registros que recaem dentro de cada segmento ficam ali por causa de terem similaridade com respeito às informações que estão sendo preditas e não somente porque elas são similares - sem que a similaridade seja bem definida. Estes segmentos preditivos, que são derivados da árvore de decisão, também vêem com uma descrição das características que definem o segmento preditivo; sendo assim, as árvores de decisões e os algoritmos que as criam podem ser complexos, mas os resultados podem ser apresentados de uma forma fácil de serem entendidos e pode ser muito útil para o usuário comercial.

“Por causa do seu alto nível de automação e a facilidade de tradução de modelos de árvore de decisão em SQL para a disposição em banco de dados relacionais, a tecnologia tem também fornecido facilidades de integração com processos existentes, requerendo pouco pré-processamento e limpeza de dados, ou a extração de um arquivo de propósito especial, especificamente para os dados fornecidos”[BERS97].

3.1 – Aplicações Data Mining para árvore de decisão

As árvores de decisões são uma forma de tecnologia fornecedora de dados que tem sido usada de uma forma muito semelhante à tecnologia atual por quase 20 anos. As primeiras versões dos algoritmos datam de 1960. Freqüentemente esta técnica era desenvolvida por estatísticos para

automatizarem um processo para determinar que campos em seus bancos de dados eram na verdade úteis ou correlacionados com o problema particular que eles estavam tentando entender. Parcialmente por conta desta história, os algoritmos de árvore de decisão tendem a automatizar o processo inteiro de geração de hipóteses de forma muito mais completa e muito mais integrada que qualquer outra técnica de *Data Mining*. Elas são também particularmente adeptas do manuseio dos dados crus com pouco ou nenhum pré-processamento. Talvez também por conta de serem originalmente desenvolvidas para fazerem mímica da forma como um analista desempenha o fornecimento de dados, elas fornecem um modelo preditivo, fácil de entender, baseado em regras tais como : 90% dos clientes de cartão de crédito de menos de 3 meses que ultrapassam seus limites de crédito, vão entrar no nível básico para empréstimo de cartão de crédito.

Por conta das árvores de decisões alcançarem níveis tão altos em muitas das características críticas de *Data Mining* elas podem ser usadas em uma ampla variação de problemas comerciais tanto para exploração como para predição. Ela tem sido usada desde predição de atritos no uso do cartão de crédito, até a predição de série de prazo do índice de intercâmbio das moedas internacionais. Geralmente os modelos a serem construídos e as interações a serem detectadas são muito mais complexas em problemas do mundo real, e é por isso que as árvores de decisão se tornam um recurso excelente.

3.1.1 – Exploração de dados

A tecnologia da árvore de decisão pode ser usada para exploração de conjuntos de dados e problemas comerciais. Isto é freqüentemente feito observando-se os preditores e valores que são escolhidos para cada divisão da árvore. Freqüentemente esses preditores fornecem visões que são úteis ou propõem perguntas que precisam ser respondidas. Por exemplo : Se forem considerados os clientes que não renovam contrato, pode-se questionar os

caminhos adotados pelos operadores de vendas e talvez mudar a forma de compensação destes.

Se o tempo do cliente for < 1 ano e o canal de vendas = vendas através de tele vendas, então a chance de não renovação é de 65%.

3.1.2 – Pré-processamento de dados

Um outro meio em que a tecnologia de árvore de decisão tem sido usada é para pré-processamento de dados para outros algoritmos de predição, por conta do algoritmo ser justamente robusto com respeito a uma variedade de tipos de preditores. Por ser um algoritmo executado relativamente rápido, as árvores de decisão podem ser usadas primeiro para a execução de *Data Mining* para criar um subconjunto de preditores úteis, possivelmente que podem ser alimentados em algoritmos de redes neurais ou *nearest-neighbor* (vizinhos próximos), que tomam grande quantidade de tempo para serem executados se houver um grande número de preditores.

3.2.3 – Predição de dados

Embora algumas formas de árvores de decisões sejam inicialmente desenvolvidas em ferramentas de *Data Mining* para refinar e processar dados para tecnologias estatísticas, elas também têm sido usadas cada vez mais para predição; isto é interessante porque muitos estatísticos ainda usarão árvores para análises exploratórias efetivamente para construir um modelo preditivo como um co-produto, mas eles ainda ignoram esse modelo em favor de técnicas com as quais eles estão mais à vontade. Algumas vezes Analistas veteranos farão isto mesmo incluindo o modelo preditivo quando este for superior ao produzido por outras técnicas.

3.2 Tecnologia da árvore de decisão

A idéia geral por trás da tecnologia das árvores de decisão é de que elas podem ser construídas a partir de dados históricos. Elas são uma forma de aprendizagem supervisionada embora sejam freqüentemente usadas também para análises exploratórias.

3.2.1 O crescimento da Árvore

O primeiro passo é do crescimento da árvore, especificamente o algoritmo busca criar uma árvore que funcione perfeitamente quanto possível com todos os dados que estejam disponíveis. A maioria do tempo não é possível ter o algoritmo trabalhando perfeitamente. Há sempre ruído no banco de dados. Há preditores que não estão sendo coletados que tem um impacto no alvo de predição.

A idéia de fazer crescer a árvore é achar a melhor pergunta possível para cada galho ou ponta de galho da árvore, o objetivo é ter as folhas das árvores tão homogêneas quanto possível com respeito ao valor de predição.

Assim a pergunta : o Cliente tem mais de 40 anos? provavelmente não distinguirá suficientemente entre aqueles que não renovam e aqueles que renovam. Assumindo-se que a porcentagem é de 40 por 60, por outro lado pode haver uma série de perguntas que fazem um bom trabalho na distinção daqueles clientes de telefones celulares que não renovarão e daqueles que renovarão o serviço. Talvez a série de perguntas fosse algo como : Tem sido cliente por menos de um ano ? Tem um telefone celular que tenha mais de 2 (dois) anos de uso ? Como o Cliente chegou aqui? Foi via televendas ? ou vendas diretas? Essas séries de perguntas definem um seguimento da população de clientes na qual 90% não renovam contrato, esses são perguntas relevantes para serem feitas em relação a predição de quem vai renovar ou não o contrato. A diferença entre uma boa pergunta e uma pergunta ruim tem a ver com quanto à pergunta pode organizar os dados ou neste caso mudar a probabilidade de uma possível desistência do serviço no seguimento de

clientes. Começando com uma população sendo de metade não renovadora e metade renovadora, então esperaria-se que uma pergunta que não organizasse os dados até um certo grau num segmento, não seria uma pergunta útil a ser feita. Por outro lado, uma pergunta que seria boa para distinguir quem vai desistir e quem não vai desistir, diz que dividindo 100 clientes num segmento de 50 que não renovam e outro segmento que renovam o serviço, seria considerada uma boa pergunta.

Os processos nos algoritmos de árvore de decisões são muito similares quando estão sendo construídas as árvores. Estes algoritmos vêem todas as perguntas possíveis que poderiam segmentar os dados em conjuntos homogêneos com respeito aos valores de predição. Alguns algoritmos de predição podem usar heurísticas para selecionar perguntas. As árvores de regressão e classificação (CART) colhem perguntas sem sofisticação, esses processos testam todas elas, depois pegam a melhor, usando-a para segmentar os dados em no mínimo dois segmentos mais organizados, então de novo faz todas as perguntas possíveis em cada um dos novos segmentos individualmente.

3.2.2 Escolha de preditores

Se o algoritmo da árvore de decisão só continuasse assim, seria concebível criar mais e mais perguntas e galhos na árvore para que eventualmente houvesse um registro no segmento. Para deixar a árvore crescer até este tamanho é computacionalmente caro e também desnecessário. Muitos algoritmos de árvore de decisão param de crescer a árvore quando um dos três critérios abaixo é encontrado :

1) O segmento contém apenas um registro ou algum número mínimo definido algoritimicamente de registros (claramente não há meio de romper com um segmento de apenas um registro em dois segmentos menores e os segmentos com poucos registros não devem ser muito

úteis na predição final, uma vez que as predições que eles estão fazendo não serão baseadas em dados históricos suficientes);

2) O segmento é completamente organizado em apenas um valor de predição. Não há razão para continuar essa segmentação mais adiante uma vez que este dado é agora completamente organizado (A árvore alcançou seu objetivo);

3) A melhoria na organização não é suficiente para garantir a segmentação em duas partes, por exemplo, se o segmento inicial fosse de 90% de pessoas desistentes e os segmentos resultantes das melhores perguntas possíveis fossem 90.001% de desistentes e 89.999% de não desistentes então não teria havido muito progresso.

Considerando o exemplo mostrado na tabela 3.1 de um segmento que pode ser dividido um pouco mais, teriam-se apenas dois exemplos. Supondo que ele foi criado a partir de um banco de dados de clientes muito grande, selecionando apenas consumidores com idade de 27 anos e olhos azuis e com salários que variam entre U$ 80.000 a U$ 81.000 anuais.

Neste caso, todas as perguntas possíveis que poderiam ser feitas sobre os dois clientes, voltariam a ter o mesmo valor : Idade, cor dos olhos e salário. Exceto pelo nome.

Tabela 3.1 – Algoritmo de segmentação de Árvore de Decisão

Nome	**Idade**	Olhos	**Salário ($)**	*Confirmado ?*
Steve	27	Azul	80.000	Sim
Alex	27	Azul	80.000	Não

Este segmento não pode ser dividido um pouco mais que isto exceto usando-se nome como preditor .

Seria possível fazer uma pergunta como : O nome do consumidor é Steve? e criar os segmentos que seriam bons em separar aqueles que não renovam daqueles que renovam o serviço. O problema é que tem-se uma intuição de que o nome do cliente não vai ser um bom indicador para informar

se aquele consumidor renova ou não o contrato; ele pode funcionar muito bem para este segmento particular de dois registros mais é improvável que funcione para outros bancos de dados de clientes ou mesmo para o mesmo banco de dados em um outro espaço de tempo. Este exemplo particular tem a ver com super ajuste do modelo, neste caso, ajustar o modelo muito próximo as indicicrazias dos dados de treinamento.

3.2.3 Escolha do valor correto para o preditor

Como as árvores de decisão selecionam um preditor particular sobre um outro para fazer uma divisão no conjunto de dados ? É feito algum tipo de escolha, uma vez que apenas um preditor é usado em cada ponta de galho da árvore. Para fazer isso, é criada uma medida numérica de divisão que parece mapear uma forma razoável ao que se está buscando em termos da diminuição da desordem do conjunto de dados. Por exemplo , a divisão mostrada na tabela 3.2 é muito boa, já a divisão mostrada na tabela 3.3 não é muito útil.

Tabela 3.2 – Exemplo de melhor segmentação dos dados

Nome	Confirmado renovação de contrato?
SEGMENTO	1
Jim	Sim
Sally	Sim
Steve	Sim
Joe	Sim
SEGMENTO	2
Bob	Não
Betty	Não
Sue	Não
Alex	Não

um exemplo da melhor divisão possível dos dados cria dois segmentos. Cada um homogêneo nos valores de predição. [Berson 97]

Tabela 3.3 – Exemplo de divisão que não melhora nos dados

Nome	Renovação de contrato?
SEGMENTO	1
Bob	Não
Betty	Não
Steve	Sim
Joe	Sim
SEGMENTO	2
Jim	Sim
Sally	Sim
Sue	Não
Alex	Não

Ambos os segmentos misturaram valores para a predição. [Berson 97]

Na primeira divisão o segmento original de oito registros foi dividido em dois segmentos de quatro registros. Um segmento com 100% e outro com 0% de pessoas que não renovam os contratos. Na segunda divisão o segmento de desistentes originais de 50% foi dividido em dois segmentos de 50% de desistência, assim sendo, na primeira divisão foi criado um segmento no qual podia-se prever a desistência em cada nível de confidencia muito alto (100%). Para a segunda divisão podia-se prever a desistência não mais do que foi feita no segmento original de 50% e não haveria nenhum progresso.

Muitos cálculos diferentes podem ser realizados para ordenar as predições e selecionar a melhor. Como é descrito acima, a característica principal da divisão de preditor é a mudança dos valores na densidade de predição depois da divisão ser feita. Uma pessoa poderia pensar nisto como sendo efetivamente a redução da desordem do segmento original em segmentos menores que são mais concentrados em valores de predição particular. Quando a predição tem apenas dois valores, uma pessoa pode efetivamente achar este processo como um isolamento dos dois valores, como

um pastor separando as ovelhas negras das ovelhas brancas em cercados diferentes.

3.3 Algoritmos para árvore de decisão

O algoritmo ainda precisa saber que valor naquele preditor usar. Por exemplo : se o algoritmo fosse usar idade para dividir um segmento dado em duas partes, que valor de idade ele deveria usar ? Idade de 50, porque essa é a metade do caminho entre zero e a idade da pessoa mais velha no banco de dados ou talvez dividir em todas as idades possíveis e dividir em até 100 segmentos menores, como 1,2,3 até 100 ? ou apenas pegar uma idade aleatória, talvez isso não importe tanto.

Isto importa, mais somente à medida que houver meios de medir a eficácia de cada preditor em fazer uma divisão. É também possível medir a eficácia de cada valor de divisão deste. Considerando-se que fossem ordenados dez registros por idade, na verdade seriam obtidos tipos diferentes de divisões, algumas mais valorizadas que outras para um dado preditor. Como exemplo apresenta-se a tabela 3.4.

O melhor valor de divisão possível tem densidades a esquerda e a direita com taxas de porcentagens tanto de 100/0, como de 0/100. Uma vez que esses valores não estão disponíveis no banco de dados, o valor de divisão que mais se aproxima daqueles valores deve ser escolhido. Um provável candidato poderia ser a idade de Sally de 46 anos, onde a divisão seria de 80/20 uma vez que as idades maiores ou menores em torno desta é a idade de 47 do Bob e a idade de 32 para Joe não são tão boas. Tudo depende da medida no algoritmo que esta sendo usado para construir a árvore de decisão.

- Nem todas as idades possíveis precisam ser testadas, apenas aquelas idades que verdadeiramente apareceram no banco de dados precisariam ser testadas desde que qualquer idade que não estivesse no banco de dados terminasse com uma divisão não melhor que a idade mais próxima que estava no banco de dados.

- Não há uma divisão perfeita, não importa que valor de divisão tenha sido escolhido, os segmentos não estarão homogêneos com respeito aos valores de predição. É de fato raro encontrar uma divisão perfeita, exceto em divisões pequenas de dados, em que essa divisão não seria significante estatisticamente.
- O que é considerado uma melhor divisão vai depender da aplicação particular, algumas divisões podem ser menos apuradas mas se aplicam a mais dados e podem ser estatisticamente significantes. E os erros criados por cada segmentos poderiam ser pesados diferentemente.

Tabela 3.4 – Tabela de dados ordenada por idade

Nome	Idade	Conformado	Preditor	D esquerda	D Direita
Keren	21	Sim	<= 21	100% (1/1)	44% (4/9)
Steve	27	Sim	<= 27	100% (2/2)	38% (3/8)
Alex	27	Não	<= 27	67% (2/3)	43% (3/7)
Joe	32	Sim	<= 32	75% (3/4)	33% (2/6)
Sally	46	Sim	<= 46	80% (4/5)	20% (1/5)
Bob	47	Não	<= 47	67% (4/6)	25% (1/4)
Ray	50	Sim	<= 50	71% (5/7)	0% (0/3)
Betty	53	Não	<= 53	62% (5/8)	0% (0/2)
Jim	62	Não	<= 62	56% (5/9)	0% (0/1)
Sue	68	Não	<= 68		

3.3.1 ID3

No final de 1970, J. Ross Quinlan apresentou um algoritmo de árvore de decisão chamado ID3. Este foi um dos primeiros algoritmos de árvore de decisão, embora fossem construídos solidamente em trabalhos prévios sobre sistemas de inferência e sistemas de aprendizagem de conceitos daquela década. Inicialmente o ID3 foi usado para tarefas como aprender boas

estratégias para jogar xadrez, desde então o ID3 tem sido utilizado para uma variedade ampla de problemas tanto na universidade como na Indústria e tem sido constantemente modificado, melhorado e adaptado. O ID3 seleciona preditores e seus valores de divisão, com base no ganho das informações que as divisões fornecem. Ganho representa a diferença entre o montante de informações que são necessários para fazer uma predição correta, tanto antes quanto depois de uma divisão ter sido feita (se o montante de informações requerido for muito menor depois da divisão ter sido feita, então aquela divisão diminuiu a desordem do segmento único original) e é definida como a diferença entre a entropia do segmento original e as entropias acumuladas dos segmentos das divisões resultantes. Entropia é uma medida bem definida da desordem das informações encontradas nos dados.

As entropias dos segmentos filhos são acumuladas em se pesando a sua contribuição para a entropia completa da divisão de acordo com o número de registros que elas contem. Por exemplo : qual das duas divisões mostradas na tabela 3.5 diminuiria a entropia e forneceria um maior ganho ?

Tabela 3.5- Duas divisões possíveis com o cálculo de entropia para cada divisão

Candidato	Divisão a esquerda	Divisão a direita	Entropia esquerda	Entropia direita
Divisão A	++++-	+----	-1/4 lg(4/5) + -1/5 lg (1/5) = 0.72	-1/4 lg(1/5)+ -4/5 lg (4/5)=0.72
Divisão B	+++++---	-	-5/9 lg(5/9) + -4/9 lg (4/9)=0.99	-1/1 lg (1/1) + -0/1 lg (0/1)=0

Os valores positivos e negativos para cada objetivo de predição são representados pelos sinais de (+) e de (-).

A divisão A é na verdade uma divisão muito melhor do que a B, porque separa mais os dados, apesar do fato que a divisão B cria um novo segmento que é perfeitamente homogêneo, zero entropia. O problema é que este segmento de zero entropia perfeito, tem apenas um registro nele. E a divisão de um registro de uma vez não criará uma árvore de decisão muito útil. O número menor de registros em cada um dos segmento (ou seja um) é improvável que forneça padrões úteis. O cálculo métrico que é utilizado para determinar que divisão deve ser escolhida, deverá fazer a escolha correta. A métrica precisa levar em conta dois critérios:

- Quanto que a desordem diminui nos novos segmentos ?
- Como a desordem deve ser medida em cada segmento ?

A medida de entropia facilmente aplicada para cada um dos novos segmentos e também aplicada para o segmento mãe pode responder a primeira pergunta. Mas o segundo critério é um pouco mais difícil. Todos os segmentos que resultam de uma divisão deveriam ser tratados igualmente ? Esta pergunta deve ser respondida no exemplo acima, onde a divisão produziu um novo segmento perfeito, mas com um valor real pequeno por conta do seu tamanho. Olhando-se a entropia média para os novos segmentos, escolher-se-á a divisão B, uma vez que neste caso a média de 0.99 e 0.0 é em torno de 0.5. Pode-se também fazer este cálculo para divisão A e chegar a uma entropia média de 0.72 para os novos segmentos. Se por outro lado pesar a contribuição de cada novo segmento com respeito ao tamanho do segmento e conseqüentemente quanto do banco de dados aquele segmento explicou, obtém-se uma medida muito diferente da desordem que se cruzou através dos dois novos segmentos; neste caso a entropia medida nos dois segmentos para a divisão A é a mesma que antes mais a entropia medida da divisão B é um pouco maior (ver tabela 3.6)

Uma vez que a regra é reduzir a entropia o menor possível, existem duas opções do que seria a melhor divisão. Levando em conta a média das entropias dos dois novos segmentos, escolhe-se a divisão B, tendo como base o número de registros que são cobertos em cada divisão escolhe-se a divisão A.

O ID3 usa o método de entropia por peso para produzir predições melhores do que simplesmente alcançar a média de entropia. Parte da razão para que isto ocorra é que quanto mais dados forem usados na predição, maior a probabilidade que ela esteja correta e também que o modelo combine as razões causais verdadeiras sublinhadas, e processos que estão na verdade funcionando ao formar os valores de predição.

Tabela 3.6 – Valores de entropia para duas escalas possíveis

Candidato	Divisão Esquerda	Divisão Direita	Entropia Média	Peso da Entropia
Divisão A	++++-	+----	0.72=(0.72+0.72)/2	0.72=(1/2)*0.72+ (1/2*0.72)
Divisão B	+++++----	-	0.50=(0.99+0)/2	0.89=(9/10)*0.99+(1/10)*0.0

A melhor das duas divisões possíveis podem ser escolhidas quando as entropias são pesadas pelo tamanho dos segmentos resultantes.

3.3.2 – Preditores de alta cardinalidade no ID3

Exemplos mostrados para ID3 levaram em consideração possíveis preditores binárias. Exemplo, homem/mulher, velho/jovem, alto/baixo. Há outros preditores disponíveis que tem mais valores do que dois. Há aqueles com um pequeno número de valores tais como : cor dos olhos: castanho, azul verde; ou cor do cabelo : loiro, castanho, preto, ruivo. Mais há também preditores com grande número de valores diferentes. Tais como número de unidades de armazenamento para um armazém, que são números de identificações individuais assinalados para cada produto diferente nas prateleiras de uma loja. Um mercadinho típico tem 60.000 unidades de estocagem diferente em um determinado tempo. Estes grandes números de valores para um determinado preditor colocam um novo e único problema para o algoritmo ID3. O algoritmo, por si só, pode ser facilmente modificado para acomodar preditores multivalorados. A equação de entropia pode ser calculada da mesma forma dentro de cada segmento resultante e a equação de entropia pesada pode ser estendida para pesar cada novo segmento pela fração dos dados do segmento mãe que ele contém. Como muitas outras técnicas, o ID3 tem sido melhorado para preditores de alta cardinalidade. Um preditor de alta cardinalidade é aquele que tem muitos valores diferentes possíveis e daí muitas formas possíveis de realizar uma divisão. Um dos problemas clássicos com o uso de um preditor de alta cardinalidade seria usar o campo "nome de cliente" para não renovadores. Isto pode parecer uma coisa sem lógica obviamente, mas muitas árvores de decisão são suficientemente automatizadas para que possam se adaptar a qualquer conjunto de preditores e automaticamente selecionar o melhor, então enquanto um usuário da árvore

de decisão for esperto o bastante para lembrar o nome do cliente desta análise, o usuário da árvore poderá esquecer outros preditores de alta cardinalidade tais como : o número do CPF ou CEP. Eliminação de campos é algo que deveria ser automatizado na árvore de decisão para o usuário. Para fazer isto, a medida para selecionar a melhor divisão que a árvore de decisão está usando, precisa ser melhorada. Assumindo que a medida de desempenho para selecionar uma divisão seja baseada em :

1. O montante em que a desordem ou entropia do segmento de dados original fosse reduzida;
2. Os tamanhos relativos dos segmentos resultantes para que a redução da desordem seja pesada para dar mais peso aos segmentos resultantes maiores;

Considerando-se o seguinte exemplo, onde são divididos os registros pelo nome do cliente e criam-se dez segmentos resultantes do banco de dados. Uma vez que cada segmento é homogeneamente composto inteiramente de todos que não renovam e que renovam, mas contendo apenas um registro, a entropia para cada segmento resultante é zero. Agora não há desordem. Todos os segmentos são de tamanho igual, então seus pesos são idênticos e assim a entropia pesada é também zero. Sendo assim não há outra divisão possível melhor do que esta, incluindo o uso do próprio campo de predição como um preditor, isto é, claramente um problema, pois a métrica de divisão escolheria o nome do cliente como melhor preditor para dividir os registros. Assim, usando o nome do cliente resultará num modelo que nunca poderá ser usado exceto em dados históricos. Para acomodar corretamente preditores de alta cardinalidade a métrica de divisão usada no ID3 pode ser melhorada. Em vez de só tomar a entropia pesada como medida de ganho do sistema a métrica melhorada leva em conta a cardinalidade do preditor. A nova métrica é chamada de índice de ganho, e diferente do ganho ela é relativamente menor se a cardinalidade do preditor for maior. Por exemplo, no caso do preditor do nome do cliente acima, um décimo dos registros recaem em cada um dos novos segmentos. As frações dos conjuntos de dados que recaem em cada segmento são justas

como as frações dos valores de predição dentro de qualquer segmento dado, mas podem ser visualizadas como probabilidades.

- A primeira é a probabilidade que algum registro do segmento mãe recaia em um segmento filho particular;
- O segundo é que a probabilidade de que dado que um registro tenha recaído em um segmento filho determinado, ele tem um valor de predição particular;

Calcular a entropia neste novo conjunto de probabilidades é idêntico a maneira que foi calculada antes. O negativo da soma da somatória de cada probabilidade multiplicado pelo seu logaritmo pela sua base logarítmica 2. Para um preditor nome do cliente, o índice de ganho seria, o ganho dividido pela entropia do tamanho dos segmentos. A entropia do segmento mãe é de 1.0 (desde que ela seja maximizadamente desordenada, consistindo de metade de desistentes e metade de não desistentes). A entropia dos segmentos filhos como dito antes é de 0.0 então o ganho é de 1.0,(1.0-0.0)
A entropia da divisão por si só é justa :

3.32 = -1/10lg(1/10)-1/10lg(1/10)-1/10lg(1/10) -1/10lg(1/10)-1/10lg(1/10)
-1/10lg(1/10)-1/10lg(1/10)-1/10lg(1/10)-1/10lg(1/10)-1/10lg(1/10)

O índice de ganho é então, o ganho dividido pela entropia da divisão ou 0.3 = 1.0 / 3.32

Para se ter idéia como isto funciona e é útil, compara-se este número de índice de ganho com o valor calculado numa divisão prévia melhor(Divisão A). Neste caso a divisão A divide 10 registros em divisões iguais de dois segmentos de 5 registros. A entropia da divisão por si própria poderia ser :

1.0 = -5/10 lg(5/10)-5/10lg(5/10)

Uma vez que o ganho da divisão A foi de 0.28 o índice de ganho seria de 0.28 = 0.28 / 1.0.

Neste caso o campo de alta cardinalidade nome do cliente ainda seria escolhido como o melhor dos dois, mas é uma decisão muito mais próxima do que se o ganho tivesse sido usado em vez do índice de ganho. No entanto esta

não é uma decisão ótima, uma vez que ela é improvável de mostra que há qualquer tipo de relação causal entre o nome do consumidor com a probabilidade de desistência deste. Este relacionamento seria esperado quando a árvore fosse testada. Com algumas técnicas de árvore de decisão, tais como o CART, a métrica para escolha de divisões poderia também permitir que ocorressem erroneamente divisões de alta cardinalidade, mas por causa de um processo de correção ser usado dentro do algoritmo que testa as árvores de decisão contra os dados deixados de lado é provável que uma divisão errônea fosse eliminada durante esta fase do algoritmo.

3.3.3 C4.5

C4.5 é um ajuste do algoritmo ID3 que melhora o desempenho do algoritmo em várias áreas. Preditores com valores faltosos podem ainda ser usados. Preditores com valores contínuos podem ser usados. E ajustes de correção são apresentados e regras podem ser derivadas. Muitas destas técnicas aparecem no algoritmo CART.

3.3.4 CART

A árvores de regressão e classificação (CART) é um algoritmo de exploração e predição desenvolvido por Leo breiman, Jerome Friedman, Richard Olshen e Charles Stone e é muito bem detalhado no livro *Classification and Regression Trees*(Breiman et al.,1984). Estes pesquisadores da universidade de *Stanford* e da universidade da Califórnia, em *Berkeley*, mostraram como este novo algoritmo poderia ser usado em uma variedade de problemas diferentes tais como a detecção de clorina a partir de dados contidos num spectrum de massa.

3.3.4.1 Escolha de Preditores CART

Na construção da árvore CART cada preditor é selecionado baseado em quão bem ele isola os registros com predições diferentes. Por exemplo : uma medida que é usada para determinar se um determinado ponto de divisão para um determinado preditor é melhor do que outro é a métrica da entropia. A medida originou-se do trabalho feito por Claude Shannon e Warren Weaver na teoria de informação de 1949. Eles se preocuparam com a maneira como a informação poderia ser eficientemente comunicada através das linhas telefônicas. Positivamente os seus resultados também forneceram provas úteis na criação de árvores de decisão. A equações de informação que eles derivaram é simples :

$$- \quad \Sigma\, p \lg(p) \qquad (1)$$

onde p é a probabilidade daquele valor de predição ocorrer em um nó particular da árvore, já que p é uma probabilidade ela tem um valor menor de 0.0 e um valor máximo de 1.0, sendo assim o valor da desordem pode ir de um mínimo de 0.0 a um máximo de 1.0.

Por exemplo, tendo-se um nó na árvore em que fosse tentado predizer quem não renovaria e quem renovaria em uma lista de clientes da companhia telefônica celular e tendo-se 100 clientes no total naquele nó, 30 que desistiram e 70 que não desistiram, saber-se-ia que a probabilidade de não renovação no nó era de 30/100 ou p=0.3 e de renovação seria de 70/100 ou p=0.7. A medida da desordem de Shannon e Weaver é :

-0.3 * lg(0.3) + -0.7 * lg(0.7) = (-0.3* -1.74) + (-0.7 * -0.514) = 0.412

Considera-se o nó ótimo nos termos de predição onde cada registro é de um valor de predição ou outro, 100 clientes no total naquele nó, 100 que desistiram e 100 que não desistiram. Neste caso espera-se que a medida de desordem seja menor ou de fato tão baixa quanto pudesse, uma vez que este é o menor nó possível. O valor de desordem para este nó é na verdade 0.0

-1.0 * lg(1.0) + -0.0 * lg(0.0) = -1.0 * 0.0 + -0.0 * ∞ = 0.0

Neste outro extremo a métrica também se comporta apropriadamente. Neste caso quase nenhuma ordem é imposta ao nó, e a desordem está no máximo. 100 clientes é o total no nó, 50 que desistiram e 50 que não desistiram.

Aqui espera-se uma medida de desordem que seja mais alta, para que quando forem escolhidos os preditores das divisões a métrica favoreça as divisões de preditores que sejam importantes para a predição. No caso das árvores de decisões para modelos preditivos a medida de entropia é :

-0.5 * lg(0.5) + -0.5 * lg(0.5) = -0.5 * -1.0 + -0.5 * -1.0 = 1.0

O pior nó possível numa árvore de decisão também corresponde ao valor de desordem máxima da equação de entropia. Outras métricas que são freqüentemente usadas são as métricas de *Twoing* e o índice de diversidade *Gini*. O valor *Gini* para um segmento dado é calculado como sendo 1 menos a soma das somas das probabilidades ao quadrado para cada predição. Assim o valor *Gini* será maior quando as proporções de cada valor das predições forem equivalentes também a entropia mais alta e a mais baixa igual a zero quando o segmento for homogêneo. A métrica *Gini*, como a métrica de entropia em ID3 é pesada pela probabilidade do tamanho proporcional do segmento para comparar a redução geral no valor *Gini* devido a uma divisão particular. O critério de *Twoing* é similar ao critério *Gini*, mas tende a favorecer a divisões mais balanceadas, segmentos de tamanho mais equivalentes que podem ter uma vantagem em evitar problema de nó oscilante onde uma divisão é escolhida para criar um segmento muito pequeno, enquanto a maioria dos registros no segmento não é movida para um segmento de tamanho aproximadamente equivalente no segmento filho. Se o poder usado no cálculo *Gini* for aumentado acima de 2(exemplo, indo de um quadrado das possibilidades para o cubo), ela tende também a favorecer mais divisões balanceadas.

3.3.4.2 Divisões CART

Uma das perguntas mais difíceis para o algoritmo CART é a de como ele determina uma divisão em um preditor com valores desordenados. Por exemplo, uma divisão de duas formas sobre a cor do cabelo, poderia resultar em muitas divisões diferentes:

Castanho,loiro || preto, ruivo
Castanho || loiro, preto, ruivo
Castanho,loiro,preto || ruivo
Castanho,ruivo || preto,loiro

Por este ser um preditor de categoria desordenado não há ordem nos valores, exemplo : o castanho não pode ser menor ou maior que loiro, e o numero total de divisões pares binários que poderiam concebivelmente ser tentados pode se tornar imensa mesmo para preditores de alta cardinalidade moderados. Para resolver este problema CART impõe uma ordem nos valores que podem ser provados ao limite dramático do numero total de divisões que precisam ser testados sem a possibilidade de perder a otimização da divisão.

3.3.4.3 Divisões usando sub-rogação

O algoritmo CART é relativamente robusto com respeito aos dados que faltam. Se o valor está faltando em um preditor particular em um registro particular, esse registro não será usado em determinação da divisão otimizada quando a árvore estiver sendo construída. Com efeito, CART utilizará mais informações quanto tiver para fazer a decisão de selecionar a melhor divisão possível. Quando CART está sendo usado para predizer novos dados de valores em falta, ele manuseia os dados via sub-rogação. Sub-rogação são valores de divisão em preditores que imitam a divisão atual na árvore e podem ser utilizados quando os dados para o preditor preferido estiver faltando. Por exemplo : embora o tamanho do sapato não seja um preditor perfeito de altura,

pode ser usado com sub-rogação para tentar uma divisão baseada na altura quando aquela informação estiver faltando no registro particular que está sendo predito com o modelo CART.

3.3.5 CHAID

Uma outra tecnologia de árvore de decisão muito popular é o CHAID, (detector de interação automático quadrático de CHI). CHAID é similar ao CART na construção da árvore de decisão, mais difere na forma que escolhe suas divisões. Em vez da entropia métrica de *Gini* para escolher divisões otimizadas, a técnica depende de um teste de quadrado CHI usado em tabelas de contingência para determinar que preditor de categoria é mais distante da independência com os valores de predição. Por CHAID depender das tabelas de contingências para formar seu teste de significância para cada preditor, todos os preditores devem ser ou categóricos ou conhecidos dentro de uma forma categórica em agrupamentos (dividir a idade das pessoas em dez partes, de 0 a 9, de 10 a 19, de 20 a 29). Embora este agrupamento possa ter conseqüências deteriorantes, os desempenhos de importância atuais do CART e CHAID têm sido mostrados para serem comparados nos modelos de resposta de marketing do mundo real.

3.4 Medidas Data Mining para árvore de decisão

A tecnologia de árvore de decisão vasculha uma área grande de algoritmos que foram derivados de áreas da inteligência artificial, estatística e um híbridos entre os dois. Por exemplo o ID3 saiu do campo da inteligência artificial e o CHAID da estatística. CART por outro lado cresceu a partir destes dois.

Em geral as forças da árvore de decisão vem da capacidade de criar modelos compreensíveis que são altamente automatizados em sua construção (especificamente com o entrelaçamento de técnicas para prevenir os

superajustes). Os sistemas respondem bem aos dados com ruídos e dados incompletos, embora por causa da predição ser feita através de predições bem definidas há uma chance de que dados com ruídos e incompletos tenham um preditor que estão sendo usados numa resposta de galho seja levado para uma resposta incorreta. Acima de tudo, no entanto, as árvores de decisão recebem altos valores para algoritmos apresentados na tabela 3.7[BERS97].

tabela 3.7 – Medidas Data Mining para árvores de decisões

Medidas DM	Descrição
Precisão	Embora na prática as árvores de decisões sejam iguais ou superiores a muitas outras técnicas de *Data Mining* existem alguns problemas simples em que elas podem produzir modelos complexos.
Clareza	Por causa da representação do modelo como uma árvore, as regras podem ser extraídas e o modelo pode ser claramente visto.
Dados sujos	O algoritmo acomoda dados incompletos muito bem, tanto trabalhando com eles ou usando substitutos para imitar seus efeitos. Dados sujos podem as vezes causar a criação de árvores menos otimizadas e causarem classificação falsa.
Dimensão	As árvores de decisão manuseiam muito bem, grandes números de preditores e freqüentemente são usadas como passos de pré-processamento para fornecer apenas preditores selecionados para algoritmos de *Data Mining*, tais como redes neurais que tem dificuldades com espaços de alta dimensão.
Dados naturais	As árvores de decisão em geral e o CART em particular não requerem quase nenhum pré-processamento de dados.
RDBMS	Os modelos de árvore de decisões são facilmente traduzidos em SQL que podem ser executados diretamente em um RDBMS
Escalabilidade	As implementações de MPP e SMP altamente eficientes de árvores de decisão têm sido criados.
Velocidade	O processo de construção de modelos de árvores de decisões é comparável com outras técnicas. A aplicação do modelo é mais rápida do que muitas outras técnicas. O processo de construção do modelo geral é significantemente mais rápido já que ele mais automatizado do que qualquer outra técnica estatística ou de *Data Mining*.
Validação	A validação é construída na maioria das técnicas de árvore de decisão.

CAPÍTULO IV

INDUÇÃO DE REGRAS

A indução de regras é uma das melhores formas de *Data Mining* e a talvez a forma mais comum de descoberta de conhecimento em sistemas de aprendizagem não supervisionada; é talvez a forma de *Data Mining* que mais se assemelha do processo que muitas pessoas associam com *Data Mining* chamada mineração de ouro, através de um grande banco de dados. O ouro neste caso seria uma regra que é interessante, que diz algo contido no banco de dados que é desconhecido e que provavelmente não poderia de forma explícita aparecer. A indução de regras em um banco de dados pode ser vista como uma tarefa massiva no qual todos os padrões possíveis são sistematicamente puxados dos dados e então calculados com significância e precisão, dizendo aos usuários quão fortes o padrão é, e quão provável ele pode ocorre novamente. As regras que são trazidas do banco de dados são extraídas e ordenadas para serem apresentadas ao usuário de acordo com o número de vezes que estão corretas e pela freqüência que são aplicadas.

Sistemas de indução de regras são altamente automatizados e são provavelmente as melhores técnicas de *Data Mining* para expor todos os padrões preditivos possíveis em um banco de dados. Eles podem ser modificados para uso em problemas de predição, mais os algoritmos que combinam evidências de uma variedade de regras vêm mais das regras de busca minuciosa e experiências práticas. Quando usados para predição são como ter um comitê de conselheiros, cada um com uma opinião um pouco diferente sobre o que fazer, mais relativamente racional, com pés no chão e com boa explicação para o porque deles terem que fazer aquilo.

O sistema de indução de regras tem se mostrado relativamente fácil de entender e fácil de dispor de grande utilidade tanto para descoberta quanto predição. Por causa dos sistemas extraírem todos os padrões interessantes eles não são frágeis nem sensitivos a valores faltosos ou dados com ruídos; de algumas formas os sistemas de indução de regras são semelhantes ao de

tomar uma decisão através de um comitê, muitos membros votam e a pluralidade oferece a decisão. Para sistemas baseados em regras há muitas regras contribuindo para a predição final que asseguram que uma única regra ou um passo errado dado por algoritmo ganancioso não causaria uma predição incorreta a ser feita. Esta decisão por consenso não tem a desvantagem, entretanto, de obscurecer a simplicidade da regra individual que oferece a probabilidade condicional defensiva e muito compreensiva como a razão para a decisão ser feita.

Os sistemas de indução de regras são bem adequados para uma variedade de tarefas e quase tão bem automatizadas quanto alguns dos mais avançados algoritmos de árvore de decisão.

4.1 Definição de regra

Em sistemas de indução de regras, a regra por si só, vem numa forma simples : “SE isto e isto e isto ENTÃO isto”. Por exemplo, uma regra que um supermercado poderia encontrar em seus dados coletados a partir dos scanners seria : Se picles é comprado então ketchup é comprado, ou então:

- se pratos de papel então garfos de plástico;
- se algo para beber, então batata frita;
- se salsa, então tortas;

Para as regras serem úteis, duas partes de informações devem ser fornecidas :

1. Exatidão – Quão freqüente a regra é correta.
2. Cobertura – Quão freqüente a regra se aplica.

Pelo fato do padrão do banco de dados ser expresso como uma regra, não significa que está correto o tempo todo, assim sendo, justamente como em outros algoritmos de extração de dados, o padrão é importante de se reconhecer e de se fazer explicitar quanto a incerteza da regra. Isto é o que a

exatidão da regra significa. A cobertura da regra relata quanto do banco de dados a regra cobre ou se aplica. Exemplos destas duas medidas são mostradas na tabela 4.1.

Tabela 4.1 – Exemplos de exatidão e cobertura

REGRA	EXATIDÂO (%)	COBERTURA (%)
Se cereais de pão são comprados, então leite será comprado.	85	20
Se pão é comprado, então queijo suíço será comprado.	15	06
Se existe um aniversário de 42 anos e salgadinhos são comprados, então cerveja será comprada.	95	0,01

Tanto a exatidão quanto a cobertura são importantes para determinar a utilidade de uma regra. A primeira regra é um padrão que ocorre muito freqüentemente. A terceira regra está quase sempre certa, mas também quase nunca é aplicada.

As regras por si só consistem em duas partes, a parte esquerda é chamada de antecedente e a parte direita é chamada de conseqüente. O antecedente pode consistir de apenas uma condição ou condições múltiplas em que todas podem ser verdadeiras, a conseqüente pode ser verdadeira na exatidão dada. Geralmente a conseqüente é apenas uma condição simples.

4.1.1 Objetivos do uso das regras

Quando as regras são extraídas dos bancos de dados elas podem ser usadas tanto para melhorar a compreensão dos problemas comerciais que os dados refletem quanto para desempenhar as atuais predições contra alguns alvos de predição pré-definidos. Uma vez que haja tanto um lado esquerdo quanto um lado direito para uma regra (antecedente e conseqüente), eles podem ser usados de várias formas.

1. **Almejando o antecedente -** Neste caso todas as regras que tenham um certo valor para o antecedente são coletadas e mostrados ao usuário. Por exemplo, um mercadinho pode querer todas as regras que tem pregos, parafusos ou porcas no antecedente para tentar entender que se

descontinuar a venda destes itens de baixa margem de lucro, trará qualquer efeito sobre os outros itens de margem de lucro mais altos; por exemplo, talvez as pessoas que compram pregos também possam comprar martelos, mas não o façam numa loja onde os pregos não estiverem disponíveis.

2. **Almejando o conseqüente -** Neste caso todas as regras que tem um certo valor para o conseqüente podem ser usadas para entender o que está associado e talvez o que afeta o conseqüente. Por exemplo, poderia ser útil saber todas as regras interessantes que tem café como seu conseqüente. Estas bem que podem ser as regras que afetam a compra do café e que o proprietário do mercadinho possa querer então colocar o café junto para aumentar a venda de ambos os itens. Ou poderia ser a regra que o fabricante do café usa para determinar em que revistas deve colocar os seus próximos cupons.
3. **Almejando com base na exatidão -** Algumas vezes a coisa mais importante para um usuário é a exatidão das regras que estão sendo geradas. As regras altamente exatas de 80% a 90% implicam fortes relacionamentos que podem ser explorados mesmo se eles tem cobertura baixa do banco de dados e ocorrerem apenas um número limitado de vezes. Por exemplo, uma regra que tenha apenas 0,1% de cobertura e uma exatidão de 95% pode ser aplicada apenas uma vez em 1000, mais provavelmente estará correta. Se essa única vez for altamente lucrativa então pode valer apenas. Esta, por exemplo, é como algumas das aplicações de *Data Mining* mais bem sucedidas funcionam no mercado financeiro, buscando aquele montante de tempo limitado em que uma predição muito confiante pode ser feita.
4. **Almejando alvo baseado na cobertura -** Algumas vezes os usuários podem querer saber quais são as regras mais presentes ou aquelas regras que são mais prontamente aplicadas. Olhando as regras classificadas pela cobertura elas podem rapidamente conseguir uma visão de nível mais alto do que está dentro do seu banco de dados na maioria das vezes.
5. **Almejando alvos baseados naquilo que é interessante -** As regras são interessantes quando possuem exatidão alta e cobertura alta e desviam do

normal. Tem havido muitas formas em que regras têm sido classificadas por alguma medida de interesse para que o comércio entre a cobertura e a exatidão possa ser feita.

Desde que o sistema de indução de regras começou a ser usado para descoberta de padrões e aprendizagem não supervisionada é difícil comparar regras. Por exemplo, é fácil para qualquer sistema de indução de regras gerar todas as regras possíveis, no entanto é muito mais difícil apresentar aquelas regras que podem facilmente ser muito mais úteis para o usuário final. Quando as regras interessantes forem encontradas, centenas de milhares delas terão sido criadas para encontrar relacionamentos entre muitos valores preditores diferentes no banco de dados, não apenas um alvo bem definido de predição. Por esta razão é freqüentemente mais difícil determinar uma medida de valor para a regra independente de seu grau de interesse. Por exemplo, seria difícil determinar o valor monetário de saber se as pessoas que compram salsicha para o café da manhã, também comprariam ovos em 60% das vezes.

4.1.2 Restrição

É importante reconhecer que mesmo os padrões produzidos a partir de sistemas de indução de regras sejam dados como regras "*If / Then*", elas não necessariamente significam que a parte esquerda da regra(*If*) cause o lado direito da regra(*Then*); comprar queijo não quer dizer que se vá comprar vinho, mesmo que a regra "se queijo então vinho" possa ser muito forte.

Esta é uma particularidade importante para ser lembrada em sistema de indução de regra, porque os resultados são apresentados como: "Se isto, então aquilo" como muitas relações causais são apresentadas.

4.1.3 Tipos de dados usados para indução de regra

Tipicamente a indução de regras é usada em banco de dados tanto com campos de alta cardinalidade (muitos valores diferentes) como com muitas colunas de campos binários. O caso clássico é do carrinho de supermercado, que contem os nomes dos produtos individuais, quantidades e podem conter dezenas de milhares de itens diferentes, que com empacotamentos diferentes criam centenas de milhares de identificadores de unidades de armazenamento.

Às vezes o conceito de um registro não é facilmente identificado dentro do banco de dados. Considere o esquema estrela típico para muitos depósitos de dados que armazenam as transações de supermercado como entradas separadas na tabela de fato. As colunas na tabela de fatos representam alguns identificadores exclusivos do carrinho de compras (para que todos os itens possam ser notados como estando no mesmo carrinho de compras), tais como a quantidade, o tempo de compra e se o item foi comprado com alguma promoção especial (venda ou cupom), assim sendo cada item na lista de compra tem uma fileira diferente na tabela de fato. Esse layout dos dados não é o melhor para os algoritmos de *Data Mining* que prefeririam ter os dados estruturados como uma fileira por carrinho de compras e cada coluna representasse a presença ou ausência de um dado item. Isto pode ser uma forma cara de armazenar os dados, no entanto desde que o mercadinho típico contenha 60.000 unidades de estocagem ou itens diferentes que poderiam ser cruzados pelo contador de checagem de saída de produto. Esta estrutura de registro pode também criar um espaço dimensional muito alto. 60.000 dimensões binárias que seriam desprovidos para muitos algoritmos de *Data Mining* clássicos, como as redes neurais e as árvores de decisão. Como podemos ver, várias armadilhas são colocadas em jogo para tornar esta computação frágil para o algoritmo de *Data Mining* enquanto não requerer uma reorganização massiva do banco de dados.

4.1.4 Descoberta de informação

A fama reivindicada pelos sistemas de indução de regras é muito mais para descoberta em sistemas de aprendizagem não supervisionados do que para predição. Esses sistemas fornecem tanto uma visão detalhada dos dados, onde padrões significantes ocorrem em uma posição do tempo e podem ser encontrados apenas quando se vêem os dados em detalhes, como também numa visão geral dos dados onde alguns sistemas buscam passar para o usuário uma visão geral dos padrões contidos no banco de dados. Esses sistemas então mostram uma combinação boa, tanto das visões micro como macro:

1. **O nível macro.** Padrões que cobrem muitas situações são fornecidos aos usuários para serem usados muito freqüentemente e com alta confiabilidade podem também ser usado para resumirem o banco de dados.
2. **O nível micro.** Regras fortes que cobrem apenas poucas situações podem ainda ser recuperada pelos sistemas e serem propostas ao usuário final. Estas podem ser de grande valor se as situações que são cobertas são altamente valorizadas (talvez elas se apliquem aos consumidores mais lucrativos), ou representam uma pequena população mas que seja crescente, o que pode indicar um mercado ou a emergência de um novo competidor. (Por exemplo: clientes estão sendo perdidos apenas em uma área particular do país onde um novo competidor está emergindo).

4.1.5 Predição de informação

Depois das regras ser criado e seu grau de interesses serem medidos, há também uma chamada para medida de predição com as regras. Cada regra por si só pode desempenhar predição. O conseqüente é o alvo e a exatidão da regra é a exatidão da predição, mas por conta dos sistemas de indução de

regras produzirem muitas regras para um dado antecedente ou conseqüente, pode haver predições conflitantes com as exatidões diferentes. Esta é uma oportunidade para melhorar o desempenho geral dos sistemas combinando as regras. Isto pode ser feito com uma variedade de formas somando-se as exatidões como se elas fossem pesos ou simplesmente tomando a predição da regra com a máxima exatidão possível.

A tabela 4.2 mostra como um dado conseqüente ou antecedente pode ser parte de muitas regras com exatidão ou coberturas diferentes. O problema da predição é tentar prever se o leite foi comprado só com base nos outros itens que estavam no carrinho de compra. Se o carrinho de compras continha apenas pão, então da tabela poderia adivinhar-se que havia uma chance de 35% de que leite fosse também comprado. Se, no entanto pão, manteiga, ovos e queijo foram comprados, qual seria a predição para leite então? A resposta de 65% teria chance de que o leite fosse comprado, pois a relação entre manteiga e leite é maior que 65% ou todos os outros itens aumentariam a chance de leite ser comprado para uma porcentagem bem além de 65%. Combinar evidência de regras múltiplas é a parte chave dos algoritmos para uso de regras para predição.

Tabela 4.2 – Cobertura e exatidão nos antecedentes e conseqüentes

Antecedentes	Conseqüentes	Exatidão (%)	Cobertura (%)
Pão bengala	Queijo cremoso	80	05
Pão bengala	Suco de laranja	40	03
Pão bengala	Café	40	02
Pão bengala	Ovos	25	02
Pão	Leite	35	30
Manteiga	Leite	65	20
Ovos	Leite	35	15
Queijo	Leite	40	08

Um único antecedente pode predizer múltiplos conseqüentes, da mesma forma como muitos antecedentes diferentes podem predizer o mesmo conseqüente.

4.2 Exatidão e cobertura

A idéia geral de um sistema de classificação de regras é que as regras são criadas para mostrar a relação entre eventos capturados no banco de

dados. Essas regras podem ser simples, com apenas um elemento no antecedente, ou podem ser mais complicadas, com muitos pares de colunas no antecedente, todas juntas através de uma conjunção (item1 e item2 e ...).

As regras são usadas para encontrar padrões interessantes no banco de dados, mas também são usadas as vezes para fazer predição. Duas coisas são importantes para entender uma regra:

- **A Exatidão -** A probabilidade de que se um antecedente é verdadeiro então o conseqüente será verdadeiro. Alta exatidão significa que esta é uma regra que é altamente dependente.
- **A Cobertura** – O número de registro no banco de dados que a regra se aplica. Alta cobertura significa que a regra pode ser usada muito freqüentemente e também que é menos provável que ela seja um artefato estimulante da técnica de amostra do banco de dados

A partir de uma perspectiva de um comércio as regras exatas são importantes porque elas implicam que há informações preditivas úteis no bano de dados que podem ser exploradas. Nominalmente existe algo longe de ser independente entre o antecedente e conseqüente. Quanto menor for a exatidão, mais próximo fica a regra da suposição aleatória das coisas. Se a exatidão estiver significantemente abaixo do que seria esperado da suposição aleatória, então a negação do antecedente pode, na verdade, ser útil (por exemplo, pessoas que usam dentaduras postiças , não são compradoras de milho ou sabugo como seria normal com outras pessoas).

A partir de uma perspectiva de negócio a cobertura implica o quão freqüente uma regra útil pode ser usada. Por exemplo, pode existir uma regra que seja exata 100%, mas que só é aplicável em apenas 1 em cada 100.000 carrinhos de compras. Mesmo sendo possível rearranjar as prateleiras para levar vantagem deste fato, isto não permitirá um ganho muito grande de dinheiro, uma vez que o evento não é muito provável de acontecer. A tabela 4.3 mostra a relação entre cobertura e exatidão.

Tabela 4.3 – cobertura x exatidão

	Baixa exatidão	Alta exatidão
Alta Cobertura	A regra é raramente correta mais pode ser usada freqüentemente	A regra é geralmente correta e pode ser usada freqüentemente
	A regra é raramente correta e só pode ser usada raramente.	A regra é geralmente correta, mas pode ser usada raramente.

Tanto a exatidão quanto a cobertura terão impacto sobre a valorização da regra.

Uma analogia entre cobertura e exatidão em fazer dinheiro é a seguinte : A partir de jogos de aposta em cavalos. ter uma regra de alta exatidão com baixa cobertura, seria como possuir um cavalo de corrida que sempre ganha quando corre, mais que só pode correr uma vez por ano.

4.2.1 Avaliação de regras

Uma forma de ver a exatidão e cobertura é ver como elas se relacionam a uma estatística simples e como elas podem ser representadas graficamente. Em estatística a cobertura é simplesmente a probabilidade da ocorrência do antecedente. A exatidão é simplesmente a probabilidade da condicional conseqüente sobre o precedente, então por exemplo, olhando para o banco de dados dos carrinhos de supermercado, serão necessárias as seguintes informações para calcular a exatidão e cobertura para uma regra simples, “leite comprado implica em ovos comprados”.

- T = 100 = Número total de carrinhos no banco de dados
- E = 30 = Número de carrinhos com ovos dentro
- M= 40 = Número de carrinho com leite dentro
- B = 20 = Número de carrinhos com ovos e leite dentro

A exatidão é apenas um número de carrinho com ovos e leite, divididas pelo número de carrinhos com leite nelas. Neste caso seria 20/40=50%. A cobertura seria o número de carrinhos com leite dividido pelo número total de

carrinhos. Isto seria 40/100 = 40%. Isto pode ser visto graficamente na figura 4.1. Não usou-se **E** o número de carrinhos com ovos nestes cálculos. Uma forma em que os ovos poderiam ser usados, seria calcular o número esperado de carrinhos com ovos e leite em referência a independência de eventos. Isto daria sentido de quão improvável e quão especial o evento seria de que 20% das carrinhos continham tanto ovos quanto leite. Se dois eventos são independentes, e não tiverem efeito um sobre o outro, o produto de suas probabilidades individuais de ocorrências deveria ser igual a probabilidade de ocorrência deles juntos.

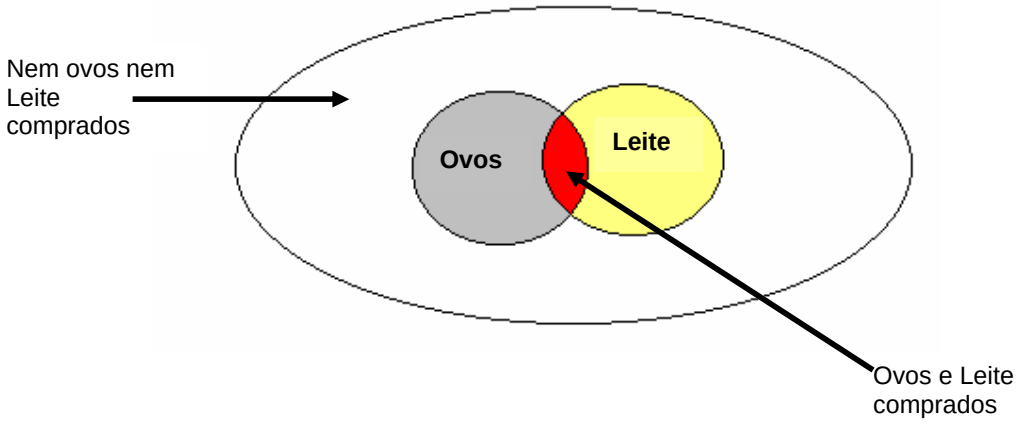


Figura 4.1 – Exatidão e Cobertura

A cobertura da regra : Se leite, então ovos, é apenas um valor ao tamanho relativo do círculo correspondente ao leite. O grau de exatidão é o tamanho do círculo maior sobreposto entre os dois, relativos ao círculo representando o leite comprado.

Se a compra de ovos ou leite for independente uma da outra, espera-se ver 12% dos carrinhos de compras tanto com ovos quanto com leite neles, já que 0.3 x 0.4 = 0.12 ou 12%. O fato de que essa combinação de produtos ocorra 20% das vezes estaria fora de ordem, se esses eventos fossem

dependentes. Em outras palavras, há uma boa chance de que a compra de um afete a compra do outro, e o grau que este seja o caso pudesse ser calculado através de testes estatísticos e testes de hipóteses.

4.2.2 Conjunções e Disjunções

Não há uma razão particular para que a parte antecedente da regra não possa ser mais complexa do que apenas um único item ou valor de predição. Por exemplo, pode haver uma regra útil que mostre que se três produtos são comprados juntos, então é altamente provável que um quarto seja comprado. Não há um limite teórico para o número de restrições (coações) que seriam enlaçadas juntas através do AND no antecedente. Há, no entanto vários limites práticos, o principal deles é que a cobertura da regra diminui drasticamente à medida que o número de restrições no antecedente seja aumentado. Mostra-se isto graficamente na figura 4.2.

Toda vez que uma nova restrição for adicionada a regra, a cobertura pode se tornar muito menor, especialmente para restrições que não se sobreponham muito. No exemplo dado para compras no supermercado, o número de pessoas comprando coca cola e batata frita podem ter uma grande sobreposição e assim uma alta cobertura. A sobreposição entre outros produtos pode ser menor. Por exemplo, uma caixa de grãos orgânicos de baixa gordura e pouco açúcar e um pacote de batatas fritas são muito pouco prováveis de serem comprados juntos no mesmo carrinho.

Toda vez que uma restrição é acrescentada a um antecedente da regra, a cobertura da regra é diminuída. Ela também pode ser dramaticamente aumentada se for usado o conector lógico *OR*, ou disjunção das restrições para formar uma regra em vez do conector lógico *AND* ou conjunção das restrições. Se uma das disjunções for usada, a cobertura da regra pode ser aumentada com as restrições adicionais do antecedente; por exemplo, a regra: “SE coca cola dietética ou coca cola comum ou cerveja ENTÃO batata frita”. Cobrirá

muito mais carrinhos de compra do que simplesmente uma das restrições sozinhas e muito mais do que restrições juntas como uma conjunção.

Embora as regras que são criadas a partir de uma disjunção de restrições sejam perfeitamente aceitas e compreendidas, a geração de regras é geralmente limitada a conjunções de restrições quando elas estiverem sendo geradas, uma vez que adição de restrições conjuntivas limite a cobertura para que ela possa auxiliar na busca por regras interessantes. A disjunção de restrições acontece implicitamente entre regras conjuntivas diferentes. Por exemplo, as duas regras :

- Se leite então ovos
- Se queijo então ovos

Podem ser separadamente gerados por um *Data Mining* mas usadas juntas para cobrir efetivamente todas as instâncias em que tanto leite quanto queijo forem comprados. Em alguns sistemas de indução de regras, estas serão geradas como conjunções de restrições e então mais tarde combinadas juntas como disjunções.

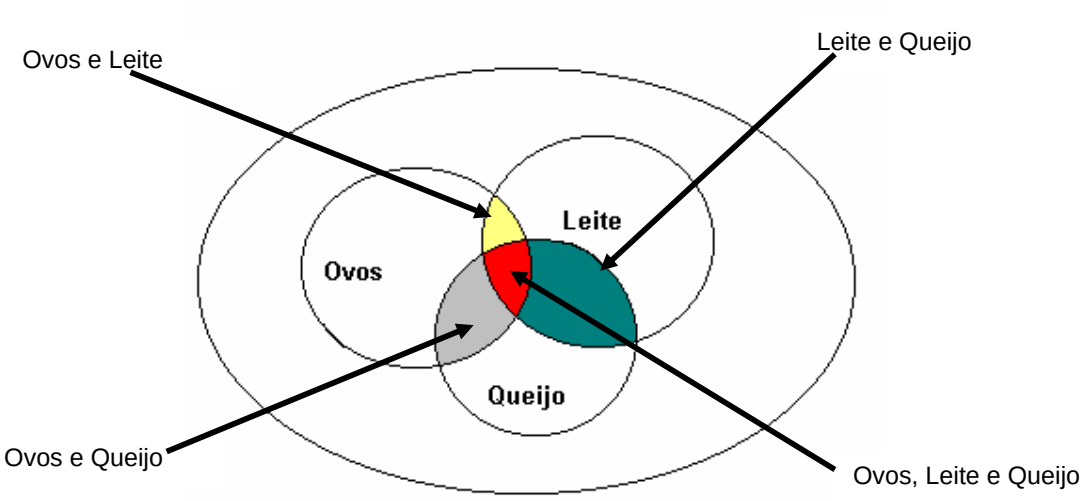


Figura 4.2 - Restrições

Cada área sobreposta representa o and lógico ou conjunção dos dois produtos que estão no mesmo carrinho de supermercado.

4.2.3 Grau de interesse da regra

Um dos maiores problemas com o sistema de indução de regras as vezes é o número exagerado de regras que são produzidas, muitas das quais não tem valor ou interesse prático. Algumas das regras são tão inexatas que não podem ser usadas. Algumas têm pouca cobertura, embora sejam interessantes, mas tem pouca aplicabilidade e finalmente muitas das regras captam padrões de informações que o usuário já está familiarizado com elas. Para combater este problema, pesquisadores implantaram meios de medir a utilidade e grau de interesse de regras. Certamente qualquer medida de grau de interesse teria algo a ver com exatidão e cobertura.

1 - Grau de interesse igual a zero. Se a exatidão da regra for igual a exatidão prévia (uma probabilidade a priori do conseqüente). O exemplo na tabela 2.6 ilustra este ponto. Neste caso uma regra para competição não seria melhor do que simplesmente adivinhar o índice geral de competição ou de desgaste.

2 - O grau de interesse aumenta a medida que a exatidão aumenta (ou diminui com a exatidão decrescente se a cobertura for fixada).

3- O grau de interesse aumenta ou diminui com a cobertura se a exatidão ficar fixa.

4- O grau de interesse diminui com cobertura para um número fixo de respostas corretas. (Exatidão é igual ao número de respostas corretas dividida pela cobertura)

Várias medidas de grau de interesse que tem essas características gerais são usadas para podar o número total de regras que poderiam ser geradas e então serem apresentadas ao usuário. A tabela 4.4 mostra regras que não comparam nenhuma regra para competição, o que por *default* tem a probabilidade a priori ou índice de precedência com outras regras com a mesma exatidão (10%) ou regras com exatidão mais altas mas com cobertura baixa e que nunca possam ser usadas. Estas não seriam regras interessantes.

Tabela 4.4 - Regras sem significância

Antecedente	Conseqüente	Exatidão (%)	Cobertura (%)
Sem restrição	então o cliente entrará em atrito	10	100
Se o saldo do cliente for de R$ 3.000,00	então o cliente entrará em atrito	10	60
Se os olhos do cliente forem azuis	então o cliente entrará em atrito	10	30
Se o CPF do cliente for 144308217	então o cliente entrará em atrito	100	0.000001

4.2.4 Outras medidas para avaliar regras

Embora a cobertura e a exatidão sejam provavelmente os dois aspectos mais importantes de uma regra, especialmente para predição (pode-se dizer a freqüência com que se estará certo e outros vão dizer com que freqüência pode-se adivinhar)

Ainda há outras medidas que podem ser levadas em conta. Por exemplo, uma outra medida que é freqüentemente usada é o suporte a uma regra, que é o percentual (%) de vezes que tanto o antecedente quanto conseqüente ocorrem. Isto pode ser calculado simplesmente com a exatidão multiplicada pela cobertura, mas o suporte é freqüentemente mostrado por si próprio para dar uma indicação da freqüência geral de uma regra dada em um banco de dados.

Uma outra medida é capturar a cobertura não como uma probabilidade ou porcentagem, mas com o número total de registro que se encaixam com o antecedente dado. Há uma boa razão para capturar a cobertura com o número de registros em vez de uma probabilidade, porque a variação estatística na regra baseada em amostra será dependente do número de registro e não a fração do banco de dados que ele cobre. Assim sendo, se a cobertura é apresentada como uma fração, de 30%, este número não significa 3/10 ou 3.000/10.000. Se a cobertura é reportada como 3 x 3.000 então o usuário consegue ter uma idéia sobre o tanto que a avaliação pode ser esperada da cobertura e exatidão da regra.

Duas regras podem ter a mesma cobertura quando medidas pela proporção, mas em níveis muito diferentes de confiabilidade. Esta outra medida de variação pode ser facilmente calculada a partir da primeira. Multiplicando-se a cobertura como uma probabilidade pelo tamanho total do banco de dados usado.

Uma outra medida que tem sido usada em todas as estatísticas e todas as técnicas de *Data Mining* é a comparação de padrões dados para chances aleatórias. Isto seria a nossa medida fora do comum. Como em estatística, dois eventos são considerados independentes se o produto de suas probabilidades for igual a probabilidade de ambos eventos ocorrerem juntos. Este fato é colocado para uso na estatística do quadrado CHI e poderia ser usada no sistema de indução de regras. Por exemplo, a regra: Se leite e ovos então pão, pode parecer muito interessante com cobertura de 30% em uma exatidão de 60%. Mas pode também não ser muito fora do comum. O suporte para esta regra seria de 0.3 x 0.6 = 0.18 (esta regra é encontrada 18% das vezes no banco de dados). Se no entanto ovos ocorrerem em 60% dos carrinhos, leite em 70% e ovos em 40% das carrinhos, então por uma simples chance aleatória, se esses eventos forem completamente dependente um do outro, nós esperaremos que a probabilidade destes três eventos ocorrendo juntos sejam de 0.17 (0.6 x 0.7 x 0.4), o que é aproximadamente a freqüência atual desses eventos no banco de dados. Sendo assim embora tenha-se encontrado uma regra com cobertura e exatidão alta, pode não ser de muito interesse por que ela meramente reflete os eventos independentes aleatórios no banco de dados, não há qualquer conexão com os eventos como a regra implica.

Uma outra medida importante é o da simplicidade da regra. Isto é importante simplesmente para o usuário final. Regras complexas, como também poderosas, podem ser difíceis de entender ou confirmar através de intuição, sendo assim o usuário tem o desejo de ver regras mais simples e esse desejo pode se manifestar diretamente nas regras que são escolhidas e fornecidas automaticamente para o usuário.

Finalmente uma medida de inovação é requerida durante a criação das regras para que as regras que forem redundantes sejam menos favorecidas de

serem buscadas do que as regras que não podem ser tão fortes, mas que cubram exemplos importantes que não sejam cobertas por outras regras fortes. Por exemplo, pode haver poucos registros históricos, para fornecerem regras para um item de armazém pouco vendido, por exemplo, geleia de menta: elas podem ter uma exatidão baixa, mais uma vez que haja tão poucas regras, mesmo quando elas não são interessantes elas serão novidades e deverão ser retidas e apresentadas para o usuário apenas por esta razão.

4.2.5 Regas x árvores de decisões

As árvores de decisões também produzem regras, mas em um formato diferente do sistema de indução de regras.

As árvores de decisão produzem regras que são mutuamente exclusivas e coletivamente exaustivas com respeito ao banco de dados. Enquanto que o sistema de indução de regras produz regras que não são mutuamente exclusivas e podem ser coletivamente exaustivas.

Isso significa que para um dado registro haverá uma regra para cobri-lo e haverá apenas uma regra para regras que venham das árvores de decisão.

Pode haver muitos registros que se encaixem com as regras dadas a partir de um sistema de indução de regras e para muitos sistemas não é garantido que uma regra existirá para cada, e todo registro possível que poderia ser encontrado (embora muitos sistemas criem muitas regras gerais padrões para capturar esses registros). A razão para essa diferença é a forma que estes dois algoritmos operam. A indução de regras busca ir de baixo para cima e coleta todos os padrões possíveis que são interessantes e só depois usa esse padrões para alvo de predição.

As árvores de decisão funcionam a partir de um alvo de predição de cima para baixo, a qual é conhecida como busca gananciosa. Buscando as melhores divisões possíveis no próximo passo (isto é, gananciosamente, colhe o melhor, sem observar qualquer outro aspecto além do próximo passo). Embora o algoritmo ganancioso possa fazer escolha em níveis mais altos da árvore que

são menos otimizados que os níveis mais baixos da árvore, ele é muito bom em comprimir efetivamente quaisquer relações entre preditores e a predição. O sistema de indução de regras, por outro lado, retém todos os padrões possíveis mesmo se eles são redundantes ou não ajudam na exatidão preditiva. Por exemplo, considerando-se que em um sistema de indução de regras, se houverem duas colunas de dados que são altamente correlacionadas ou na verdade apenas simples transformações uma das outras. elas resultariam em duas regras, ao passo que em uma árvore de decisão um preditor seria escolhido e então uma segunda seria redundante, e não seria escolhida de novo. Um exemplo poderia ser os dois encargos anuais e encargos anuais médios (os encargos mensais médios sendo os encargos anuais divididos por 12). Se o montante de encargos fosse preditivo então a árvore de decisão escolheria um dos preditores e o usaria para um ponto de divisão na árvore. A árvore de decisão efetivamente teria comprimido o valor preditivo do preditor e então movido este para o próximo. Em um sistema de indução de regra por outro lado, seriam criadas duas regras, talvez como :

- Se os encargos anuais >12.000 então o padrão seria igual a exatidão verdadeira de 90%.
- Se os encargos mensais médios fossem > 1.000 então o padrão seria igual a uma exatidão verdadeira de 90%.

Constata-se um caso extremo em que dois preditores foram exatamente os mesmos, mas pode haver outros casos menos extremos. Por exemplo, a altura poderia ser usada em vez do tamanho do sapato na árvore de decisão, ao passo que ambos seriam apresentados como regras em um sistema de indução de regras.

Nenhuma técnica é melhor que a outra, embora tendo uma variedade de regras e preditores, estes auxiliam na predição quando houverem valores faltosos. Por exemplo, se a árvore de decisão escolhesse altura como um ponto de divisão, mas se aquele preditor não fosse capturado no registro (valor nulo), mas o tamanho do sapato fosse, o sistema de indução de regras ainda teria uma regra de encaixe para capturar este registro. As árvores de decisão possuem fórmulas para superar esta dificuldade, mantendo sub-rogação em

cada ponto da divisão que implicasse na divisão mais próxima, como faz o preditor escolhido. Neste caso o tamanho do sapato poderia ter sido mantido como um sub-rogação para a altura neste galho particular da árvore. Outra coisa que as árvores de decisão e sistemas de indução de regras tem em comum é o fato de que ambos precisam descobrir meios de combinar e simplificar regras. Numa árvore de decisão isto pode ser tão simples como reconhecer se uma divisão mais baixa sobre um preditor é mais coagido do que uma divisão sobre o mesmo preditor mais acima da árvore. Ambos não precisam ser fornecidos para o usuário, apenas aquele que é mais restritivo. Por exemplo, se a primeira divisão da árvore é a idade menor ou igual a 50 anos e a divisão mais baixa para a folha dada é idade maior ou igual a 30 anos , então apenas a última restrição precisa ser capturada na regra para aquela folha.

As regras do sistema de indução de regras são geralmente criadas tomando-se uma simples regra de alto nível e adicionando novas restrições para ela. Até que a cobertura fique tão pequena quanto a de menor significância. Isto significa que as regras na verdade tem famílias, o que é chamado de cones de especialização, onde uma regra mais geral pode ser a mãe de muitas regras especializadas. Estes cones então podem ser apresentados ao usuário como visualizações de nível alto das famílias de regras e podem ser visualizados de uma maneira hierárquica para ajudar na compreensão.

4.3 Algoritmo para indução de regras

A forma que as regras são construídas é similar as das árvores de decisão. Exceto por uma diferença maior, em árvores de decisão apenas a restrição possível é adicionada à árvore enquanto que em indução de regras todas as restrições possíveis são adicionadas à regra existente. Há uma variedade de formas de fazer crescer regras, mantendo-se registros de que regras deverão se expandir e como cortar e organizar as regras, uma vez que

elas são criadas. Em geral as restrições na forma de novas conjunções são adicionadas a regra tanto como valores específicos para preditores categóricos de baixa cardinalidade ou como restrições baseadas em intervalos particulares para um preditor ordenado continuamente. Embora os algoritmos sejam avaliados eles tem os passos principais em comum :

1. Pré-processamento de dados para que cada preditor tenha intervalo bem definidos em vez de valores contínuos;
2. Regras iniciais geradas a partir dos dados de apenas uma restrição;
3. A partir dos registros, gerar regras que tenham uma restrição adicional a partir das regras dadas;
4. Manter o grupo de regras que são bons candidatos para ter restrições acrescentadas;
5. Continuar a acrescentar restrições nas regras ate que os critérios de parada tenham sido encontrados para todas as regras;
6. Organizar as regras com base na sua utilidade (ou seja, exatidão, cobertura, apoio a significância, simplicidade e inovação) para o usuário final;

No primeiro passo os dados são pré processados para que tanto os preditores contínuos quanto os categóricos tenham um número pequeno de valores diferentes, esses valores são então as restrições que podem ser acrescentadas a regra uma por vez para criar regras mais complexas. Esse passo também é usado em algoritmo de árvore de decisão quando a pré-seleção do preditor precisar ser feita. A direção das regras iniciais é realizada simplesmente pegando-se cada valor de cada preditor em cada registro e emparelhando-o com todos os outros pares de valores do preditor. Estes pares de valores representam os primeiros passos das regras if-then simples; a partir dessas regras é possível definir que regras são boas candidatas para expansão dos novas restrições. Geralmente estas são as regras que passam algum limiar mínimo de exatidão e cobertura; por exemplo, se uma regra tem uma exatidão muito baixa (a compra de maçã prediz compra de biscoito recheado 0.1 % das vezes) é muito menos provável que a regra se torna muito

mais exata com as restrições adicionais (embora isto seja possível). Um início ainda melhor para ser empregado para determinar que regras deveriam ser consideradas e encerradas e quais deveriam continuar sendo expandidas é ver a cobertura e o suporte. Regras com restrições adicionais no seu melhor não afetam a cobertura e o suporte de uma determinada regra, mas muito provavelmente elas diminuirão significativamente estas variáveis. As restrições adicionais nunca podem aumentar a cobertura ou suporte para uma regra. Se uma regra tem uma cobertura bastante baixa, ela não pode fazer sentido para expandir a regra, uma vez que haverá alguma cobertura mínima depois da qual a regra não terá nenhum valor, mesmo se ela possuir uma exatidão alta (por exemplo, se a regra nunca é usada então não importa qual é a exatidão dela). Quando o suporte para a regra cresce pouco, há poucos registros históricos para substanciar a regra e ela por si própria pode ser inteiramente impulsionada e um artefato de ruído aleatório ou variações estatísticas devido a mostra pequena. Como um exemplo, as regras iniciais seriam criadas entre duas colunas do banco de dados, “receita inicial, olhos e gêneros”. Serão criadas regras para a cor dos olhos e para a receita inicial no antecedente e gênero no conseqüente. Uma vez que existam três valores diferentes para a receita inicial, 3 para cor dos olhos e 2 valores diferentes para gênero haverá uma possibilidade de 18 regras diferentes (3 x 3 x 2) com duas restrições no antecedente. Isto representaria todas as combinações diferentes dos valores de preditores e contaria corretamente para o fato de que mesmo se a ordem das restrições da regra fossem mudados, a regra ainda seria a mesma e contaria apenas uma vez. Por exemplo :

- Se for de renda media e olhos castanhos então mulher
- Se olhos castanhos e renda média então mulher

São equivalentes mesmo que a ordem das restrições sejam mudadas no antecedente. A tabela 4.5 mostra o número de regras diferentes que poderiam ser criadas com duas restrições no antecedente a partir dos dados históricos da tabela 4.6. O número de regras é apenas 8 em vez de 18 regras possíveis que foram preditas, a razão para isto é que muitas das regras possíveis nunca ocorreram no banco de dados. Em um banco de dados do mundo real, poderia

haver centenas de preditores cada um com 100 valores diferentes, isto resultaria em aproximadamente 10 bilhões (100 x 100 x 100 x 100) de regras possíveis, se apenas quatro dos preditores fossem usados para o antecedente. A geração de regra desta forma é dupla:

- número de regras possíveis diferentes podem rapidamente crescer até se tornarem enormes;
-
- número de regras atuais serão muito menores, uma vez que muitas das regras possíveis não ocorrem nos dados;

Depois dessa regras serem geradas, é possível serem geradas outras restrições adicionando-se outros preditores. Uma vez que estas regras venham de um exemplo de um banco de dados de 10 registros, o suporte para cada um deles é muito pequeno. Um ou dois registros, o que está longe de estar abaixo do considerado estatisticamente significante. Se no entanto as regras tiverem um suporte maior, elas poderiam ser expandidas com restrições adicionais a partir de outros preditores. No limite, se o algoritmo não apresentar problema para baixo suporte, o sistema poderia criar regras que contivessem cada valor para cada preditor e então reduzir o espaço que apenas um registro do histórico do banco de dados seria capturado para cada regra. A exatidão seria de 100% mas o sistema sofreria superajuste e produziria regras que não o generalizariam bem. Os inícios requeridos para o suporte e cobertura são de fato tentativas de limitar superajustes na geração de regras e estas com os melhores suportes seriam mais provavelmente as que generalizariam novas situações.

Tabela 4.5 – Dados Históricos

Id	Nome	Predição	Idade	Orçamento	Receita	Olhos	Gênero
1	Amy	No	62	0,00	Media	Castanhos	F
2	Al	No	53	1.800,00	Media	Verdes	M
3	Betty	No	47	16.543,00	Alta	Castanhos	F
4	Bob	Yes	32	45,00	Media	Verdes	M
5	Carla	Yes	21	2.300,00	Alta	Azuis	F
6	Carl	No	27	5.400,00	Alta	Castanhos	M
7	Donna	Yes	50	165,00	Baixa	Azuis	F
8	Don	Yes	46	0,00	Alta	Azuis	M
9	Edna	Yes	27	500,00	Baixa	Azuis	F
10	Ed	No	68	1.200,00	Baixa	azuis	M

Tabela 4.6 – Regras geradas com duas restrições no antecedente e uma no conseqüente

Regra	Antecedente	Antecedente	Conseqüente	Suporte
1	Baixa renda	Olhos azuis	Mulher	2
2	Baixa renda	Olhos azuis	Homem	1
3	Renda média	Olhos castanhos	Mulher	1
4	Renda média	Olhos verdes	Homem	2
5	Renda alta	Olhos castanhos	Mulher	1
6	Renda alta	Olhos castanhos	Homem	1
7	Renda alta	Olhos azuis	Mulher	1
8	Renda alta	Olhos azuis	Homem	1

4.3.1 Algoritmo de força bruta

O algoritmo descrito acima prossegue acrescentando restrições para regras existentes e então medindo sua exatidão, cobertura e suporte. Há um *host* (provedor) de métodos heurísticos pelos quais essas adições de novas restrições podem ser desempenhadas. Mas como a velocidade dos computadores é rápida e o espaço em disco se tornou mais barato, os métodos de força bruta estão se tornando cada vez mais apelativos. O algoritmo mais simples de força bruta seria:

1. gerar todos os pares preditor/valor para cada registro como primeiro grupo de regras;
2. contar o numero de ocorrência para cada regra e o antecedente por si próprio;

3. calcular a exatidão, cobertura, suporte e eliminar aquelas regras que não passam do mínimo requerido;
4. para cada registro, ver que regras são aplicas e acrescentar uma restrição adicional, um valor preditor para a regra a partir do registro;
5. voltar ao passo 1 até que as regras sejam tão complexas ou nenhuma regra passe da cobertura mínima ou dos limites do suporte;.

Este algoritmo pode parecer computacionalmente muito caro, mas pode ser eficientemente implantado através de um algoritmo classificador ou muito subdividido. Isto é eficiente quando realizado em um computador paralelo e faz uso de processamento paralelo. O algoritmo é realizado em paralelo, primeiro produzindo todas as regras para um dado registro, baseado em regras prévias (simplesmente passa por cada regra associada com o registro dado e acrescenta uma restrição a partir dos valores preditores disponíveis). Toma a nova lista de regras possíveis e as etiqueta (rotula) com o numero de registros do qual ela foi gerada, então, cataloga todas essas regras formadas recentemente, para que as regras com os mesmos precedentes e antecedentes exatos estejam próximos umas das outras. Por conta de cada nova regra a partir de um registro dado estar próxima da mesma regra gerada a partir de um outro registro (devido a ordem classificada), todos os valores necessários para serem contados para computar a exatidão, cobertura e suporte podem ser facilmente computados através de uma soma executada (algo que é eficientemente executado em computadores paralelos massivamente). Os limiares podem então ser aplicados a essas novas regras e aquele passo pode então ser classificado de novo de acordo com o número de registros do qual eles vieram e quando eles forem locais para aquele registro o próximo conjunto de regras pode ser gerado. Sendo assim, qualquer sistema que tenha otimizado rotina de classificação pode desempenhar esta geração de regras muito rapidamente.

4.4 Medidas Data Mining para Indução de regras

O sistema de indução de regras tem se mostrado relativamente fácil de entender e é de grande utilidade tanto para descoberta quanto predição. Por causa dos sistemas extraírem todos os padrões interessantes, não são frágeis nem sensitivos a valores faltosos ou dados com ruídos. Para sistemas baseadas em regras há muitas regras contribuindo para a predição final ou seja, uma única regra ou um passo errado dado por um algoritmo ganancioso não causaria uma predição incorreta.

Os sistemas de indução de regras são bem adequados para uma variedade de tarefas e quase tão bem automatizados quanto alguns dos mais avançados algoritmos de árvore de decisão como pode ser visto na tabela 4.7[BERS97]

Tabela 4.7 – Medidas Data Mining para Indução de Regras

Medidas DM	Descrição
Exatidão	Quando sistemas de indução de regras são usados para predição suas exatidões preditivas podem estar entre as melhores, entretanto pelo acúmulo de evidências de encaixe múltiplo é ainda mais arte do que ciência e a falta de um processo para o cobrimento de validade cruzada, os sistemas podem oferecer resultados surpreendentes se eles não forem cuidadosamente dispostos.
Clareza	Sistemas baseados em regras são relativamente compreensíveis para as regras próprias mas o número completo de regras podem dominar totalmente o usuário, a clareza das predições pode ser perdida desde que o acúmulo de evidências utilize muitas regras, todas as quais influenciam a determinação da predição.
Dimensão	Por causa delas construírem regras independentemente, os sistemas de indução de regras podem acomodar espaços altamente dimensionais e dimensões com alta cardinalidade. Embora muitas dimensões retardem o processo, o aumento no tempo é geralmente linear com o numero do valor do preditor, que pode ser experimentado nas regras.
Dados Sujos	As regras funcionam bem em superar os valores faltosos e dados sujos em geral.
Dados Puros	Estes sistemas podem requerer algum pré-processamento de dados numéricos em partes ou intervalos inicialmente, mas não são piores que algoritmos de árvore de decisão e são melhores do que alguns algoritmos de redes neurais.

Medidas DM	Descrição
RDBMS	Os sistemas de indução de regras geralmente utilizam algoritmos que são difíceis de serem implementados diretamente dentro de um RDBMS sem uma extração da base de dados. As regras por si só podem ser expostas contra um RDBMS através de SQL, mas aquelas técnicas que usam a evidências acumuladas para predição podem ser mais difíceis e menos eficientes de serem implementadas.
Escalabilidade	Os sistemas de indução de regras escalam bem com números maiores de registros e dimensionalidade. Eles também mapeiam bem para arquitetura de computadores paralelos.
Velocidade	Estes sistemas podem ser lentos, em geral isto se dá por conta de estarem extraindo todos os padrões possíveis do banco de dados. Embora, em computadores paralelos, A aplicação das regras pode ser muito rápida.
Validade	Estes sistemas raramente têm validação cruzada embutida ou validação de conjuntos de testes. Alguns dos mais sofisticados sistemas usam testes estatísticos e ajustes para análises ou avaliações repetidas.

CAPÍTULO V

DATA MINING UMA ABORDAGEM UTILIZADA PARA DESCOBERTA DE CONHECIMENTO

Quando o assunto é Sistema de Apoio a Decisão – SAD, ou outro processo tecnológico que vise o mapeamento de informações estratégicas para apoio à decisão, pensa-se sempre nos seguimentos: MARKETING, CONTROLE INDUSTRIAL, FINANÇAS, VENDAS, ENGENHARIA e MANUFATURA. O foco é sempre minimizar custos e aumentar o lucro gerando benefícios em produtos e prestação de serviço. Em um mercado altamente competitivo em que a informação é o maior bem e este é disputado por todos, há na realidade um grande nicho sendo explorado.

Existe, entretanto um grande número de instituições públicas permanentes com compromissos e missões bem definidos e que respondem com seus serviços a milhares de pessoas. Suas informações, embora não gerem lucro no sentido de retorno financeiro, geram bem estar e benefícios geralmente que não podem ser quantificados e justificados financeiramente.

Com o aumento da demanda por ferramentas *Data Mining*, os preços destas tem se tornados mais atraentes para aqueles que querem se beneficiar dos grandes acervos de dados colhidos ao longo dos anos. Os algoritmos implementados pelas ferramentas *Data Mining* exigem cada vez mais poder de processamento, pois, realizam milhares de comparação na busca de padrões, necessitam tirar proveito do recursos do hardware (o preço tem se tornado mais acessível) para melhorar o tempo de resposta. Isto tem permitido que instituições repensem suas aplicações e busquem dar valor ao grande acervo legado por aplicações transacionais.

De posse de máquinas e programas cada vez mais sofisticados, profissionais tomadores de decisões tais como executivos, diretores e analistas, exigem dos sistemas de suporte a decisão mais recursos. Alguns aspectos que tem deixado esses profissionais frustrados são :

- Trabalho complexo de coleta de dados que são introduzidos e armazenados em lugares diferentes;
- Modelagem de dados inapropriada para análise dos assuntos da instituição;
- Aplicações operacionais desenvolvidas interna e externamente gerando redundância, inconsistência e insegurança nas informações disponíveis em consultas;
- Baixo desempenho de acesso aos dados. O privilégio é dos dados gerados para aplicações transacionais (*OLTP* – Processo Transacional On Line);
- Inexistência de contexto histórico. Dados históricos indicam tendências, padrões.

5.1 Análise Exploratória e Confirmatória

Descoberta é o resultado esperado num processo de busca em que fica claro a intenção que é encontrar padrões escondidos, sem que sejam conhecidas as idéias ou hipóteses sobre o que os padrões podem ser. A aplicação tem a função de encontrar padrões interessantes, liberando o usuário do esforço de pensar primeiro em perguntas pertinentes. Grandes bases de dados, contém milhares de padrões que dificilmente seriam cobertas por pensamentos do usuário. Do ponto de vista estatístico, há dois tipos de análises de dados:

- Análise confirmatória
- Análise exploratória.

Em análise confirmatória, o usuário tem uma hipótese e confirma-a ou refuta-a por uma linha de raciocínio estatística de inferência. Porém, o gargalo para análise confirmatória é a escassez de hipóteses por parte do analista.

Em análise Exploratória, o usuário deseja achar hipóteses satisfatórias para confirmar ou refutar. Descoberta automática melhora o processo de

análise de dados exploratória, enquanto permite aos analistas inexperientes explorar *datasets* muito grandes, porém de forma mais efetiva.

Cada vez mais, os volumes de informação excedem a capacidade de sua análise pelos métodos tradicionais (planilhas, consultas, gráficos e relatórios sintéticos). Esses métodos não podem ser analisados sob o enfoque do conhecimento.

5.2 Data Warehouse

"Um estudo do International Data Corporation (IDC) em 1996 com companhias dos Estados Unidos, apontava o *Data Warehouse* como solução estratégica para os próximos anos. A média do ROI (retorno do investimento) das companhias que investiram naquele ano em *Data Warehouse* foi de 401%, metade das companhias reportaram o ROI acima de 160%, e 25% tiveram um ROI de mais de 600%"[BERS97].

"*Data Warehouse* suporta processamento informatizado provendo uma plataforma sólida e integrada, com dados históricos dos quais se faz análises. Provê as facilidades para integração em um mundo de sistemas de aplicações não integrados. Organiza e armazena os dados necessários para processamento informatizado e analítico sobre perspectivas históricas ao longo do tempo.[INMO97b]"

Um *Data Warehouse* é um ambiente projetado, estruturado e extensível para análise de dados não-voláteis, logicamente e fisicamente transformados, permitindo acesso a múltiplas fontes via aplicações estruturadas para negócios, demonstrando condições empresariais refletidas em longos períodos de tempo e resumidas para análise rápida.

Um processo de *Data Warehouse* envolve uma quantidade significante de profissionais, que estarão envolvidos ao longo do ciclo de desenvolvimento, segundo William Inmon [INMO98], os profissionais exigidos são : Administrador de *Data Warehouse*, Gerente de Mudanças Organizacionais, Administrador de Banco de Dados, Gerente de Metadados, Analista de necessidades de

negócios, Arquiteto de *Data Warehouse*, Desenvolvedor de Aquisição de Dados, Desenvolvedor de Acessos a dados, Desenvolvedores de Manutenção de *Data Warehouse*, Responsável Executivo de Sistema de Informação e Analista de Qualidade de Dados.

Por *Data Warehouse* integrar os dados de uma organização, o hardware deve ser robusto o suficiente em questões pertinentes a capacidade de armazenamento e a realização de i/o de alta performance.

“As características do Hardware necessário para armazenar um *Data Warehouse* deve prever : redundância, escalabilidade, capacidade de auto recuperação, velocidade de recuperação de falhas, Sistema baseado em multiprocessamento simétrico e uma grande capacidade de armazenamento de dados com alta taxa de transferência. Grandes servidores com tecnologia de processamento paralelo”[DODG98].

A justificativa de custos do *Data Warehouse* não é feita mediante critério previamente estabelecido de retorno sobre o investimento. Para isso os benefícios deveriam ser conhecidos antes da construção do *Data Warehouse* .

Quanto aos benefícios do *Data Warehouse*, estes só são conhecidos após sua construção e implantação. As técnicas clássicas de análise de retorno sobre o investimento não podem ser aplicadas ao ambiente do *Data Warehouse*. Somente após as interações dos Analistas de negócio e da avaliação de quão bem estas informações satisfazem as necessidades de consultas é que o *Data Warehouse* pode ser avaliado para responder a justificativa do investimento.

O propósito da maioria dos *Data Warehouse* é reunir grandes quantias de dados históricos de várias fontes e os usar para apoio à decisão. As atividades executadas em um grande repositório de dados corporativos são normalmente diversas, mas freqüentemente inclui as tarefas distintas como consulta e relatório, análise multidimensional e Data Mining. Estas tarefas são quebradas em grupos de usuários separados, como também processos computacionais distintos. Um *Data Warehouse* é assim o lugar natural para armazenar “espaço de dados.” É onde armazena-se elementos de dados básicos nivelados que são depois analisados para gerar informação.

5.2.1 Data Warehouse - Considerações

Diversos fatores devem ser mensurados na implantação de um *Data Warehouse* em uma instituição pública:

- Na maioria das vezes não há volume de dados (na ordem de TeraBytes) que justifique sua implantação;
- Retorno do investimento obscuro;
- Tempo necessário para o processo de implantação muito longo em relação as mudanças políticas comuns em instituições públicas;
- Investimento em equipamentos/softwares/pessoal muito alto;
- Operacionalização dispendiosa, o que deveria ter previsão orçamentária bem definida, ponto difícil a ser atingida por uma instituição pública;
- Qualificação profissional necessária deficiente no mercado;
- Falta de cultura na construção de sistemas a médio e longo prazo em instituições públicas;

5.3 Data Mart

"*Data Mart* pode ser considerado a extensão do *Data Warehouse*. Ele é direcionado para particionamento de dados e é criado para o uso de grupos de usuários. Seu conjunto de dados é criado para ser desnormalizado, resumido ou agregado" [DODG98].

Os Data Marts geralmente são uma tecnologia agregada ao *Data Warehouse*, que pode ser implantado baseado em duas metodologias : TOP DOWN ou BOTTOM UP, na primeira o *Data Warehouse* é criado e então pelas necessidades específicas dos departamentos (ou grupos de usuários) de uma organização os Data Marts surgem; na Segunda abordagem, por questões de tempo ,custo ou por se justificar a criação de *Data Warehouse* segmentado (o que facilita a assimilação do novo processo por todos da organização), são

criados *Data Marts* com informações departamentais, que depois serão integradas formando o *Data Warehouse*.

Tamanho não determina um *Data Mart*, eles existem desde alguns megabytes até terabytes[INMO98]. O que define um *Data Mart* é o tipo de estrutura de dados ele armazena. Ele geralmente possui informações de um grupo de usuários ou de um departamento, implementado para responder rapidamente a questões específicas, visa a performance, redundância gerenciada, enfatiza categorização e é parcialmente desnormalizado.

5.3.1 Data Mart - Considerações

Data Marts são projetados para dar inicio a um processo de Data Warehousing, que pela dificuldade em justificar os investimentos iniciais, geralmente começam com uma estrutura departamental disposto e estruturado a servir de protótipo para um grande projeto.

Uma questão importante quanto aos *Data Marts* diz respeito as ferramentas utilizadas para transformar dados em informação. A metodologia é baseada em análise confirmatória, o usuário poderá confirmar ou refutar uma hipótese, mas no máximo estará preparado seus dados a responder questões conhecidas. É bem verdade que a infra estrutura requerida já não é a mesma do *Data Warehouse*. Ferramentas OLAP são utilizadas para tirar o maior proveito possível dessa arquitetura.

Grandes volumes de dados criam dificuldades para exploração interativa, pois exigem a presença do usuário, além de que ele tem que fornecer dados úteis. Um outro ponto negativo para a validação é a dificuldade na identificação de todos os padrões possíveis.

5.4 OLAP

OLAP – On Line Analytcal process (Processo Analítico On Line) permite ao usuário verificar uma hipótese visualizando os dados de forma mais natural, usando um modelo multidimensional, de modo que os dados possam ser verificados de forma *drill-down* (aprofundamento da pesquisa) ou por *slice-and-dice* (sumarização). Esta tecnologia pode ser aplicada a *Data Warehouse* e a Data Mart, que são estruturas de dados agregadas por categorias ou dimensões com níveis de detalhamentos bem definidos.

"OLAP é uma categoria de tecnologia de software que permite que analistas, gerentes e executivos obtenham, de maneira rápida, consistente e interativa, acesso a uma variedade de visualizações possíveis de informação que foi transformada de dados puros para refletir a dimensão real do empreendimento do ponto de vista do usuário"[INMO98].

OLAP pode ser empregado em modelagem de verificação, que considera uma hipótese do usuário validando-a contra os dados. O foco está com o usuário que é responsável por formular a hipótese e executar a consulta nos dados e assim afirmar ou negar a hipótese.

Para William Inmon [INMO98] "Há muito o que dizer sobre consultas no nível OLAP :

- É econômico;
- É altamente flexível;
- Permite a personalização de dados necessários por um determinado departamento;
- Permite o isolamento significativo de partes de dados;
- Visão conceitual multidimensional;
- Transparência;
- Acessibilidade;
- Relatórios consistentes".

Por características ligadas aos Sistemas Gerenciadores de Bancos de Dados (SGBDs) utilizados com a tecnologia OLAP, surgiram algumas sub-classes com a finalidade de diferenciação [BERS97]:

- ROLAP – Relational On Line Analytical Processing. Tecnologia OLAP que acessa SGBDs Relacionais;
- MOLAP – Multidimensional On Line Analytical Processing. Tecnologia que acessa SGBDs Multidimensionais, o que permite acesso direto, sem mapeamentos das aplicações OLAP;
- HOLAP – Hybrid On Line Analytical Processing. Tecnologia híbrida, o que permite benefícios quanto a alta performance do MOLAP com a escalabilidade do ROLAP.

Embora importante, pois apresenta as informações em nível de detalhamento ideal, em linguagem natural do negócio analisado, OLAP não mostrará nenhuma informação nova, pois o processo descrito somente recupera registros para verificar ou negar a hipótese do usuário. O processo de busca é interativo, os refinamentos são sempre sugeridos pelo usuário e são interrompidos quando este se satisfaz.

5.5 KDD e Data Mining

Em 1989, foi convencionado o termo KDD (*Knowledge Discovery in Database*) como uma referência geral ao processo de descoberta de conhecimento em Bancos de Dados. É um processo interdisciplinar que envolve áreas da estatística, da inteligência artificial e do aprendizado de máquinas (*Machine Learning*), reconhecimento de padrões e visualização de dados.

Data Mining é uma das fases do processo de descoberta de conhecimento, para o qual têm sido descritos vários métodos. Dentre estes : técnicas estatísticas, visualização, árvore de decisão, *nearest negbohor*, agentes, algoritmos genéticos e redes neurais.

Data Mining compreende os métodos referentes a aplicação dos algoritmos para extração de padrões a partir dos dados, esta pode ser feita pelo usuário, por exemplo só executando consultas, ou pode ser auxiliado por um programa inteligente que automaticamente procura no banco de dados por si só e acha padrões significativos.

Data Mining era visto como um “subconjunto” das atividades associadas com o *Data Warehouse*. Enquanto o ultimo pode ser uma boa fonte para os dados a serem minerados, o primeiro foi reconhecido como uma autentica tarefa com direitos próprios, e já não é considerado como uma colônia do *Data Warehouse*.

Data Mining não só ganhou independência, mas está diretamente e significativamente influenciando o desenho e implementação de grandes *Data Warehouse*. Construir um *Data Warehouse* primeiro, minerar depois, parecia uma regra simples e intuitiva, porém as técnicas de *Data Mining* possuem um conjunto de ferramentas que não necessariamente tem que estar associados a um *Data Warehouse* para serem utilizadas.

Embora *Data Warehouse* e *Data Mining* sejam atividades indubitavelmente relacionadas, o ultimo requer estruturas de dados diferentes, satisfaz a processos computacionais para um grupo diferente de usuários que o *Data Warehouse* típico.

5.6 Etapas do processo KDD

Data Mining é um componente do processo KDD, e é um processo interativo, repetitivo e envolve: Seleção, Pre-processamento, Transformação, mineração de dados e Interpretação e Avaliação, como apresentados na figura 5.2[BERS98].

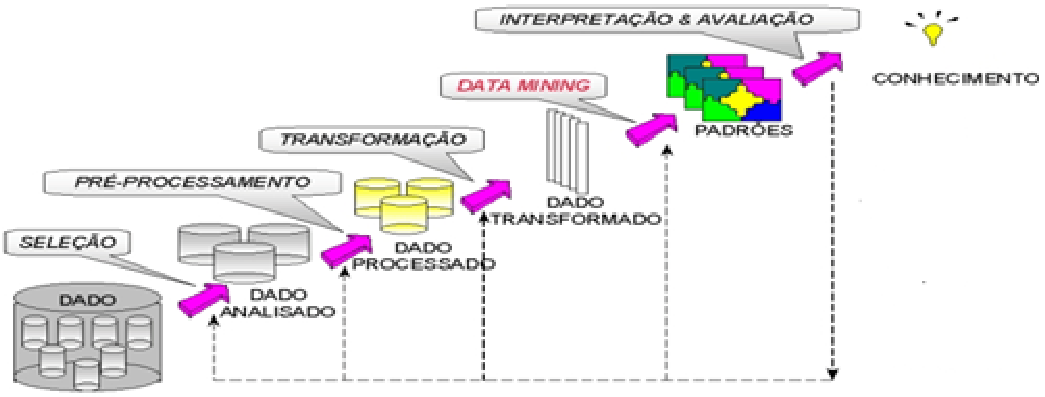


Figura 5.1 – Processo de Descoberta

Existem 10 regras básicas a serem seguidas na criação do modelo a ser minerado [PYLE99] :

1. Selecionar claramente a definição do problema a ser resolvido;
2. Especificar a solução requerida;
3. Definir como a solução especificada será usada;
4. Entender as muitas possibilidades entre o problema e o *dataset*;
5. Nortear o modelo de acordo com o problema;
6. Estipular suposições;
7. Redefinir o modelo interativamente
8. Construir um modelo simples, mas não simplificado;
9. Definir a instabilidade do modelo (áreas críticas);
10. Definir as incertezas do modelo

As três primeiras recomendações estão associadas ao processo do conhecimento do domínio a ser explorado. A quarta recomendação é básica para qualquer atividade que envolva Banco de Dados. A quinta recomenda a escolha de ferramentas que se adeqüem tanto ao modelo construído quanto a possíveis variações; A sexta confirma a necessidade de *brainstorming*. A sétima apresenta o processo como interativo e circular, sendo possível retornar a fases anteriores para correções e ajustes. A oitava recomendação é

fundamental para a compreensão do processo. A nona e a décima recomendação dizem respeito as limitações que devem ser conhecidas para evitar surpresas ou esperar muito mais do que o modelo possa fornecer.

5.6.1 Seleção de Dados

Na fase de seleção dos dados os estudos são para definir o foco a ser Minerado. Responder a questão “Quais são as necessidades específicas do negócio ?”, constitui a chave do sucesso. Os dados de origem são os encontrados no nível estruturado organizacional e nos sistemas transacionais.

Uma vez que o domínio sobre o qual se pretende explorar é conhecido, passa-se a selecionar e coletar os conjuntos de dados necessários para constituírem a base a ser explorada pelas aplicações *Data Mining*.

5.6.2 Pré-processamento

Uma grande quantidade dos dados a serem submetidos ao *Data Mining* podem estar armazenados em tecnologias ultrapassadas e que certamente não retratam o modelo necessário para o processo de exploração.

Nesta fase os dados que farão parte do novo modelo de dados deverão ser migrados. É comum a necessidade de limpeza dos dados obtidos nos sistemas transacionais, por não estarem íntegros ou por não sofrerem nenhum tipo de crítica nos sistemas de armazenamento de origem. Embora muitas técnicas possam tratar dados sujos ou incompletos, quanto melhor a qualidade do dados melhor será o resultado obtido.

5.6.3 Transformação

O processo de transformação é necessário quando os dados de origem estão apresentados em um nível de detalhe não adequado para a mineração. Como exemplo pode-se assumir que não é convencional utilizar a data de nascimento como preditor, pois um algoritmo provavelmente iria segmentar o *dataset* considerando o dia, mês e ano, enquanto que o usual é a criação de agrupamentos por idade ou faixas etárias. Neste caso a data de nascimento deveria ser codificada em idade ou faixa etária.

Transformar dados em informações, significa carregar um nível superior a partir de um nível inferior e requer [INMO98] :

- Sumarização dos dados para categoriza-los em dimensões;
- Um subconjunto de dados baseados nas necessidades de informações, selecionando :
  - Linhas específicas que satisfaçam critérios adequados, ou
  - Apenas as colunas de dados necessárias, ou
  - Uma combinação entre os dois;
- Desnormalização de dados em um número menor de linhas para torná-los “planos” e assim mais fáceis de serem compreendidos;
- Aplicação de regras de negócios para derivar dados que satisfaçam as necessidades de informações específicas;
- Ou uma combinação de qualquer um item mencionado acima.

5.6.4 Data Mining

Esta fase se preocupa com a extração de padrões dos dados. Um padrão pode ser definido como determinado conjunto de fatos (dados), e alguma medida de certeza. Um padrão é uma declaração que descreve relações entre um subconjunto de dados com um grau de certeza.

Existe um grande número de ferramentas para *Data Mining* no mercado implementando predição e descoberta com uma diversidade de algoritmos e técnicas. A avaliação quanto a natureza do negócio, complexidade, volume dos dados, hardware disponível, recursos humanos, suporte e preço, sempre devem ser considerados quando da definição da ferramenta.

Data Mining envolve ainda a determinação de características para serem usadas nos ajustes dos dados, pois o conhecimento é deduzido. A decisão se o modelo produz ou não conhecimento significante parte do julgamento subjetivo humano. Os modelos para *Data Mining* levam em consideração as funções :

- Classificação – Classifica os dados dentro de uma categoria ou várias categorias predefinidas;
- Regressão – Mapeamento dos dados em uma variável de valor real;
- Particionamento – Mapeamento dos dados em classes (agrupamento);
- Sumarização – Mapeamento compacto para um subconjunto de dados;

5.6.4.1 Árvores de decisão

Os modelos construídos pelas árvores de decisões podem ser facilmente visualizados com uma árvore de simples decisões baseadas em preditores familiares ou como um conjunto de regras. O usuário pode na verdade confirmar o modelo de árvore de decisão em mãos ou modifica-lo e dirigi-lo com base em seus próprios conhecimentos. As árvores implementam aprendizagem supervisionada.

Árvores de decisão apresentam uma técnica muito favorável para automatizar a maioria dos dados fornecidos e o processo de modelagem preditiva. Elas são engajadas em soluções automáticas para coisas como o super ajuste preventivo de manuseio de dados fornecidos que muitas outras técnicas deixam como ônus para o usuário.

“Por conta das árvores de decisões trabalharem bem com bancos de dados relacionais, elas fornecem soluções bem integradas com modelos altamente acurados, o que qualifica o ROI empregado nestas ferramentas”[BERS97].

Exemplo de ferramentas para árvore de decisão estão no anexo 6.

5.6.4.2 Indução de Regras

Sistemas de indução de regras tendem a ser altamente automatizados na ordenação de construção e apresentação das regras. O usuário é requerido a passar por muitas regras que são propostas como interessantes para na verdade determinar se a regra é importante.

As regras são geralmente simples e fáceis de entender, embora o porque delas ocorrerem pode não ser facilmente justificável. Também, por conta do grande número de regras que são retornadas ao usuário final, alguma clareza do sistema pode ficar perdida, uma vez que o usuário está envolvido com regras obscuras que não fazem sentido ou regras obvias que já eram conhecidas.

“Embora os sistemas de indução de regras possam ser usados para predição eles são na maioria das vezes usados para aprendizagem supervisionada, para descobrir coisas que não são conhecidas. Mesmo sendo difícil quantificar o retorno do investimento para regras o ROI é bem compreendido”[BERS97].

Exemplos de ferramentas para indução de regras estão no anexo 6.

5.6.5 Escolhendo a técnica Certa

A técnica escolhida para aplicação de *Data Mining* deve considerar : rendimento, aumento da renda, diminuição dos custos ou o retorno do

investimento e muito mais do que encontrar padrões nos dados ele deve ser usado para o sucesso do negócio.

Uma técnica ou um conjunto de técnicas pode ser selecionado para o processo de *Data Mining*. Todas possuem pontos fortes e fracos e geralmente podem ser usadas em conjunto, uma vez que nenhuma consegue atender a todas as necessidades de todas as aplicações.

A escolha da ferramenta ou ferramentas *Data Mining* deve considerar [BERS97] :

EXATIDÃO – A ferramenta *Data Mining* deve produzir um modelo que seja exato o quanto possível, mas reconheça as pequenas melhorias que podem ser implementadas.

EXPLICAÇÃO – A ferramenta *Data Mining* deve explicar como o modelo trabalha para o usuário final, tornando claras a permissão e construção de intuições facilmente testadas. Deve permitir também o cálculo do retorno do investimento (ROI).

INTEGRAÇÃO – A ferramenta *Data Mining* deve permitir integração com o atual processo de negócio, dados e o fluxo de informação da organização.

5.4.2.4.5 Interpretação e Avaliação

Data Mining é uma metodologia para encontrar uma descrição lógica ou matemática, eventualmente de natureza complexa, de padrões e regularidade em um determinado conjunto de dados.

A interpretação do modelo descoberto deve ser realizado, e nesta fase são mensurados o valor dos padrões encontrados, sua relevância. É possível avaliar a necessidade de retorno a um passo anterior, removendo modelos redundantes ou sem significado para o usuário.

As características exatidão, explicação e integração importantes na definição da ferramenta *Data Mining* devem ser quantificados para determinar o resultado do processo de descoberta de conhecimento.

CAPÍTULO VI

EXTRAÇÃO DE CONHECIMENTO DOS DADOS DO MINISTÉRIO PÚBLICO CONTROLE DE INQUÉRITOS POLICIAIS

O Ministério Público do Estado de Rondônia (Organograma no anexo I) conta com sistemas transacionais há mais de 10 anos. Com o advento da microinformática nos anos 80, foi vislumbrada pelos diretores da Instituição a necessidade de informatização dos processos das áreas meio e fim.

Com a constituição de 1988, os Ministérios Públicos Estaduais ganharam a responsabilidade de zelar pelo cidadão e pela sociedade. Em Rondônia essa obrigação veio acompanhada de uma necessidade de construir um sistema de informação que controlasse as distribuições dos inquéritos policiais e todas as ações em que houvesse acompanhamento do Ministério Público. Em 1989 o Controle de Inquéritos Policiais - CIPO, foi desenvolvido e implantado em linguagem de Quarta geração ZIM em ambiente EDIX (*UNIX LIKE*).

Ao longo dos anos o CIPO vem armazenando dados para retratar a atividade dos Membros da instituição dentro de um período determinado (nunca superior a três meses). A Corregedoria-Geral, órgão da alta administração da Instituição faz uso das informações para acompanhar o desempenho dos promotores de Justiça. O CIPO também permite o acompanhamento dos prazos estabelecidos no encaminhamento e recepção de inquéritos policiais entre o Ministério Público e a Secretaria de Segurança Pública (através de suas delegacias) e o Judiciário.

A Figura 6.1 apresenta o modelo lógico do CIPO concebido em 1988 e ainda em uso.

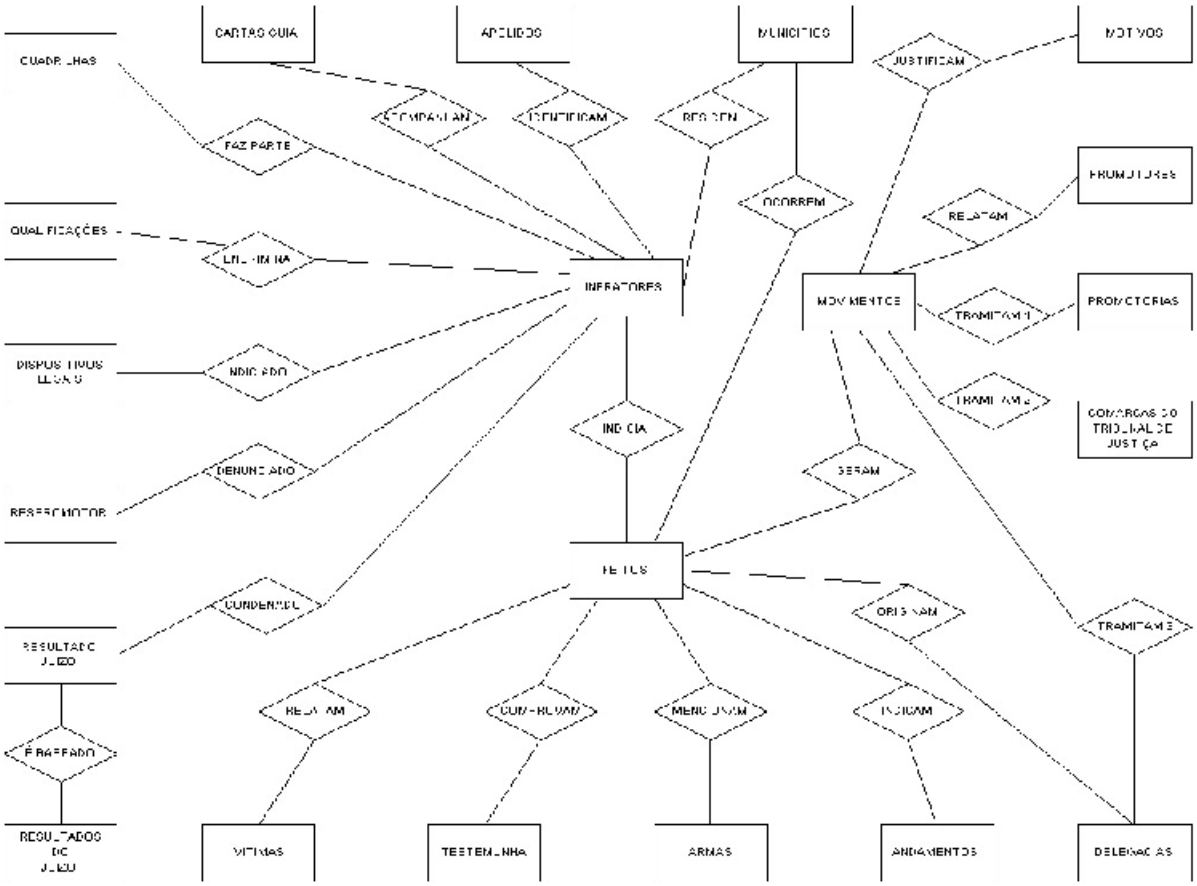


Figura 6.1 – Modelo conceitual do CIPO

O foco de visualização de dados do CIPO está na movimentação dos feitos, muito embora informações referentes aos delitos cometidos, infratores, testemunhas, vítimas entre outras, estejam armazenadas.

Com todas as mudanças ocorridas na legislação ao longo dos últimos dez anos, além da reestruturação constante do Ministério Público para que exerça sua missão de forma adequada, o CIPO permanece como ferramenta de mensuração da produtividade dos Promotores de Justiça.

O Ministério Público de Rondônia conta atualmente com 20 promotorias instaladas em 20 municípios, todas com estrutura básica de informática. Algumas Promotorias então conectadas via rede remota ao CIPO e as demais, realizando o envio de formulários para serem alimentados no banco de dados, gerando um volume considerado de dados com informações de todo o estado.

No anexo II é apresentado um quadro resumo de microcomputadores das Promotorias de Justiça.

6.1 Formulação do Problema

Uma das missões do Ministério público é propor a adoção de medidas que venham beneficiar sempre o coletivo. Para a proposta de novas medidas é necessário o estudo do comportamento dos diversos segmentos sociais e do impacto que tal medidas podem causar. Dentre estas medidas pode-se citar: a construção de presídios de segurança máxima, a construção de penitenciárias, a mudança no sentido de tráfego de automóveis em vias públicas, a intervenção de órgãos governamentais e privados e etc. Embora não seja responsável pela implantação das medidas sugeridas o Ministério Público age como fiscal da lei, buscando sempre o interesse coletivo.

O conhecimento das infrações cometidas, ou o mapeamento da situação das ações que envolvam o Ministério Público, tem sido alvo de discussões permanentes. Há necessidade de respaldar a sociedade com informações que possam melhorar a confiabilidade e a qualidade da justiça brasileira e consequentemente a melhoria da qualidade de vida.

O Colégio Nacional de Corregedores Gerais de Justiça, composto por todos os Corregedores Gerais dos Ministérios Públicos estaduais, vem definindo um modelo, com o intuito de mapear os dados estatísticos e sociais levantados nos mais diversos pontos do país e com isso traçar políticas e planos de atuação. Estes dados são fruto do trabalho desenvolvido pelo Ministério Público nas áreas de atuações :

- Criminal
- Cível
- Infância e Juventude
- Controle Externo da Atividade Policial
- Patrimônio Público

- Meio Ambiente
- Cidadania
- Defesa das Vítimas
- Consumidor
- Atendimento ao Público
- Recursos Constitucionais

Como resultado dos encontros chegaram a conclusão que um modelo satisfatório seria aquele que reunisse informações especializadas para as áreas acima citadas. A idéia básica é reunir informações acerca das infrações cometidas. Um primeiro esboço classificou as informações a serem formatadas como :

Área Criminal :

Crimes praticados por prefeitos

Código Penal

Tóxicos

Licitações

Parcelamento do solo Urbano

Crimes Ambientais

Crimes praticados por Autoridades Policiais

Código Penal

Código Penal Militar

Abuso de autoridade

Tortura

Crimes contra a ordem tributária

Área Cível

Ações propostas de Alimentos

Ações propostas de Investigação de Paternidade

Ações propostas de interdição

Ações propostas de Intervenção do Estado nos Municípios

Ações propostas de Inconstitucionalidade

Área da Infância e Juventude

- Ato infracional
- Medidas sócio-educativas
- Remissões
- Interesse Difusos e Coletivos
- Conselhos tutelares da Criança e do Adolescente
- Programas de Atendimentos
- Colocação em família substituta
- Infrações Administrativas

Área de Controle Externo da Atividade Policial

- Requisições
- Improbidade de Autoridades Policiais
- Inspeção em estabelecimentos policiais e carcerários

Área do Patrimônio Público e Social

- Ação civil Pública
- Penalidades Aplicadas

Área do Meio Ambiente

- Poluição Sonora
- Recursos Hídricos
- Tratamento do Esgoto
- Poluição industrial
- Combate ao depósito de lixo a céu aberto
- Extração Irregular de Minérios
- Urbanismo
- Patrimônio Histórico Cultural
- Fauna
- Flora
- Reserva Florestal
- Agrotóxicos

Área da Cidadania

Defesa de Interesses Difusos e Coletivos

Saúde

Educação

Transporte Coletivo

Deficientes Físicos/Idosos/indígenas

Área das Vítimas

Ações “EX DELICTO”

Danos advindos de erros médicos

Danos adivindos de acidente de transito

Danos adivindos de Violência Policial

Danos adivindos de Delitos de imprensa

Danos adivindos de Racismo

Danos adivindos de outros Delitos

Ações Acidentarias

Área do Consumidor

Interesses Difusos e Coletivos

Atendimento ao Público

Segurança

Transporte

Área dos Recursos Constitucionais

Recursos Especiais

Recursos Extraordinários

A mensuração destas informações passa necessariamente pela implantação de uma grande infra-estrutura que armazene os dados a nível nacional.

Com uma proposta de mensuração das ações do Ministério Público definida, com o conhecimento do negócio e suas implicações, seria possível construir um modelo de verificação para extração de informações baseadas no esboço apresentado acima, considerando-se a base de dados populada.

Como exemplo prático da necessidade da mensuração das informações da justiça, o Supremo Tribunal Federal (STF), anunciou que todo o Poder Judiciário (Justiça comum, do Trabalho, Militar, Eleitoral) será informatizado no país em todas as instâncias. Existe uma verba de R$ 50 milhões, prevista para o atingimento desta meta. Será criada uma rede mundial informatizada entre poderes judiciários de diversos países. Há interesse do Bird (Banco Mundial) em financiar a informatização total da Justiça brasileira(Folha on-line 24/07/00 em www.uol.com.br/folha).

Deve ser considerado que o esboço apresentado não está fundamentado em dados colhidos por sistemas transacionais em um período determinado. Portanto o modelo é fruto de abstrações realizadas com base na *expertise* dos autores. Entende-se que se o modelo fosse construído a partir de um grande banco de dados das ações do Ministério Público nacional, haveria uma grande chance dos dados não confirmarem esse modelo, haja vista que humanos tem limitações em abstrair grandes quantidades de variáveis com infinitas condições[BERS97].

A busca de conhecimento sendo realizada de forma automática é um dos caminhos para a avaliação do modelo definido.

Como o CIPO reúne dados dos últimos 12 anos, pensou-se na viabilidade de construir um *Data Warehouse* para armazenar informações captadas ao longo desse período, e a partir deste seriam criadas as consultas para dar suporte as informações do Ministério Público de Rondônia, servindo ainda como estudo de caso aos demais Ministérios Públicos Estaduais.

As exigências para um processo de *Data Mining* são muito mais próximas à realidade do Ministério Público do Estado de Rondônia. A tarefa de reestruturar os dados do CIPO para permitir análise das ações, vai identificar padrões de dados que provavelmente não foram considerados no modelo proposto pelo Colégio de Corregedores Gerais de Justiça.

Para efeito desta pesquisa e pela limitação apresentada as ferramentas *Data Mining* empregadas nos dados do CIPO, implementam algoritmos de árvore de decisão e indução de regras.

Busca-se confirmar a praticidade do uso desse tipo de ferramenta, adicionada a outras tecnologias empregadas com sucesso nas instituições públicas. Pretende-se apresentar aos Ministérios Públicos Estaduais um exemplo de mensuração das ações desenvolvidas no Ministério Público de Rondônia e que este venha a ser seguido para a criação de um grande banco de dados das ações nacionais. O fato gerador dessa necessidade está centrado na falta de informações sobre os delitos cometidos contra o cidadão e a sociedade. Estima-se que a criminalidade está crescendo, mas quantos por centos ao ano ? em que regiões são mais incidentes ?

O problema a ser resolvido considera o grande volume de dados armazenados por mais de uma década Pelo Ministério Público do Estado de Rondônia em seus sistemas transacionais. Até então os dados foram usados para responder questões relativas a períodos de tempo muito pequeno (não superior a três meses) sem, contudo estes dados terem sido transformados em informações que servissem de base para a tomada de decisão estratégica organizacional.

A aplicação do modelo proposto leva em consideração os dados captados no Sistema de Controle de Inquéritos Policiais – CIPO. O escopo principal está nas ações acompanhadas pelo Ministério Público no período de janeiro de 1995 a dezembro de 1999, considerando-se infrações referentes a Homicídios (Artigo 121 do CPB), Tóxicos (Lei 6368/76) e estupros (Artigo 213 do CPB), ocorridos no município de Porto Velho, estado de Rondônia.

6.2 Extração primária dos dados

O primeiro passo para implantação do processo *Data Mining*, considera a necessidade de extração primária dos dados, ou seja, mover os dados do ambiente operacional para um SGBD.

Os dados do CIPO encontra-se armazenados em um banco de dados proprietário(ZIM). A leitura direta dos dados por utilitários é impraticável, dado o fato que não existem serviços de conexão aos dados como ODBC ou outra

fonte que permita o uso de utilitários de extração. (Anexo III apresenta as estruturas internas do CIPO em ZIM)

Como solução para o problema, considerou-se os seguintes passos :

- Utilização de comandos ZIM para exportar os dados para arquivos em formato textual, sem formatação e sem *header*;
- Promoção de uma limpeza preliminar de caracteres inseridos sem crítica prévia;
- Carga do arquivo textual para uma tabela em um SGBD;

6.2.1 Utilização de comandos ZIM

ZIM é uma Linguagem de 4ª Geração dos anos 80, que possui cinco componentes principais : Os comandos – elementos fundamentais da linguagem; Os operadores – Palavras chave ZIM; As funções – Palavras chave que realizam tarefas; As variáveis – Representam constantes e os Utilitários do Sistema Operacional – Programas que rodam fora do ZIM e que realizam tarefas como a criação de novas bases de dados.

Todos os componentes citados acima trabalham em função de um dicionário de dados que mantém o Banco de Dados e todos os objetos necessários para uma aplicação.

No anexo IV estão descritos todos os comandos ZIM utilizados para a migração dos dados.

6.2.2 Limpeza Prévia

Os arquivos texto gerados através de comandos ZIM, criaram linhas com caracteres sem significado. Um fator importante de ser lembrado é que, sistemas alimentados por terminais de vídeo, podem ser executados em microcomputadores, desde que esses façam uso de um programa emulador de

terminal. Por ZIM não ser um SGBD não há a possibilidade de criação de *constraints* (restrições) o que deixa a cargo do programador da aplicação a restrição quanto aos dados digitados. Como os teclados dos microcomputadores possuem caracteres estendidos e o sistema não foi concebido para tratar tais erros, um grande número de caracteres estranhos foram encontrados nos dados.

A limpeza prévia não leva em consideração a verificação da integridade dos dados, uma vez que os dados não estão portados para um SGBD. Somente após o banco de dados ser criado é possível validar a integridade referencial e validar e integridade semântica quanto aos dados terem sido inseridos corretamente.

Uma das características dos sistemas baseados em arquivos é que todo o controle dos dados é de responsabilidade do programador da aplicação. Humanos erram e possuem limitações para analisar inúmeras possibilidades de ocorrência de eventos que podem provocar a inconsistência de dados.

6.2.3 Carga dos dados em um SGBD

"Uma vez que o modelo relacional é baseado em princípios matemáticos e lógicos, existem diversos SGBD que oferecem ótimas soluções para uma grande variedade de aplicações comerciais e científicas. Os requisitos exigidos de um SGBD estão fundamentados na necessidade deste suportar eficientemente um grande número de acessos simultâneos de leitura e gravação[BERS97]".

Na fase de extração primária dos dados, há uma preocupação em retratar os dados no mesmo esquema adotado no banco de dados de origem. Necessariamente nesta primeira fase do processo de *Data Mining*, não há necessidade da adoção de um grande SGBD, uma ferramenta como o Microsoft Access pode ser adotado tranqüilamente.

O Ministério Público do estado de Rondônia mantém aplicações residentes em três SGBRs que estão instalados em plataforma UNIXWARE, WINDOWS NT e WINDOWS 98 , a saber :

- *Sybase Adaptive Server Anywhere Database Version 6.0;*
- *Microsoft SQL Server 6.5*
- *ORACLE 8i*

Todos os três SGBDs instalados em seus respectivos ambientes operacionais atendem os requisitos para hospedagem do esquema de dados necessários para o *Data Mining*.

A carga dos dados foi gerada em uma aplicação Power Builder que realizava a leitura das linhas de um arquivo texto, em seguida separava a linha em valores para colunas efetuando as conversões de tipos de dados e depois estas eram gravadas no banco de dados.

6.3 Codificação

As informações usadas para exploração de dados, necessitam estar em um formato diferente do encontrado no ambiente operacional. Os dados precisam sofrer uma codificação que os enriqueçam e os preparem para o processo de descoberta.

A definição dos níveis de granularidade dos dados para o ambiente projetado dos dados é vital nesta etapa. Os dados serão explorados categorizados por assunto e de forma histórica. Para atender estas necessidades o modelo deve considerar :

- Os principais assuntos a serem minerados;
- Os relacionamentos entre os assuntos;
- Os atributos dos assuntos principais ;
- As chaves dos assuntos principais;

- Criação de atributos temporais (data);

Os dados devem ser estruturados para satisfazer as necessidades de informação de um grupo distinto de usuários, identificados por uma função de negócio específico. Há um nível razoável de resumo.

Desnormalizar dados é outra técnica utilizada no processo de codificação dos dados, um dos principais fatores é que com o passar do tempo informações de referência podem ser alteradas, neste caso, guarda-se a descrição. Dados desnormalizados são mais fáceis de serem entendidos pelos usuários, pois estes não precisarão juntar tabelas para entender o significado de um punhado de dados relacionados. Porém, atributos de referência que não representam um percentual significativo de entradas na tabela principal, não tem razão de serem desnormalizados, uma vez que vão gerar uma grande quantidade de ocorrências nulas.

"Antes de começar a codificação, deve-se estudar técnicas de modelagem dimensional, planejar a desnormalização sempre que possível, reduzir o número de tabelas por 10 ou mais, e pré-agregar para criar dados de negócio com uma granularidade adequada em cada dimensão"[INMO97].

O processo de codificação dos dados do CIPO para um SGBD foi realizado e gerou inúmeros problemas de integridade. Pode-se afirmar que nas principais tabelas foram encontradas duplicações de linhas(tuplas). Um erro grave uma vez que a integridade da chave primária não foi respeitada. Duas linhas(em alguns casos até mais de duas), estavam escritas no arquivo texto, proveniente do banco de dados ZIM.

Com a definição de chaves primárias e chaves estrangeiras nas tabelas do Banco de dados que receberia as tabelas migradas do ZIM, muitos problemas ocorreram, uma vez que o banco de dados de migração não permitia a inserção de linhas que quebrassem a integridade definida.

O CIPO desenvolvido em ZIM permitiu na captação de dados duplicação de linhas (tuplas) no processo transacional, o que causou grande quantidade de erros no programa de carga do banco relacional. Como solução, programou-se o aplicativo de carga para ignora-las.

A tabela 6.1 apresenta a quantidade de registros povoados em cada tabela do SGBD.

Tabela 6.1 – Quantitativo de registros povoados

Tabela	Registros	Tabela	Registros
APELIDO	662	ARMAS	852
DISPOLEG	3.203	FEITOS	2.686
INFRATOR	3.478	PESFISICA	1.592
QUALIFICAÇÃO	6.250	RESJUIZO	3053
TABCLASSE	8	TABCOR	5
TABDELEGACIA	32	TABESTADOCIVIL	6
TABGRUPO	30	TABRESJUIZO	7
TESTEMUNHAS	2.509	VITIMA	2.965

O primeiro modelo de dados derivado do CIPO é apresentado na figura 6.2. O modelo retrata a estrutura encontrada no ambiente de origem (ZIM), sem terem sido exportados os dados referentes a movimentação de feitos.

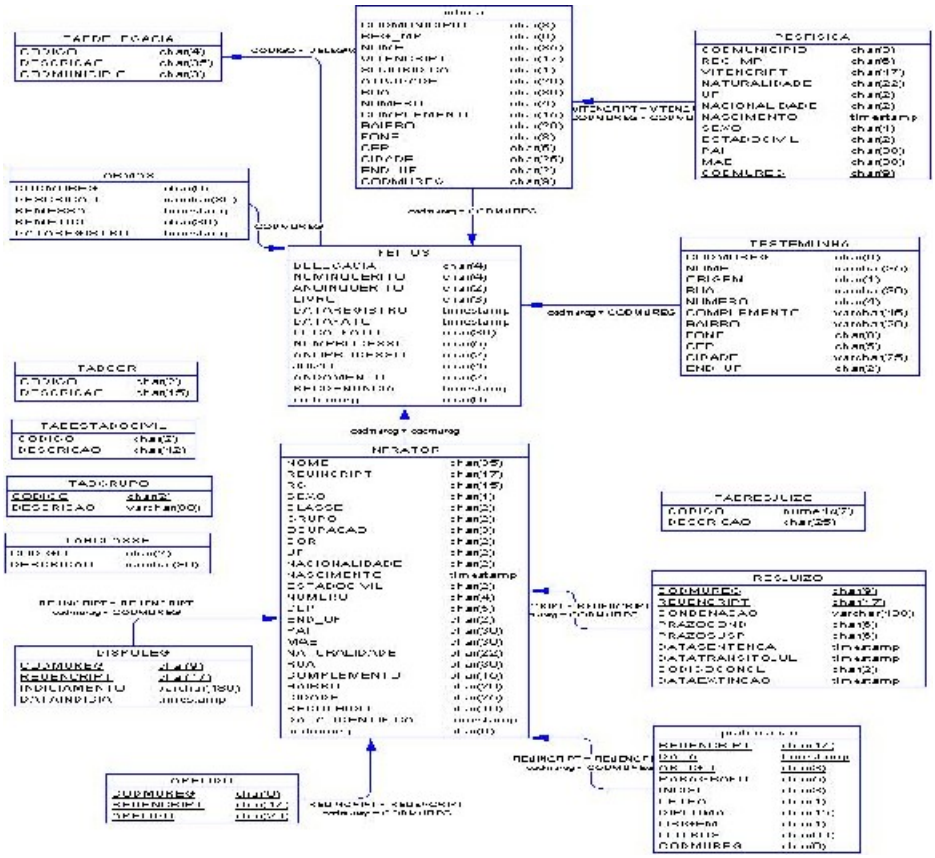


Figura 6.2 – Modelo gerador a partir do CIPO

A figura 6.1 apresentou o modelo lógico contendo tabelas referentes a tramitação (movimentação) dos feitos entre o Ministério Público e a Secretaria de Segurança Pública e o Judiciário. Contudo, embora estas tabelas possuam a maior quantidade de registros, uma vez que um feito pode sofrer centenas de movimentações, seus dados são analisados constantemente na Instituição, para mensurar a produtividade dos Membros (Promotor de Justiça).

Nesta fase foram definidas as agregações dos dados e o nível de granularidade para o processo de *Data Mining*. Baseado no processo de desnormalização eliminou-se as tabelas de referência.

Com o auxílio de comandos SQL foram geradas novas tabelas derivadas do modelo lógico apresentado na figura 6.2. Estas tabelas resumem assuntos relevantes para o Ministério Público, dos quais pouco se conhece e que praticamente não foram utilizadas para fortalecer uma decisão na Instituição.

A figura 6.3 apresenta o modelo lógico a ser submetido as ferramentas *Data Mining* para a exploração de padrões.

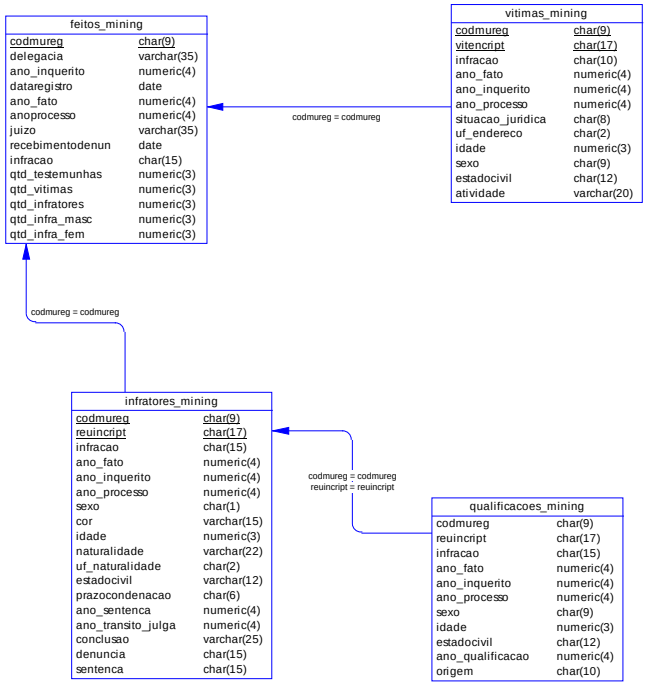


Figura 6.3 – Modelo transformado para extração

Quatro tabelas resultantes da fase de codificação são destinadas ao processo de *Data Mining*. Embora algumas ferramentas possuam limitações para trabalhar somente com dados numéricos, considerou-se relevante para o entendimento do usuário final a adoção de campos alfanuméricos.

O processo de carga para o banco de dados foi realizado por uso de comandos SQL, o que só foi possível pela adoção de um modelo intermediário, entre o modelo de origem (ZIM) e o modelo a ser minerado.

O elemento tempo foi adicionado em cada tabela a ser minerada. O Atributo DATADOFATO foi convertido do tipo DATA para o tipo ANO, passando assim a ser um preditor mais significativo no processo. O Atributo INFRAÇÃO, retrata a infração cometida, de acordo com a qualificação dada.

Os atributos : ANO_FATO, ANO_INQUERITO e ANO PROCESSO foram derivados da tabela FEITO. Observa-se que os atributos estão em formato diferente do encontrado na origem. O atributo INFRAÇÃO foi derivado da tabela QUALIFICAÇÃO.

Construiu-se um dicionário de dados para refletir as diversas mudanças ocorridas. Registra-se que um METADADOS auxilia no processo regular de *Data Mining*, pois mantém informações relativas a descrição do conteúdo, estrutura e definição dos objetos do banco de dados; a origem dos dados; nomes comerciais e técnicos dos dados e etc.

6.4 Ferramentas Data Mining Aplicadas ao CIPO

As ferramentas utilizadas no processo *Data Mining* foram selecionadas na internet e satisfazem o processo de exploração de dados, modelados para investigação.

6.4.1 WizRule

Ferramenta da *WizSoft inc*, é baseada em tecnologia *Data Mining* e realiza análise complexa revelando inconsistências, erros e casos a serem auditados.

WizRule é uma ferramenta de indução de regras capaz de trabalhar com praticamente todos os SGBDs do Mercado via fonte de dados ODBC (*Open Database Connectivity*), além de realizar leitura direta em Gerenciadores de Arquivos *.DBF, Tabelas do MS Access, MS SQL SERVER e ORACLE.

Para o software descobrir exceções, primeiro ele descobre todas as regras de um *dataset* (ponto forte do *WizRule*), pois ele é baseado em um algoritmo matemático capaz de revelar todas as regras contidas no *dataset*. Como resultado é apresentado uma lista de registros que são improváveis em referência as regras descobertas. Estes registros são suspeitos de conterem erros, ou no mínimo devem ser examinados.

Utilizou-se uma cópia de demonstração do *WizRule* com limitação de número de registros (1000). Para efeito de demonstração foram analisados dados referente a infrações cometidas nos anos de 1998 e 1999.

Como é padrão em um aplicativo Windows, o *WizRule* faz uso de botões de controle o que o torna bastante intuitivo. A figura 6.4 apresenta a tela inicial do *WizRule*.

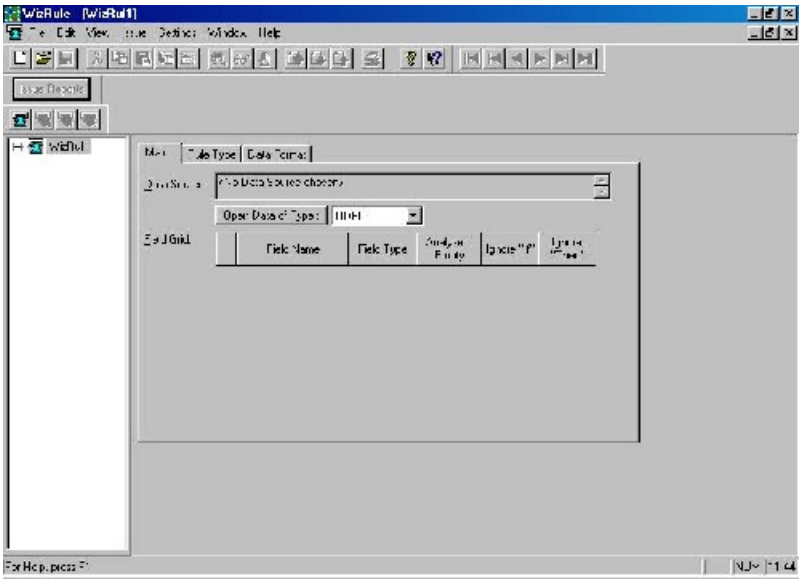


Figura 6.4 – Tela inicial do WizRule

A figura 6.5 apresenta a seleção de uma fonte de dados que contém o *dataset* a ser verificado pelo WizRule.

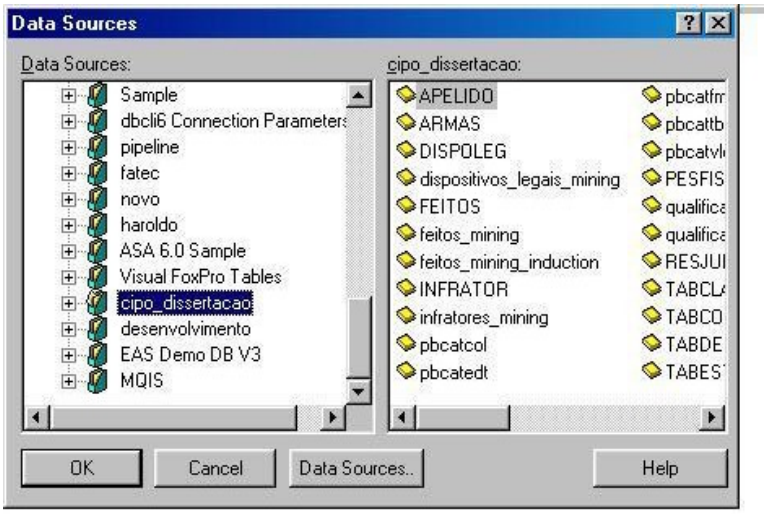


Figura 6.5 – Escolha do *dataset*

A figura 6.6. apresenta um *dataset* selecionado. Três abas devem ser configuradas com valores que serão considerados pelo algoritmo na geração das regras. É nesta interação que o usuário refina o processo de extração. Os valores das abas podem ser alterados para refletir o interesse do usuário, permitindo ajustes.

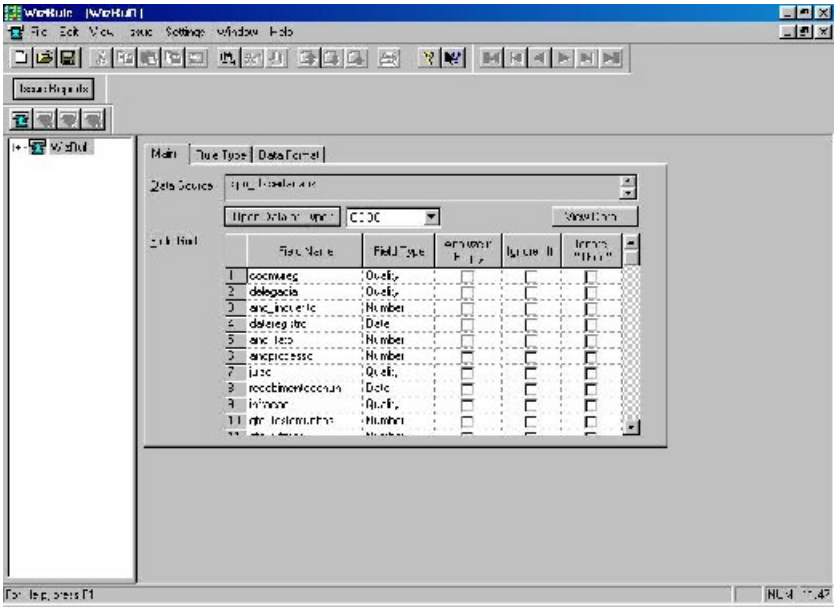


Figura 6.6 – Condições para extração

A figura 6.7 apresenta os resultados obtidos pelo *WizRule* após a geração das regras. Pela adoção de janelas simples e intuitivas, pode-se notar :

- O conteúdo de cada registro dos dados;
- O gráfico localizado a direita da janela apresenta a nível de improbabilidade do desvio corrente apresentado, em comparação a todos os desvios encontrados.
- Nível máximo da improbabilidade, é a probabilidade máxima que é desviada inexplicavelmente da descoberta das regras. O número de botões indica o total de números de desvios que tem um nível de improbabilidade superior a 0,6;
- A lista de regras na parte direita da janela, apresenta as regras que desviam do padrão. Contendo informações sobre o nome do dataset, o número total de registros lidos e as regras geradas. Após cada regra contendo desvio é apresentada a relação com o número de cada registro que desvia da regra. As regras podem ser apresentadas como: incondicionais, if-then e como fórmulas;

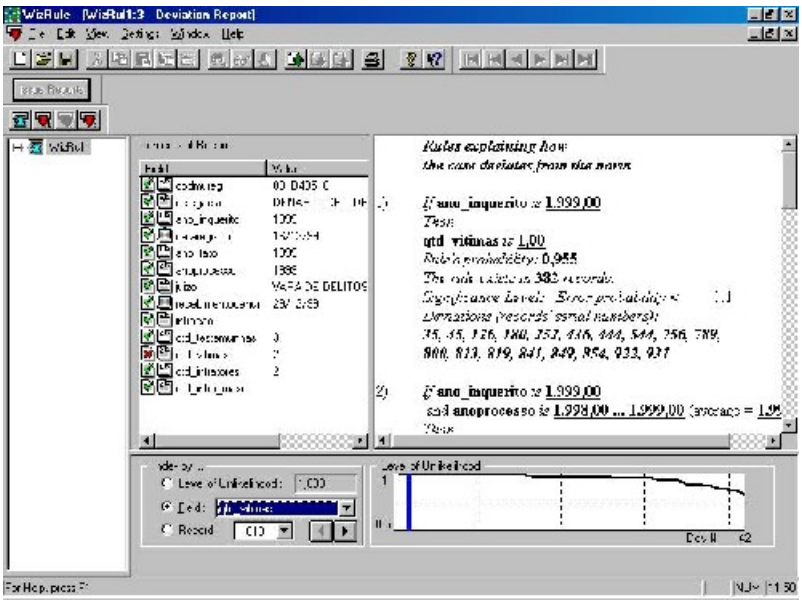


Figura 6.7 – Geração de regras

6.4.2 CART for Windows

A Ferramenta da Salford Systems, é baseada em tecnologia de árvore de decisão para análise, predição e pre-processamento de dados. Ela foi primeira colocada no concurso internacional de *Data Mining* organizada pela ACM (Association for Computing Machinery) no ano 2000.

CART na versão atual, trabalha com variáveis cujo nome possua no máximo 8 caracteres (sem conter caracteres especiais ou espaço), e não é configurado para trabalhar com dados alfanuméricos, sendo necessário realizar a conversão destes dados para numérico antes da importação.

CART trabalha com todos os bancos de dados de mercado fazendo uso de um utilitário chamado DBMS copy, que é capaz de converter dados alfanuméricos.

CART é o único sistema de árvore de decisão baseado no código original CART desenvolvido pelo universidade de Stanford e Universidade da Califórnia, possuindo extensões que foram desenvolvidas pela Salford System. A figura 6.7 apresenta a janela inicial do CART.

A aba model é requerida para todas as árvores; é aonde as variáveis alvo (target) e do preditor são especificadas, elas devem ser inicializadas antes da construção da árvore. A figura 6.8 apresenta a aba model.

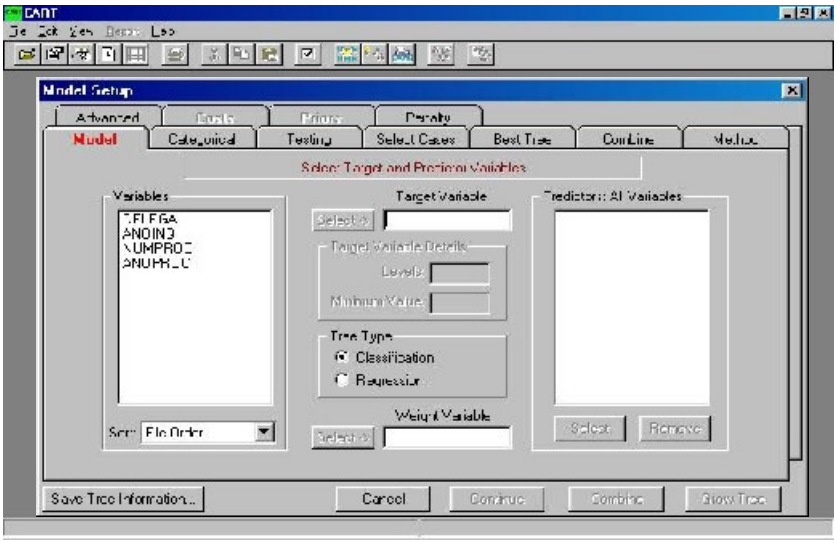


Figura 6.8 – Aba model

A aba Method, especifica a regra de divisão usada para classificação e regressão em árvores de decisão, e avalia quando são computadas combinações lineares e se são usadas como divisores. Estes valores devem ser assinalados antes da construção da árvore. Uma regra de divisão é escolhida para cada tipo de árvore. A regra apropriada será usada quando CART começar a construir a árvore, e ele sabe se é uma árvore de classificação ou de regressão. A figura 6.9 apresenta a aba method.

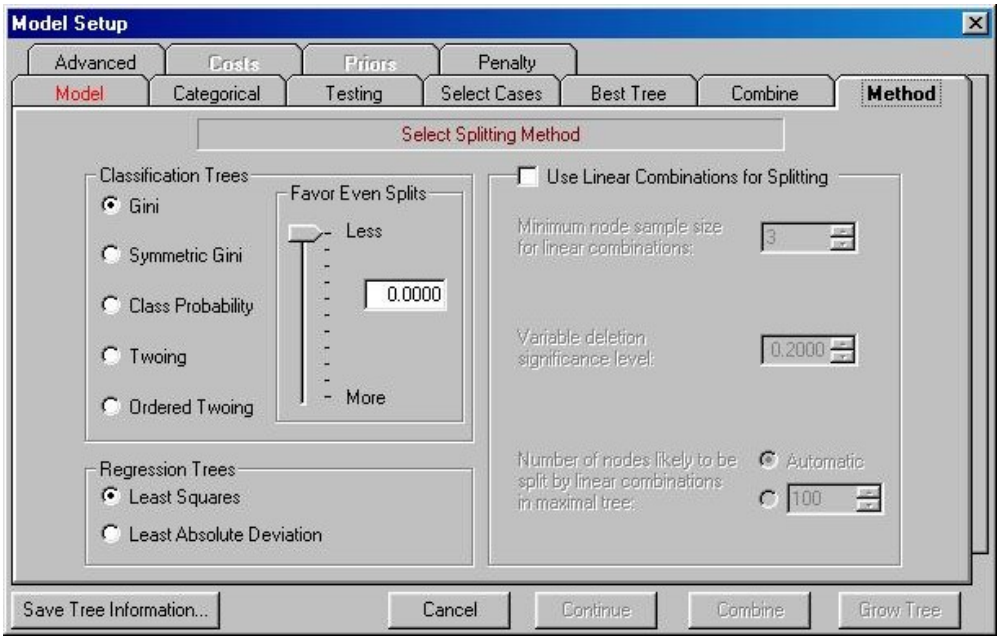


Figura 6.9 – Aba Method

Árvores de classificação

As cinco opções de divisão da regra para uma árvore de classificação são Gini (default), Symmetric Gini, Probabilidade de Classe, Twoing, e Twoing ordenada.

Árvores de regressão

CART oferece duas regras de regressão: Least squares e least absolute deviation.

Escolhendo uma Regra de divisão

As regras seguintes estão baseado na experiência da Salford System com empresas de telecomunicações, bancos e áreas de pesquisa de mercado, e pode não ser aplicadas literalmente a outros assuntos ou até mesmo a outros datasets. Não obstante, eles representam um conjunto consistente de achados empíricos :

1.) para uma variável dependente nível 2 que pode ser predita com um erro relativo menor que 0.50, a regra de divisão *Gini* é geralmente melhor.

2.) para uma variável dependente nível 2 que pode ser predita com um erro relativo de 0.80 ou pior, *Twong* tende a ser melhor.

3.) para variáveis designadas de 4 a 9 níveis, *Twoing* tem uma chance boa de ser a melhor regra de divisão.

4.) para variáveis dependentes categóricas de alto nível com 10 ou mais níveis, *Twoing* e *Power—Modified Twoing* são freqüentemente consideravelmente mais precisas que Gini.

A janela de visualização CART contem :

- Estatísticas Descritivas (*mean, SD, n, sum*) para aprender e testar amostras.
- Seqüência da Árvore
- Custo inicial e designação de classes (para classificação) ou significado Inicial e Discrepância (para regressão)
- Gráficos de erro e de complexidade
- Diagrama. de Baixa-resolução da árvore
- Detalhes para cada nodo e divisão.
- Detalhes para cada nodo terminal.

A figura 6.10 apresenta a janela com o resultado do processo de construção de uma árvore.

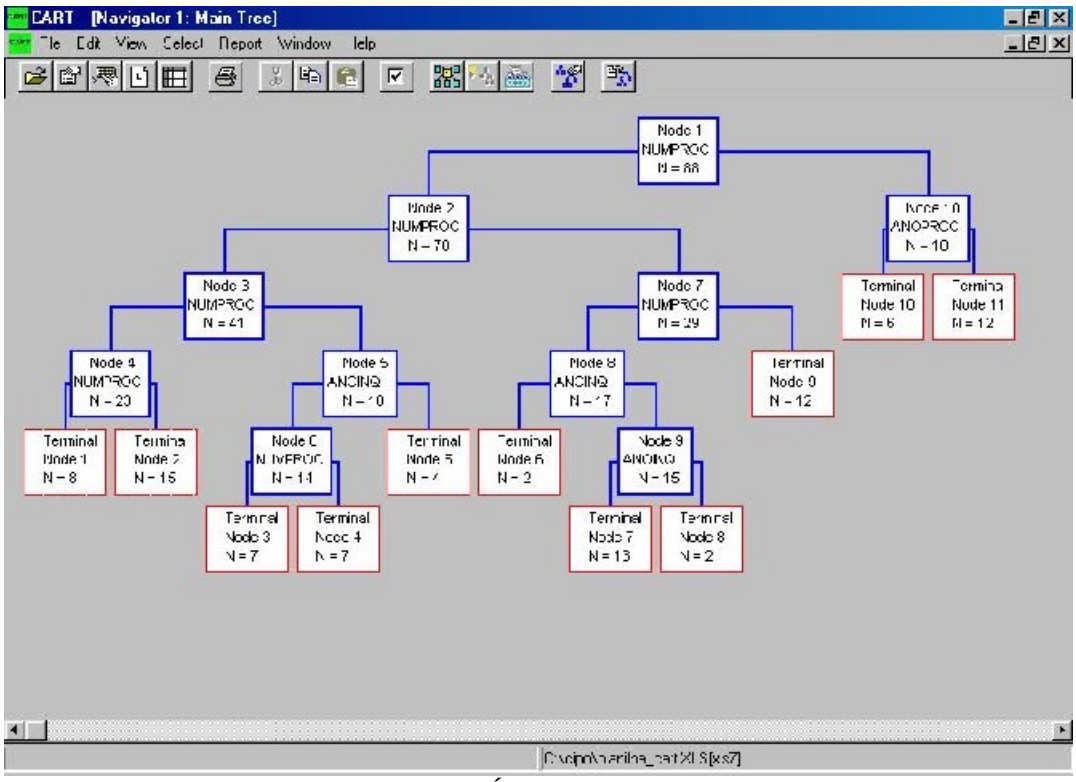


Figura 6.10 – Árvore CART gerada

6.4.3 XpertRule Miner

Ferramenta da Attar Software, para descoberta de padrão e exploração de dados.

O algoritmo usado pelo XpertRule está baseado no algoritmo *Dynamic* item set *counting* (DIC) descrito no artigo "*Dynamic Item set Counting and Implication Rules*" escrito por S. Brin, R. Motwani, J.D. Ullman, e S. Tsur: que foi apresentado na conferência internacional de Procedimentos da ACM SIGMOD (Tucson, Arizona, E.U.A., em maio de 1997).

O algoritmo consiste em duas fases : uma de conta que inclui o leitura de todas as transações em uma única partição e outra de geração de candidato.

As duas fases são processadas em alternância até que todo o conjunto de itens descoberto seja confirmado (raro ou freqüente).

Para o XpertRule Miner, *Data Mining* é um processo que pode ser dividido nos seguinte passos:

- Seleção de *datasets* ;
- Transformação, limpeza e preparação de dados;
- Visualização de dados;
- Mineração e exploração de padrões;
- Desenvolvimento de padrões descobertos.

XpertRule apóia todos esses passos do projeto de dados, usando interface gráfica que permite ao usuário definir uma seqüência de operações de mineração que usam uma interface intuitiva. Há um número pequeno de conceitos que o usuário precisa entender para obter o máximo do XpertRule. A figura 6.11 apresenta a janela incial do XpertRule na operação de seleção de uma fonte de dados.

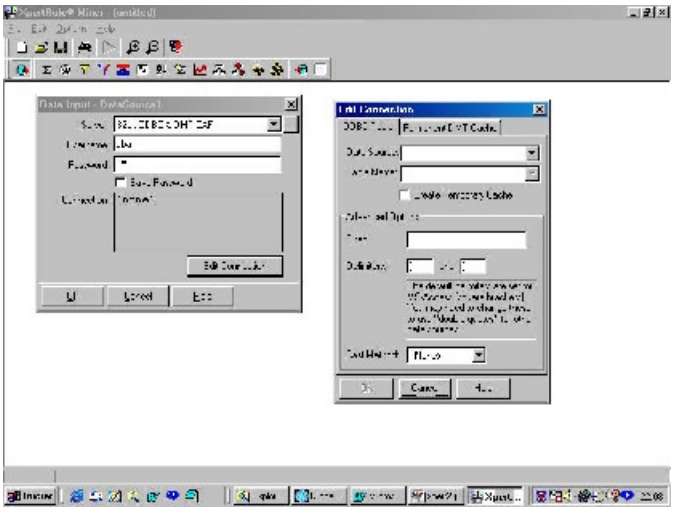


Figura 6.11 – Seleção do dataset no XpertRule

Definindo a fonte de dados

Qualquer fonte de dados ODBC pode ser usada como uma operação de entrada no XpertRule.

Transformando os dados

Na maioria dos casos a fonte de dados selecionada precisa de uma sucessão de operações de transformações (manipulações de tabelas e campos) antes de estarem prontos para a mineração de padrões.

Conferindo os resultados de transformação

Relatórios podem ser anexados a fonte de dados ou depois das operações de transformação. Estes podem ser usados para conferir se as operações de transformação estão corretas.

Minerando os dados transformados

Operações de mineração de padrões(Árvore Miner, Associação de Caso e Associação de Transação) devem ser unificadas para uma fonte de dados física . Uma seqüência de operações de transformação devem ser realizadas para uma operação de fonte de dados de saída.

Uma seqüência de operação de mineração é referenciada como um script miner. Uma vez desenvolvido um script, o usuário pode executa-lo. XpertRule só permitirá executar scripts com interconecções entre operações válidas. A figura 6.12 Apresenta a definição dos valores para a geração de uma árvore.

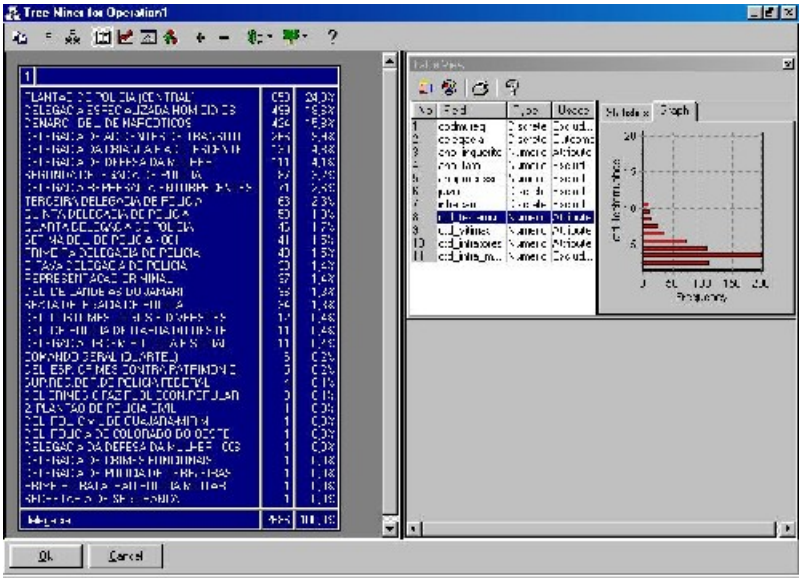


Figura 6.12 – Configuração da árvore XpertRule

Desenvolvendo uma árvore com resultados interativos discreto, o sistemas apresenta uma lista de todos os atributos a toda as divisão. Estes atributos são ordenados por Entropia ou *Chi-square*, como definido pelo usuário. Consequentemente pode-se usar Entropia ou *Chi-square* como base para seleção de atributo ou forçar qualquer outro atributo de divisão escolhido pelo usuário.

O valor do galho relacionamento/numérico limiar de uma divisão de atributo também pode ser mudado para o que é considerado melhor pelo algoritmo. Isto forçará a reavaliação da Entropia e *Chi-square* e se necessário uma reclasificação dos atributos.

O Nível de Significado e a taxa de redução de erro de cada atributo de uma divisão são apresentadas. Esta informação pode ser usada para decidir quando interromper o crescimento da árvore.

A figura 6.13 apresenta uma árvore gerada pela XpertMiner.

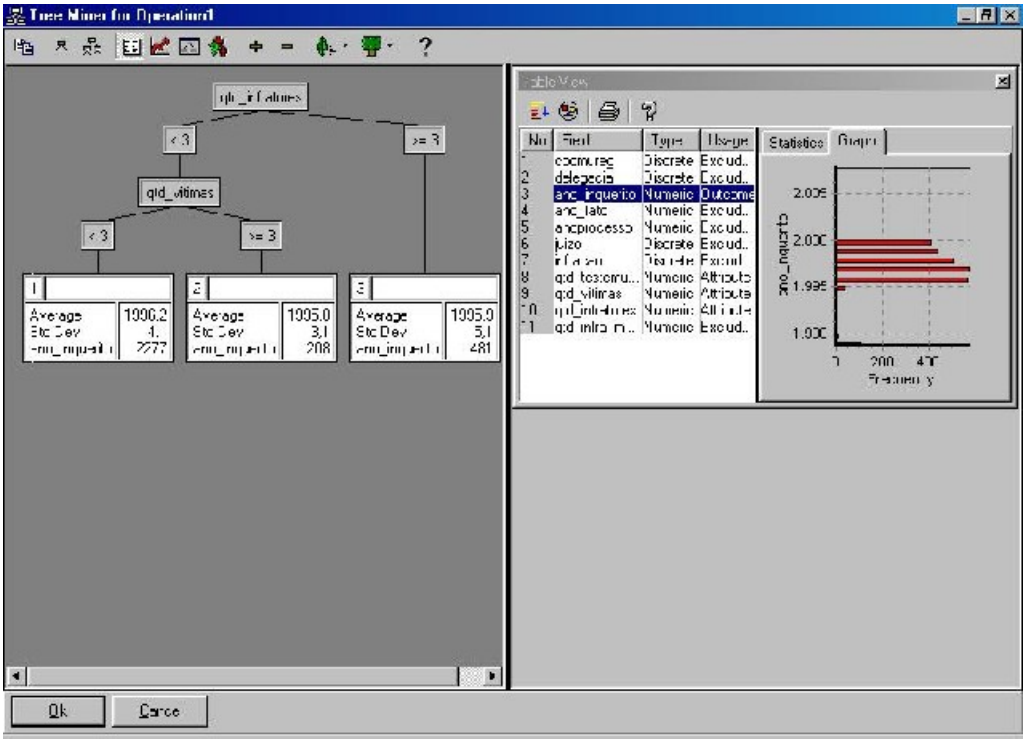


Figura 6.13 – Definindo as variáveis para mineração

6.5 Resultados obtidos

Muito mais do que comparar ferramentas, buscou-se demonstrar que o volume de dados são passíveis de gerar conhecimento significativo para Instituições Públicas, e ainda, demonstrar o estágio em que as ferramentas estão e da possibilidade de integração com outros aplicativos.

Os resultados obtidos pelas ferramentas mencionadas foram apresentados aos Órgãos do Ministério público para mensuração do seu valor. Muito embora tenha se mostrado viável, somente os Diretores dos órgãos podem ratificar a riqueza do conhecimento encontrado e consequentemente decidir pela adoção dessas ferramentas no processo decisório do Ministério Público de Rondônia.

A constatação da necessidade de uma auditoria completa nos dados coletados ao longo dos anos é a primeira medida tomada a partir deste trabalho.

As ferramentas WizRule e XpertRule possuem acesso direto as fontes de dados, facilitando a realização de operações entre *datasets* como *join*, e gerando informações a partir desse. A ferramenta CART necessita de um aplicativo para acesso a Bancos de Dados e trabalha com um arquivo em formato proprietário. CART apresenta uma interface sofisticada para configuração da árvore, porém necessitou-se transformar o *dataset* em uma planilha excell (aplicativo do pacote Microsoft Office) para minera-lo.

O processo de preparação dos dados deve considerar a ferramenta a ser utilizada, haja vista, que existem peculiaridades específicas de cada uma que devem ser respeitadas. É difícil preparar o modelo a ser minerado e então submete-lo a uma ferramenta qualquer. CART é um bom exemplo disso. Conhecer a ferramenta e sua forma de atuação é um ponto relevante no processo *Data Mining*.

O Ministério Público do Estado de Rondônia conta com aproximadamente 100.000 feitos (inquéritos) armazenados nos últimos 12 anos, distribuídos em 20 comarcas, sendo que 50% destes originados na capital. Adicionando-se a esse número a quantidade de infratores, testemunhas, vítimas, qualificações e etc, o volume de dados torna-se inviável para avaliações por técnicas tradicionais como relatórios resumidos e gráficos.

O grande resultado obtido diz respeito a decisões tomadas pela alta administração a partir da apresentação do processo *Data Mining* no Ministério Público:

- Realizar auditoria completa no Banco de dados do CIPO, para identificação dos tipos de erros e para definição de uma política de correção e integração dos dados;
- Autorizar a adoção de uma ferramenta de indução de regras capaz de auxiliar o processo de auditoria;

- Autorizar o desenvolvimento de um novo aplicativo capaz de integrar todas as ações envolvendo o Ministério Público em todas as fases aonde as ações acontecem, em um nível de detalhamento que permita exploração;
- Autorizar a aquisição de hardware específico e robusto o suficiente para acomodar os dados do novo sistema;
- Autorizar estudo de viabilidade para integração de todas as comarcas da Instituição ao prédio central do Ministério Público, o que vai permitir a unificação de vários aplicativos em uso atualmente e que são fonte de redundância e erros;
- Normatizar o uso de ferramentas baseadas em conhecimento aos órgãos da alta administração;

No anexo V é apresentado o modelo lógico que está sendo abstraído para o novo sistema. Ressalta-se que é um modelo preliminar.

Pretende-se ainda apresentar o novo sistema aos demais Ministérios Públicos Estaduais, como ferramenta capaz de mensurar e sumarizar as atividades do Ministério Público Brasileiro.

6.5.1 Resultados da base minerada

6.5.1.1 Árvore de decisão

Como mencionado no Capítulo III, as árvores de decisão são um modelo cujo resultado é visualizado como uma árvore, sendo que os galhos são perguntas de classificação e as folhas são partições dos dados classificados.
A figura 6.14 apresenta uma árvore de decisão que considera INFRAÇÃO como resultado do processo de *Data Mining*. Conclui-se facilmente que Homicídio e Estupro foram infrações mais evidentes no período do que infrações de Tóxico.

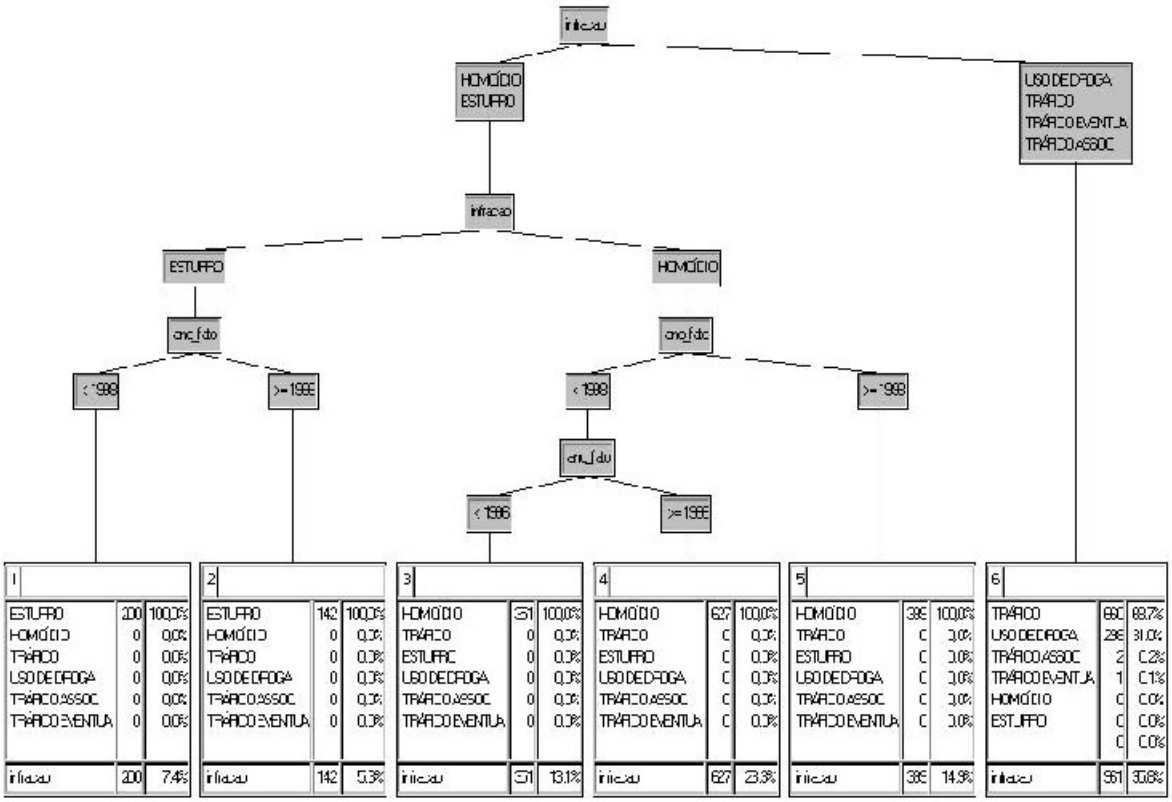


Figura 6.14 – Árvore tendo INFRAÇÃO como preditor

A figura 6.15 apresenta uma árvore de decisão obtida utilizando-se a cor do infrator como critério de classificação. É importante frisar que a ferramenta busca automaticamente os melhores preditores para a classificação dos dados, ou seja a densidade é testada para encontrar o melhor divisor que classifica os dados.

A figura 6.16 apresenta uma árvore cuja classificação desejada é por sexo do infrator. Obtém-se como resultado que para crimes de homicídio, estupro e tóxico a maioria dos infratores e significativamente composta por homens.

Árvore é um modelo que permite visualizar de forma simples o resultado e suas conseqüências, é possível concluir que um volume considerável de infratores de estupro, tráfico, uso de drogas e tráfico associado não possui o sexo informado, o que gerou 36% de infratores com sexo ignorado. Isto pode ser um fator importante dependendo da mineração executada pelo analista.

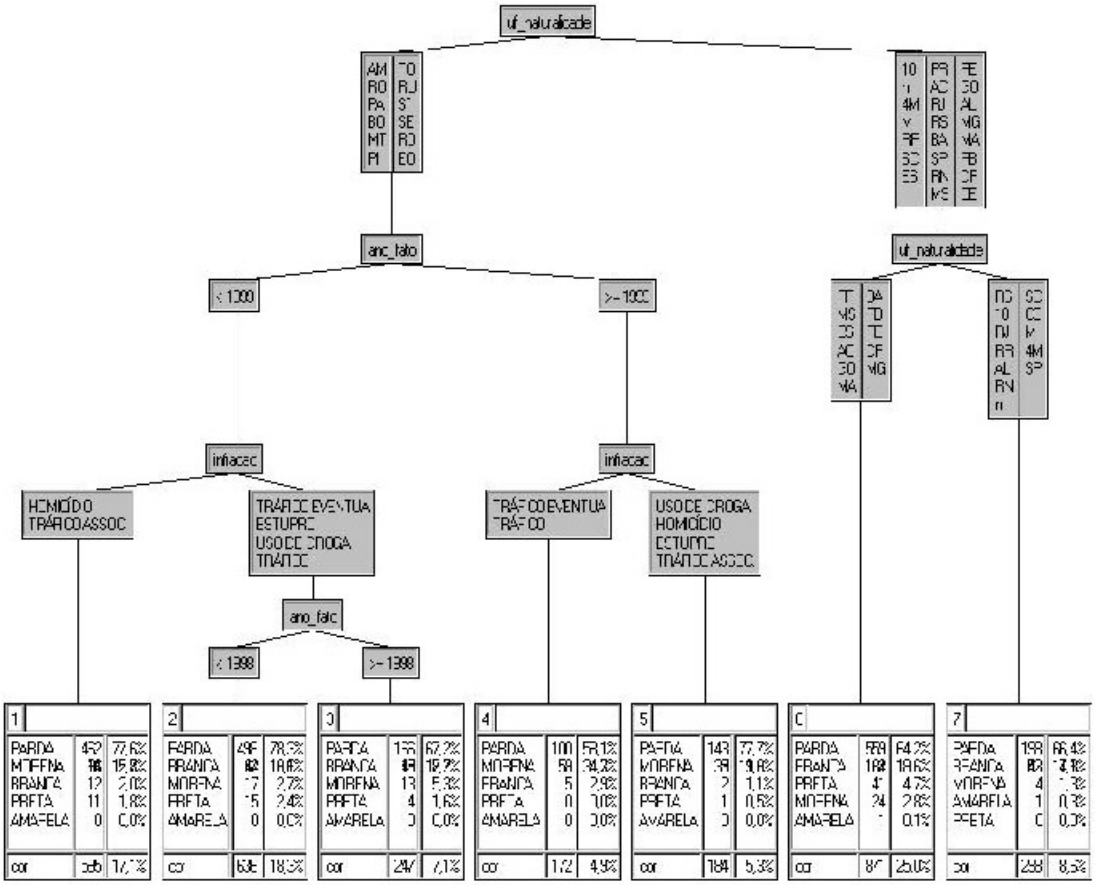


Figura 6.15 – Árvore tendo COR como preditor

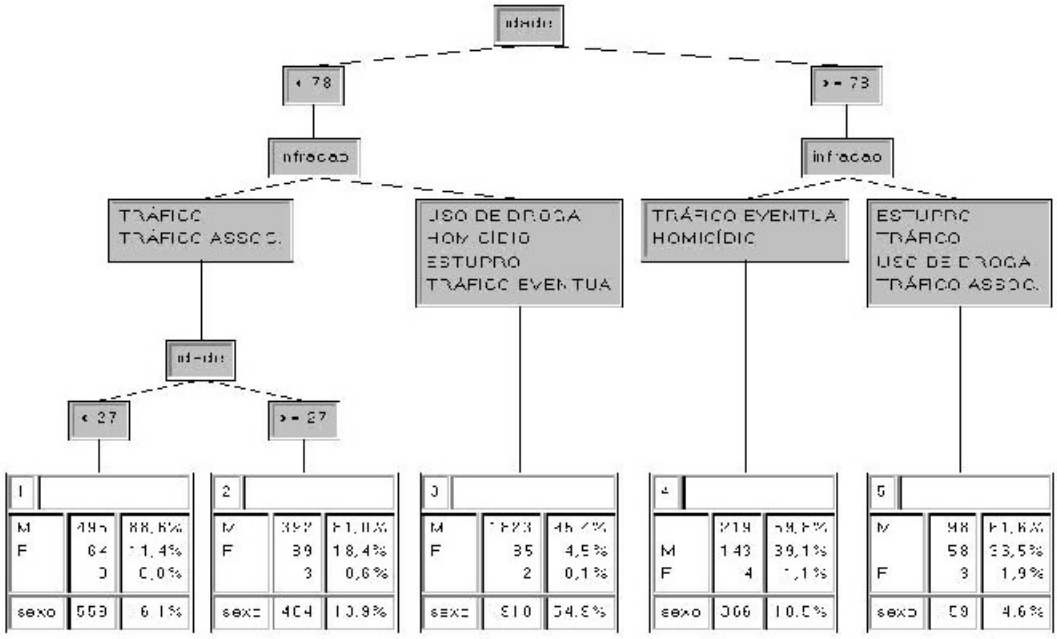


Figura 6.16 – Árvore tendo SEXO como preditor

6.5.1.2 Regras

Regras são uma das formas mais comuns e mais conhecidas de ferramentas *Data Mining*. É um modelo intuitivo com a qual as pessoas estão acostumadas a manusear no dia a dia. Possuem a característica de poder expor todos os padrões possíveis contidos em um *dataset*.

São apresentadas a seguir regras mineradas considerando-se FEITOS que são as infrações cometidas e INFRATORES que são as pessoas que cometeram os delitos. Observa-se ainda que embora estas regras existam fortemente no *dataset*, isto não significa que devam ser aplicadas sempre, ou seja, a regra SE infração é ESTUPRO ENTÃO cor é PARDA ocorre em 98% dos casos contidos no *dataset*, não é possível assumir que "toda" pessoa PARDA é um "estuprador".

Regras geradas a partir do dataset FEITOS

If **delegacia** *is* **DELEGACIA DA CRIANCA E ADOLESCENTE**
Then
infracao *is* **ESTUPRO**
Rule's probability: **0,989**
The rule exists in **87** *records.*
Significance Level: Error probability is almost 0
Deviations (records' serial numbers):
558

If **delegacia** *is* **DELEGACIA DA CRIANCA E ADOLESCENTE**
and **ano_fato** *is* **1.999,00**
Then
infracao *is* **ESTUPRO**
Rule's probability: **1,000**
The rule exists in **40** *records.*
Significance Level: Error probability is almost 0

If **infracao** *is* **TRÁFICO**
and **qtd_vitimas** *is* **1,00**
and **qtd_infratores** *is* **2,00**
Then
juizo *is* **VARA DE DELITOS DE ENTORPECENTES**
Rule's probability: **1,000**
The rule exists in **49** *records.*
Significance Level: Error probability is almost 0

Regras geradas a partir do dataset INFRATORES

If **infracao** *is* **ESTUPRO**
Then
cor *is* **PARDA**
Rule's probability: **0,978**
The rule exists in **89** *records.*
Significance Level: Error probability < 0,0001
Deviations (records' serial numbers):
113, 690

If **infracao** *is* **USO DE DROGA**
Then
sexo *is* **M**
Rule's probability: **0,985**
The rule exists in **131** *records.*
Significance Level: Error probability < 0,001
Deviations (records' serial numbers):
352, 694

If **infracao** *is* **HOMICÍDIO**
and **ano_fato** *is* **1.999,00**
and **sexo** *is* **M**
and **estadocivil** *is* **SOLTEIRO**
Then
cor *is* **PARDA**
Rule's probability: **0,968**
The rule exists in **60** *records.*
Significance Level: Error probability < 0,01
Deviations (records' serial numbers):
297

If **infracao** *is* **HOMICÍDIO**
and **ano_fato** *is* **1.999,00**
and **estadocivil** *is* **SOLTEIRO**
Then
cor *is* **PARDA**
Rule's probability: **0,956**
The rule exists in **65** *records.*
Significance Level: Error probability < 0,01
Deviations (records' serial numbers):
114, 297

If **infracao** *is* **ESTUPRO**
Then
sexo *is* **M**
Rule's probability: **0,967**
The rule exists in **88** *records.*
Significance Level: Error probability < 0,1
Deviations (records' serial numbers):
605, 732, 762

If **idade** *is* **28,00**
Then
cor *is* **PARDA**
Rule's probability: **0,969**
The rule exists in **31** *records.*
Significance Level: Error probability < 0,1
Deviations (records' serial numbers): ***456***

CAPITULO VII

CONCLUSÕES E RECOMENDAÇÕES

7.1 Conclusões

Este trabalho abordou o uso de ferramentas *Data Mining* aplicadas a Bancos de Dados de instituições Públicas permanentes, dentre estas instituições encontra-se o Ministério Público, com uma missão assegurada pela constituição federal e que recaem sobre toda a população do pais.

Os progressos obtidos com novas tecnologias baseadas em informação, têm contribuído para a redução no custo de softwares de mineração de conhecimento, o que tem atraído organizações de vários segmentos a buscarem nestes softwares conhecimentos estratégicos sobre seus negócios.

Nos capítulos anteriores foram analisadas técnicas que estão sendo empregadas com sucesso em grandes bases de dados, com foco na descoberta de padrões para aprendizagem, oferecendo aos analistas de negócios das instituições públicas, informações que podem auxiliar na criação de atos e medidas preventivas e corretivas que afetem a sociedade.

Como aplicação prática, testou-se soluções de mercado em uma base de dados do Ministério Público de Rondônia, usando ferramentas que implementam algoritmos de árvore de decisão e indução de regras.

O resultado obtido pela aplicação validado pelos Órgãos do Ministério Público de Rondônia demonstrou que existe conhecimento útil em bases de dados de Instituições públicas (muito mais do que o retorno do investimento (ROI) *Data Mining* deve auxiliar no sucesso do negócio), tanto que, a partir deste trabalho, o Ministério Público de Rondônia adotou um conjunto de medidas (seção 6.5) que criarão o ambiente necessário para a utilização de ferramentas *Data Mining* integradas aos sistemas tradicionais de tomada de decisão em suas ações.

7.2 Recomendações para futuros trabalhos

Recomenda-se para trabalhos futuros a ampliação das áreas cobertas pelo Controle de Inquéritos Policiais, podendo ser explorado para as diversas áreas armazenadas no banco de dados : Criminal, Cível, Infância e Juventude, Meio Ambiente, Cidadania, Defesa das vítimas, Consumidor, Atendimento ao público, Controle externo da atividade Policial e Recursos constitucionais, Considerando todos os municípios e regiões cobertos pela aplicação.

Recomenda-se o estudo de aplicações baseadas em algoritmos que implementem redes neurais, agentes, *nearest neighbor*, algoritmos genéticos, mensurando de forma mais abrangente os parâmetros de exatidão, explicação e integração, importantes na definição de qual a melhor técnica a ser adotada para mineração.

Recomenda-se o estudo para integração dos resultados obtidos por ferramentas *Data Mining* com OLAP, o que permitirá ao usuário visualizar os resultados de forma mais natural.

Recomenda-se ainda, o estudo de ferramentas *Data Mining* aplicado a internet como meio para o conhecimento do comportamento dos internautas e para ser aplicado a sistemas de detecção de invasões na rede corporativa ou em computadores servidores.

REFERÊNCIAS BIBLIOGRÁFICAS

[BERS97] **BERSON,** Alex **SMITH,** Stephen J.. Data Warehousing, Data Mining, & OLAP. 1ed. New York: McGraw-Hill, 1997.

[CAMP98] **CAMPOS,** Maria Luiza **FILHO,** Data Warehouse . Universidade Federal do Rio de Janeiro - 1998;

[DODG98] **DODGE,** Gary **GORMAN,** Tim.. Oracle8 Data Warehousing.1 ed. New York: Wiley Computer Publishing, 1998.

[GUPT97] **GUPTA,** Vivek R. An Introduction to Data Warehousing, System Services corporation, Chicago, Illinois, August 1997

[HAME98] **HAMERMESH,** Daniel S. Crime and the timing of Work, National Bureau of Economic Research - June 1998

[HERB00] **HERBERT**, Simon **GALLAGHER**, Simon **QUICK**, Joe **REPPART**, Ken. PowerBuilder 7.0. 1 ed. : Sams Publishing, 2000.

[INMO97] **INMON,** W.H. **HACKATHORN,** Richard D.. Como Usar O Data Warehouse, 2 ed. IBPI Press, 1997.

[INMO98] **INMON,** W. H. **WELCH,** J. D. **GLASSEY,** Katherine L.. Gerenciando Data Warehouse. 1 ed. São Paulo: Makron Books, 1999.

[INMO97b] **INMON,** W. H. Como Construir o Data Warehouse. 2 ed. Rio de Janeiro: Editora Campus, 1997.

[KIMB96] **KIMBALL,** R., "Mastering Data Extraction", DBMS Magazine, junho 1996.

[ORAC98] **ORACLE,** ORACLE Applications DATA WAREHOUSE, July 1998 Disponível na WWW no endereço: http://www.oracle.com/

[PARS95] **PARSAYE**, K. Sandwich Paradigm, Data Warehousing and Mining, Database Programming and Design, April 1995. Disponível na WWW no endereço : http://www.datamining.com

[PARS96] **PARSAYE,** K., Rules Are Much More Than Decision Trees**.** The Journal of Data Warehousing, December 1996.

[PARS97] **PARSAYE**, K. A Characterization of *Data Mining*, Technologies and Processes. Journal of Data Warehousing – Dezembro 1997. Disponível na WWW no endereço : http://www.datamining.com

[PYLE99] **PYLE,** Dorian. Data Preparation For *Data Mining*. 1 ed. San Francisco: Morgan Kaufmann Publishers, 1999.

[ROLL] **ROLLEIGH,** Louis **THOMAS,** Joe. Data Integration: The Warehouse Foundation Disponível na WWW no endereço: http://www.acxiom.com/

[SCHE00] **SCHERER**, Douglas **GAYNOR**, Jr William **VALENTINSEN**, Arlene **CURSETJEE**, Xerxes. ORACLE 8i Dicas & Técnicas. 1. Ed. : Oracle Press, 2000.

[VERI97] **VERITY,** John W. Sistema Selecionando dados estratégicos – Business Week - fevereiro 1997;

Building a Corporate information System : The Role of the Data Mart Disponível na WWW no endereço: http://www.system-services.com

DataMines for DataWarehouses. By Information Discovery Disponível na WWW no endereço: http://www.datamining.com/dm4dw.htm

ROLLEIGH, Louis TOMAS, Joe. Data Integration : The warehouse Foundation. www.acxion.com

Endereço Eletrônico de Artigos :

http://www.acxion.com

http://www.datamining.com/

http://www.datawarehousing.com/

http://www.dmreview.com/

http://www.kdd.org/

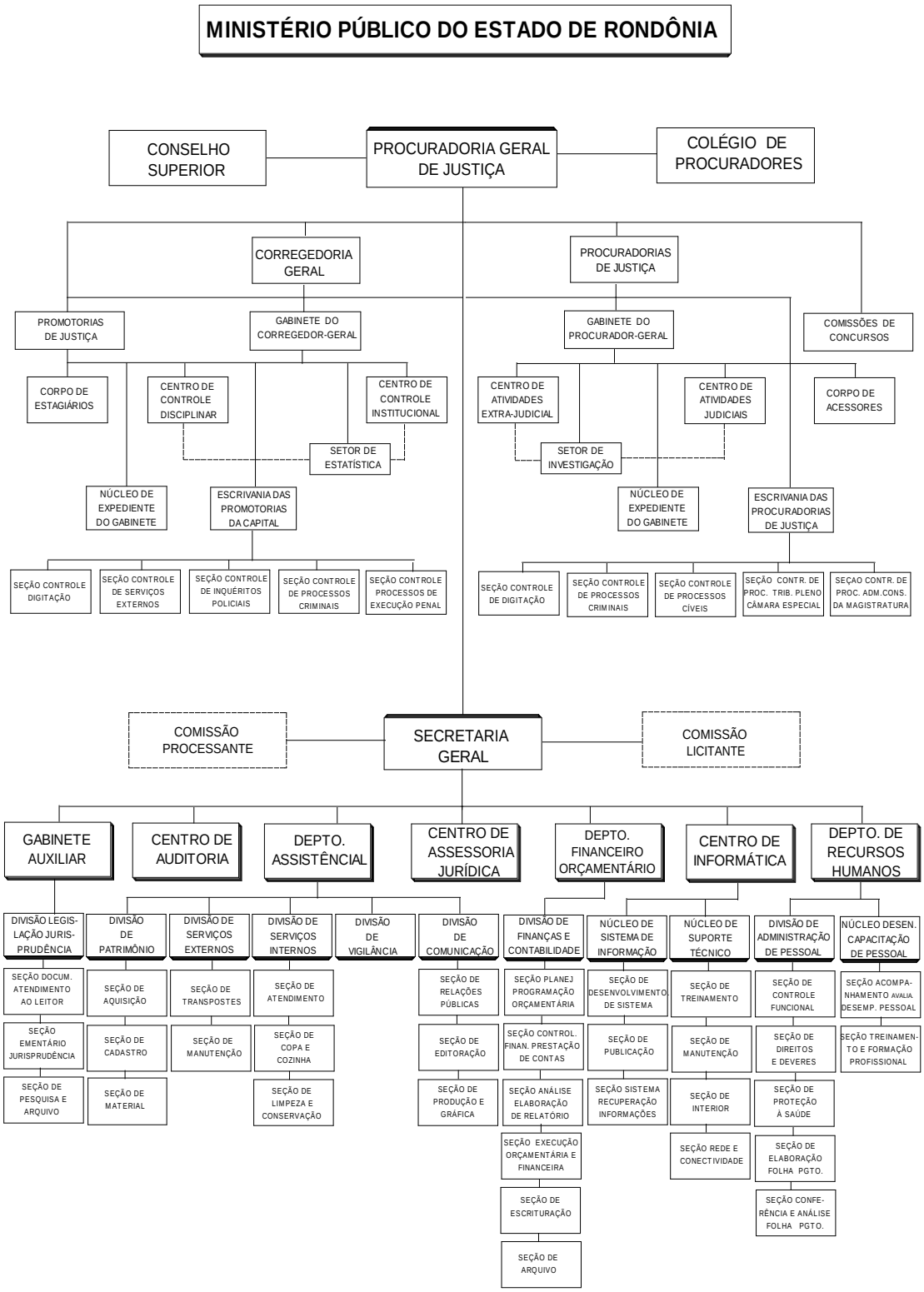
http://www.kdnuggets.com/

http://www.techguide.com/

http://idc.com

ANEXOS

ANEXO 1


MINISTÉRIO PÚBLICO DO ESTADO DE RONDÔNIA
CONSELHO SUPERIOR
PROCURADORIA GERAL DE JUSTIÇA
COLÉGIO DE PROCURADORES
CORREGEDORIA GERAL
PROCURADORIAS DE JUSTIÇA
PROMOTORIAS DE JUSTIÇA
GABINETE DO CORREGEDOR-GERAL
GABINETE DO PROCURADOR-GERAL
COMISSÕES DE CONCURSOS
CORPO DE ESTAGIÁRIOS
CENTRO DE CONTROLE DISCIPLINAR
CENTRO DE CONTROLE INSTITUCIONAL
CENTRO DE ATIVIDADES EXTRA-JUDICIAL
CENTRO DE ATIVIDADES JUDICIAIS
CORPO DE ACESSORES
SETOR DE ESTATÍSTICA
SETOR DE INVESTIGAÇÃO
NÚCLEO DE EXPEDIENTE DO GABINETE
ESCRIVANIA DAS PROMOTORIAS DA CAPITAL
NÚCLEO DE EXPEDIENTE DO GABINETE
ESCRIVANIA DAS PROCURADORIAS DE JUSTIÇA
SEÇÃO CONTROLE DIGITAÇÃO
SEÇÃO CONTROLE DE SERVIÇOS EXTERNOS
SEÇÃO CONTROLE DE INQUÉRITOS POLICIAIS
SEÇÃO CONTROLE DE PROCESSOS CRIMINAIS
SEÇÃO CONTROLE PROCESSOS DE EXECUÇÃO PENAL
SEÇÃO CONTROLE DE DIGITAÇÃO
SEÇÃO CONTROLE DE PROCESSOS CRIMINAIS
SEÇÃO CONTROLE DE PROCESSOS CÍVEIS
SEÇÃO CONTR. DE PROC. TRIB. PLENO CÂMARA ESPECIAL
SEÇAO CONTR. DE PROC. ADM.CONS. DA MAGISTRATURA
COMISSÃO PROCESSANTE
SECRETARIA GERAL
COMISSÃO LICITANTE
GABINETE AUXILIAR
CENTRO DE AUDITORIA
DEPTO. ASSISTÊNCIAL
CENTRO DE ASSESSORIA JURÍDICA
DEPTO. FINANCEIRO ORÇAMENTÁRIO
CENTRO DE INFORMÁTICA
DEPTO. DE RECURSOS HUMANOS
DIVISÃO LEGIS-LAÇÃO JURIS-PRUDÊNCIA
DIVISÃO DE PATRIMÔNIO
DIVISÃO DE SERVIÇOS EXTERNOS
DIVISÃO DE SERVIÇOS INTERNOS
DIVISÃO DE VIGILÂNCIA
DIVISÃO DE COMUNICAÇÃO
DIVISÃO DE FINANÇAS E CONTABILIDADE
NÚCLEO DE SISTEMA DE INFORMAÇÃO
NÚCLEO DE SUPORTE TÉCNICO
DIVISÃO DE ADMINISTRAÇÃO DE PESSOAL
NÚCLEO DESEN. CAPACITAÇÃO DE PESSOAL
SEÇÃO DOCUM. ATENDIMENTO AO LEITOR
SEÇÃO EMENTÁRIO JURISPRUDÊNCIA
SEÇÃO DE PESQUISA E ARQUIVO
SEÇÃO DE AQUISIÇÃO
SEÇÃO DE CADASTRO
SEÇÃO DE MATERIAL
SEÇÃO DE TRANSPOSTES
SEÇÃO DE MANUTENÇÃO
SEÇÃO DE ATENDIMENTO
SEÇÃO DE COPA E COZINHA
SEÇÃO DE LIMPEZA E CONSERVAÇÃO
SEÇÃO DE RELAÇÕES PÚBLICAS
SEÇÃO DE EDITORAÇÃO
SEÇÃO DE PRODUÇÃO E GRÁFICA
SEÇÃO PLANEJ PROGRAMAÇÃO ORÇAMENTÁRIA
SEÇÃO CONTROL. FINAN. PRESTAÇÃO DE CONTAS
SEÇÃO ANÁLISE ELABORAÇÃO DE RELATÓRIO
SEÇÃO EXECUÇÃO ORÇAMENTÁRIA E FINANCEIRA
SEÇÃO DE ESCRITURAÇÃO
SEÇÃO DE ARQUIVO
SEÇÃO DE DESENVOLVIMENTO DE SISTEMA
SEÇÃO DE PUBLICAÇÃO
SEÇÃO SISTEMA RECUPERAÇÃO INFORMAÇÕES
SEÇÃO DE TREINAMENTO
SEÇÃO DE MANUTENÇÃO
SEÇÃO DE INTERIOR
SEÇÃO REDE E CONECTIVIDADE
SEÇÃO DE CONTROLE FUNCIONAL
SEÇÃO DE DIREITOS E DEVERES
SEÇÃO DE PROTEÇÃO Á SAÚDE
SEÇÃO DE ELABORAÇÃO FOLHA PGTO.
SEÇÃO CONFE-RÊNCIA E ANÁLISE FOLHA PGTO.
SEÇÃO ACOMPA-NHAMENTO AVALIA. DESEMP. PESSOAL
SEÇÃO TREINAMEN-TO E FORMAÇÃO PROFISSIONAL

ANEXO 2

Computadores Distribuídos por Comarca – Rondônia

Comarcas/Promotorias	Quantidade
001 Porto Velho	189
002 Guajará-Mirim	2
003 Ji-Paraná	4
004 Vilhena	4
005 Pimenta Bueno	4
006 Cacoal	3
007 Ariquemes	5
008 Ouro Preto do Oeste	5
009 Jaru	2
010 Presidente Médici	1
011 Espigão do Oeste	1
012 Colorado do Oeste	1
013 Costa Marques	2
014 Rolim de Moura	2
015 Cerejeiras	2
016 Santa Luzia do Oeste	2
017 Alvorada do Oeste	2
018 Alta Floresta do Oeste	1
019 Nova Brasilândia do Oeste	1
020 Machadinho do Oeste	1
TOTAL	**234**

ANEXO 3

Nomes das Entidades do CIPO

Campos
Cartaguia
Etiqueta
Feitos
Infrator
Menus
Movimento
Pesfisica
Qualificacao
Tabandamento
Tabclasse
Tabcor
Tabdelegacia
Tabestadocivil
Tabgrupo
Tabjuizo
Tabmunicipio
Tabnacionalidade
Tabocupacao
Tabpromotor
Tabpromotoria
Tabresjuizo
Testemunha
Vitima
Dispoleg
Respromoto
Resjuizo
Usuarios
Tabmotivo
Tabetiqueta
Nocartorio
Docmov
Resumo

Atributos das Entidades do CIPO

OwnerName	SN	FieldName	Type	Length	Decimals	Reqd	Index
cartaguia	1	codmureg	char	9	0	yes	yes
cartaguia	2	processo	char	8	0	yes	no
cartaguia	3	comarca	char	20	0	no	no
cartaguia	4	vara	char	15	0	no	no
cartaguia	5	reu	char	35	0	no	no
cartaguia	6	cor	char	2	0	no	no
cartaguia	7	juiz	char	35	0	no	no
cartaguia	8	datasente	date	8	0	no	no
cartaguia	9	incursao	char	66	0	no	no
cartaguia	10	pena	char	69	0	no	no
cartaguia	11	multa	char	25	0	no	no
cartaguia	12	custas	char	25	0	no	no
cartaguia	13	acessoria	char	25	0	no	no
cartaguia	14	medseguranca	char	22	0	no	no
cartaguia	15	consiste	char	34	0	no	no
cartaguia	16	sursis	char	9	0	no	no
cartaguia	17	condsursis	char	128	0	no	no
cartaguia	18	suspcondic	date	8	0	no	no
cartaguia	19	trasjulgado	date	8	0	no	no
cartaguia	20	acordao	char	256	0	no	no
cartaguia	21	prisao	date	8	0	no	no
cartaguia	22	interrupcoes	char	108	0	no	no
cartaguia	23	vencprisao	date	8	0	no	no
cartaguia	24	personalidade	char	58	0	no	no
cartaguia	25	periculosida	char	57	0	no	no
cartaguia	26	condenacoes	char	60	0	no	no
cartaguia	27	obs	char	60	0	no	no
cartaguia	28	reuencript	char	17	0	no	yes
CHAMADOPOR	1	CODMUREG	CHAR	9	0	YES	YES
CHAMADOPOR	2	REUENCRIPT	CHAR	17	0	NO	YES
CHAMADOPOR	3	APELIDO	CHAR	20	0	NO	YES
DISPOLEG	1	CODMUREG	CHAR	9	0	YES	YES
DISPOLEG	2	REUENCRIPT	CHAR	17	0	YES	NO
DISPOLEG	3	DATAINDICIA	DATE	8	0	NO	NO
DISPOLEG	4	INDICIAMENTO	VARCHAR	180	0	NO	NO
docmov	1	CODMUNICIPIO	CHAR	3	0	YES	NO
docmov	2	REG_MP	CHAR	6	0	YES	NO
docmov	3	ENTIDADE	CHAR	4	0	NO	no
docmov	4	DISTRIBUICAO	DATE	8	0	NO	no
docmov	5	DIST_HORA	char	8	0	NO	NO
docmov	6	RECEBIMENTO	DATE	8	0	NO	no
docmov	7	REC_HORA	char	8	0	NO	NO
docmov	8	CADPROMOTOR	VARCHAR	5	0	NO	no
docmov	9	DIAS	INT	3	0	NO	NO
docmov	10	CODMOTIVO	CHAR	2	0	NO	no
docmov	11	MOTIVO	VARCHAR	120	0	NO	NO
etiqueta	1	numero	char	6	0	no	no
FAZPARTEQUAD	1	CODMUREG	CHAR	9	0	YES	YES
FAZPARTEQUAD	2	REUENCRIPT	CHAR	17	0	NO	NO
FAZPARTEQUAD	3	QUADRILHA	CHAR	20	0	NO	NO
FAZPARTEQUAD	4	DATAVERIFIC	DATE	8	0	NO	NO
FEITOS	1	CODMUNICIPIO	CHAR	3	0	YES	NO
feitos	1	codmureg	char	9	0	no	yes
FEITOS	2	REG_MP	CHAR	6	0	YES	NO
feitos	2	inquerito	char	6	0	no	yes

FEITOS	3	DELEGACIA	CHAR	4	0	NO	NO
feitos	3	processo	char	7	0	no	yes
FEITOS	4	NUMINQUERITO	CHAR	4	0	NO	NO
FEITOS	5	ANOINQUERITO	CHAR	2	0	NO	NO
FEITOS	6	LIVRO	CHAR	3	0	NO	NO
FEITOS	7	DATAREGISTRO	DATE	8	0	NO	NO
FEITOS	8	DATAFATO	DATE	8	0	NO	NO
FEITOS	9	LOCALFATO	CHAR	30	0	NO	NO
FEITOS	10	NUMPROCESSO	CHAR	5	0	NO	NO
FEITOS	11	ANOPROCESSO	CHAR	2	0	NO	NO
FEITOS	12	JUIZO	CHAR	4	0	NO	NO
FEITOS	13	ANDAMENTO	CHAR	2	0	NO	NO
FEITOS	14	RECDENUNCIA	DATE	8	0	NO	NO
INFRATOR	1	CODMUNICIPIO	CHAR	3	0	YES	NO
infrator	1	codmureg	char	9	0	no	yes
INFRATOR	2	REG_MP	CHAR	6	0	YES	NO
INFRATOR	3	NOME	CHAR	35	0	NO	YES
INFRATOR	4	REUENCRIPT	CHAR	17	0	NO	NO
INFRATOR	5	RG	VARCHAR	15	0	NO	NO
INFRATOR	6	SEXO	CHAR	1	0	NO	NO
INFRATOR	7	CLASSE	CHAR	2	0	NO	NO
INFRATOR	8	GRUPO	CHAR	2	0	NO	NO
INFRATOR	9	OCUPACAO	CHAR	3	0	NO	NO
INFRATOR	10	COR	CHAR	2	0	NO	NO
INFRATOR	11	UF	CHAR	2	0	NO	NO
INFRATOR	12	NACIONALIDADE	CHAR	2	0	NO	NO
INFRATOR	13	NASCIMENTO	CHAR	8	0	NO	NO
INFRATOR	14	ESTADOCIVIL	CHAR	2	0	NO	NO
INFRATOR	15	NUMERO	CHAR	4	0	NO	NO
INFRATOR	16	CEP	CHAR	5	0	NO	NO
INFRATOR	17	END_UF	CHAR	2	0	NO	NO
INFRATOR	18	DATA_IDENTIFICA	DATE	8	0	NO	NO
INFRATOR	19	PAI	VARCHAR	30	0	NO	NO
INFRATOR	20	MAE	VARCHAR	30	0	NO	NO
INFRATOR	21	NATURALIDADE	VARCHAR	22	0	NO	NO
INFRATOR	22	RUA	VARCHAR	30	0	NO	NO
INFRATOR	23	COMPLEMENTO	VARCHAR	15	0	NO	NO
INFRATOR	24	BAIRRO	VARCHAR	20	0	NO	NO
INFRATOR	25	CIDADE	VARCHAR	25	0	NO	NO
INFRATOR	26	RECOLHIDO	VARCHAR	10	0	NO	NO
MENCIONINSTRU	1	CODMUREG	CHAR	9	0	YES	YES
MENCIONINSTRU	2	DESCRICAO	VARCHAR	80	0	NO	NO
mencioninstru	3	remessa	date	8	0	no	no
mencioninstru	4	remetido	char	30	0	no	no
mencioninstru	5	dataregistro	date	8	0	no	no
menuss	1	telasubm	int	2	0	no	no
menuss	2	seqsubm	int	2	0	no	no
menuss	3	descricao	char	20	0	no	no
MOVIMENTO	1	CODMUNICIPIO	CHAR	3	0	YES	NO
movimento	1	codmureg	char	9	0	no	yes
MOVIMENTO	2	REG_MP	CHAR	6	0	YES	NO
MOVIMENTO	3	ENTIDADE	CHAR	4	0	NO	YES
MOVIMENTO	4	DISTRIBUICAO	DATE	8	0	NO	YES
MOVIMENTO	5	DIST_HORA	char	8	0	NO	NO
MOVIMENTO	6	RECEBIMENTO	DATE	8	0	NO	YES
MOVIMENTO	7	REC_HORA	char	8	0	NO	NO
MOVIMENTO	8	CADPROMOTOR	VARCHAR	5	0	NO	YES
MOVIMENTO	9	DIAS	INT	3	0	NO	NO
MOVIMENTO	10	CODMOTIVO	CHAR	2	0	NO	YES
MOVIMENTO	11	MOTIVO	VARCHAR	120	0	NO	NO

```
nocartorio                1 cad_promotor       char            5        0 yes yes
nocartorio                2 quantidade         int             4        0 no  no
nocartorio                3 mes                int             2        0 no  no
nocartorio                4 ano                int             4        0 no  no
------------------------------------------------------------------------------
PESFISICA                 1 CODMUNICIPIO       CHAR            3        0 YES NO
pesfisica                 1 codmureg           char            9        0 no  yes
PESFISICA                 2 REG_MP             CHAR            6        0 YES NO
PESFISICA                 3 VITENCRIPT         CHAR           17        0 NO  YES
PESFISICA                 4 NATURALIDADE       CHAR           22        0 NO  NO
PESFISICA                 5 UF                 CHAR            2        0 NO  NO
PESFISICA                 6 NACIONALIDADE      CHAR            2        0 NO  NO
PESFISICA                 7 NASCIMENTO         DATE            8        0 NO  NO
PESFISICA                 8 SEXO               CHAR            1        0 NO  NO
PESFISICA                 9 ESTADOCIVIL        CHAR            2        0 NO  NO
PESFISICA                10 PAI                CHAR           30        0 NO  NO
PESFISICA                11 MAE                CHAR           30        0 NO  NO
------------------------------------------------------------------------------
QUALIFICACAO              1 CODMUNICIPIO       CHAR            3        0 YES NO
Qualificacao              1 codmureg           char            9        0 no  yes
QUALIFICACAO              2 REG_MP             CHAR            6        0 YES NO
QUALIFICACAO              3 REUENCRIPT         CHAR           17        0 NO  YES
QUALIFICACAO              4 DATA               DATE            8        0 NO  NO
QUALIFICACAO              5 ARTIGO             CHAR            3        0 NO  NO
QUALIFICACAO              6 PARAGRAFO          CHAR            5        0 NO  NO
QUALIFICACAO              7 INCISO             CHAR            3        0 NO  NO
QUALIFICACAO              8 LETRA              CHAR            1        0 NO  NO
QUALIFICACAO              9 DIPLOMA            CHAR           15        0 NO  NO
QUALIFICACAO             10 ORIGEM             CHAR            1        0 NO  NO
QUALIFICACAO             11 OUTROS             CHAR           60        0 NO  NO
------------------------------------------------------------------------------
RESJUIZO                  1 CODMUREG           CHAR            9        0 YES YES
RESJUIZO                  2 REUENCRIPT         CHAR           17        0 YES NO
RESJUIZO                  3 PRAZOCOND          CHAR            6        0 NO  NO
RESJUIZO                  4 PRAZOSUSP          CHAR            6        0 NO  NO
RESJUIZO                  5 DATASENTENCA       DATE            8        0 NO  NO
RESJUIZO                  6 DATATRANSITOJUL    DATE            8        0 NO  NO
RESJUIZO                  7 CODIGOCONCL        CHAR            2        0 NO  NO
RESJUIZO                  8 DATAEXTINCAO       DATE            8        0 NO  NO
RESJUIZO                  9 CONDENACAO         VARCHAR       180        0 NO  NO
------------------------------------------------------------------------------
RESPROMOTO                1 CODMUREG           CHAR            9        0 YES YES
RESPROMOTO                2 REUENCRIPT         CHAR           17        0 YES NO
RESPROMOTO                3 DATADENUNCIA       DATE            8        0 NO  NO
RESPROMOTO                4 DENUNCIA           VARCHAR       180        0 NO  NO
------------------------------------------------------------------------------
resumo                    1 cadastro           char            5        0 yes yes
resumo                    1 cadpro             char           15        0 no  yes
resumo                    2 promotoria         char            4        0 yes yes
resumo                    3 anomes             char            6        0 yes yes
resumo                    4 inqrecebido        int             4        0 no  no
resumo                    5 prorecebido        int             4        0 no  no
resumo                    6 feitoanter         int             4        0 no  no
resumo                    7 inqpendent         int             4        0 no  no
resumo                    8 propendent         int             4        0 no  no
resumo                    9 inqcartorio        int             4        0 no  no
resumo                   10 procartorio        int             4        0 no  no
resumo                   11 feicartorio        int             4        0 no  no
------------------------------------------------------------------------------
sainfra                   1 CODMUNICIPIO       CHAR            3        0 YES NO
sainfra                   2 REG_MP             CHAR            6        0 YES NO
sainfra                   3 NOME               CHAR           35        0 NO  YES
sainfra                   4 REUENCRIPT         CHAR           17        0 NO  NO
sainfra                   5 RG                 alpha          15        0 NO  NO
sainfra                   6 SEXO               CHAR            1        0 NO  NO
```

```
sainfra                    7 CLASSE             CHAR            2        0 NO  NO
sainfra                    8 GRUPO              CHAR            2        0 NO  NO
sainfra                    9 OCUPACAO           CHAR            3        0 NO  NO
sainfra                   10 COR                CHAR            2        0 NO  NO
sainfra                   11 UF                 CHAR            2        0 NO  NO
sainfra                   12 NACIONALIDADE      CHAR            2        0 NO  NO
sainfra                   13 NASCIMENTO         CHAR            8        0 NO  NO
sainfra                   14 ESTADOCIVIL        CHAR            2        0 NO  NO
sainfra                   15 NUMERO             CHAR            4        0 NO  NO
sainfra                   16 CEP                CHAR            5        0 NO  NO
sainfra                   17 END_UF             CHAR            2        0 NO  NO
sainfra                   18 DATA_IDENTIFICA    numeric         8        0 NO  NO
sainfra                   19 PAI                alpha          30        0 NO  NO
sainfra                   20 MAE                alpha          30        0 NO  NO
sainfra                   21 NATURALIDADE       alpha          22        0 NO  NO
sainfra                   22 RUA                alpha          30        0 NO  NO
sainfra                   23 COMPLEMENTO        alpha          15        0 NO  NO
sainfra                   24 BAIRRO             alpha          20        0 NO  NO
sainfra                   25 CIDADE             alpha          25        0 NO  NO
sainfra                   26 RECOLHIDO          alpha          10        0 NO  NO
------------------------------------------------------------------------------
sinfra                     1 CODMUNICIPIO       CHAR            3        0 YES NO
sinfra                     2 REG_MP             CHAR            6        0 YES NO
sinfra                     3 NOME               CHAR           35        0 NO  YES
sinfra                     4 REUENCRIPT         CHAR           17        0 NO  NO
sinfra                     5 RG                 alpha          15        0 NO  NO
sinfra                     6 SEXO               CHAR            1        0 NO  NO
sinfra                     7 CLASSE             CHAR            2        0 NO  NO
sinfra                     8 GRUPO              CHAR            2        0 NO  NO
sinfra                     9 OCUPACAO           CHAR            3        0 NO  NO
sinfra                    10 COR                CHAR            2        0 NO  NO
sinfra                    11 UF                 CHAR            2        0 NO  NO
sinfra                    12 NACIONALIDADE      CHAR            2        0 NO  NO
sinfra                    13 NASCIMENTO         CHAR            8        0 NO  NO
sinfra                    14 ESTADOCIVIL        CHAR            2        0 NO  NO
sinfra                    15 NUMERO             CHAR            4        0 NO  NO
sinfra                    16 CEP                CHAR            5        0 NO  NO
sinfra                    17 END_UF             CHAR            2        0 NO  NO
sinfra                    18 DATA_IDENTIFICA    numeric         8        0 NO  NO
sinfra                    19 PAI                alpha          30        0 NO  NO
sinfra                    20 MAE                alpha          30        0 NO  NO
sinfra                    21 NATURALIDADE       alpha          22        0 NO  NO
sinfra                    22 RUA                alpha          30        0 NO  NO
sinfra                    23 COMPLEMENTO        alpha          15        0 NO  NO
sinfra                    24 BAIRRO             alpha          20        0 NO  NO
sinfra                    25 CIDADE             alpha          25        0 NO  NO
sinfra                    26 RECOLHIDO          alpha          10        0 NO  NO
------------------------------------------------------------------------------
sumario                    1 cadastro           char            5        0 yes yes
sumario                    1 cadpro             char           15        0 no  yes
sumario                    2 promotoria         char            4        0 yes yes
sumario                    2 cadprocod          char           17        0 no  yes
sumario                    3 anomes             char            6        0 yes yes
sumario                    4 codigo             char            2        0 yes yes
sumario                    5 quantidade         int             4        0 no  no
------------------------------------------------------------------------------
tabandamento               1 CODIGO             char            2        0 yes unq
tabandamento               2 DESCRICAO          char           80        0 no   no
------------------------------------------------------------------------------
TABCLASSE                  1 CODIGO             CHAR            2        0 YES unq
TABCLASSE                  2 DESCRICAO          VARCHAR        80        0 NO  NO
------------------------------------------------------------------------------
TABCOR                     1 CODIGO             CHAR            2        0 YES unq
TABCOR                     2 DESCRICAO          CHAR           15        0 NO  NO
------------------------------------------------------------------------------
```

```
TABDELEGACIA              1 CODIGO             CHAR            4        0 YES unq
TABDELEGACIA              2 DESCRICAO          CHAR           35        0 NO   NO
TABDELEGACIA              3 CODMUNICIPIO       CHAR            3        0 NO   NO
------------------------------------------------------------------------------
TABESTADOCIVIL            1 CODIGO             CHAR            2        0 YES unq
TABESTADOCIVIL            2 DESCRICAO          CHAR           12        0 NO   NO
------------------------------------------------------------------------------
TABETIQUETA               1 CODIGO             CHAR            3        0 YES YES
TABETIQUETA               2 ULTIMAETIQUETA     CHAR            6        0 NO   NO
------------------------------------------------------------------------------
TABGRUPO                  1 CODIGO             CHAR            2        0 YES unq
TABGRUPO                  2 DESCRICAO          VARCHAR        80        0 NO   NO
------------------------------------------------------------------------------
TABJUIZO                  1 CODIGO             CHAR            4        0 YES unq
TABJUIZO                  2 DESCRICAO          CHAR           35        0 NO   NO
------------------------------------------------------------------------------
TABMOTIVO                 1 CODMOTIVO          CHAR            2        0 YES YES
TABMOTIVO                 2 DESCRICAO          CHAR           35        0 NO  NO
------------------------------------------------------------------------------
TABMUNICIPIO              1 CODIGO             CHAR            3        0 YES unq
TABMUNICIPIO              2 DESCRICAO          CHAR           22        0 NO   NO
TABMUNICIPIO              3 CEP                CHAR            5        0 NO   NO
TABMUNICIPIO              4 UF                 CHAR            2        0 NO   NO
------------------------------------------------------------------------------
TABNACIONALIDADE          1 CODIGO             CHAR            2        0 YES unq
TABNACIONALIDADE          2 DESCRICAO          CHAR           15        0 NO   NO
------------------------------------------------------------------------------
TABOCUPACAO               1 CODIGO             CHAR            3        0 YES unq
TABOCUPACAO               2 DESCRICAO          CHAR           80        0 NO   NO
------------------------------------------------------------------------------
----
TABPROMOTOR               1 CADASTRO           CHAR            5        0 YES unq
TABPROMOTOR               2 CODMUNICIPIO       CHAR            3        0 NO   NO
TABPROMOTOR               3 CODPROMOTORIA      CHAR            4        0 NO   NO
TABPROMOTOR               4 NOME               CHAR           35        0 NO   NO
------------------------------------------------------------------------------
TABPROMOTORIA             1 CODIGO             CHAR            4        0 YES unq
TABPROMOTORIA             2 DESCRICAO          CHAR           35        0 NO   NO
TABPROMOTORIA             3 CODMUNICIPIO       CHAR            3        0 NO   NO
------------------------------------------------------------------------------
TABRESJUIZO               1 CODIGO             int             2        0 yes unq
TABRESJUIZO               2 DESCRICAO          char           25        0 no   no
------------------------------------------------------------------------------
TESTEMUNHA                1 CODMUREG           CHAR            9        0 YES YES
TESTEMUNHA                2 ORIGEM             CHAR            1        0 NO   NO
TESTEMUNHA                3 NUMERO             CHAR            4        0 NO   NO
TESTEMUNHA                4 FONE               CHAR            8        0 NO   NO
TESTEMUNHA                5 CEP                CHAR            5        0 NO   NO
TESTEMUNHA                6 END_UF             CHAR            2        0 NO   NO
TESTEMUNHA                7 RUA                VARCHAR        30        0 NO   NO
TESTEMUNHA                8 COMPLEMENTO        VARCHAR        15        0 NO   NO
TESTEMUNHA                9 BAIRRO             VARCHAR        20        0 NO   NO
TESTEMUNHA               10 CIDADE             VARCHAR        25        0 NO   NO
TESTEMUNHA               11 NOME               VARCHAR        35        0 NO   NO
------------------------------------------------------------------------------
usuarios                  1 nom_usuario        char           20        0 no  Yes
usuarios                  2 codmunicipio       char            3        0 No  no
usuarios                  3 identificacao      Numeric         8        0 yes yes
usuarios                  4 faz_audit          Numeric         8        0 no   no
usuarios                 14 GRUPO              Char            2        0 NO   NO
usuarios                 24 senha              VarChar        20        0 No  Yes
------------------------------------------------------------------------------
VITIMA                    1 CODMUNICIPIO       CHAR            3        0 YES NO
vitima                    1 codmureg           char            9        0 no  yes
VITIMA                    2 REG_MP             CHAR            6        0 YES NO
```

VITIMA	3	NOME	CHAR	35	0 NO	YES
VITIMA	4	VITENCRIPT	CHAR	17	0 NO	NO
VITIMA	5	SITJURIDICA	CHAR	1	0 NO	NO
VITIMA	6	ATIVIDADE	CHAR	20	0 NO	NO
VITIMA	7	RUA	CHAR	30	0 NO	NO
VITIMA	8	NUMERO	CHAR	4	0 NO	NO
VITIMA	9	COMPLEMENTO	CHAR	15	0 NO	NO
VITIMA	10	BAIRRO	CHAR	20	0 NO	NO
VITIMA	11	FONE	CHAR	8	0 NO	NO
VITIMA	12	CEP	CHAR	5	0 NO	NO
VITIMA	13	CIDADE	CHAR	25	0 NO	NO
VITIMA	14	END_UF	CHAR	2	0 NO	NO

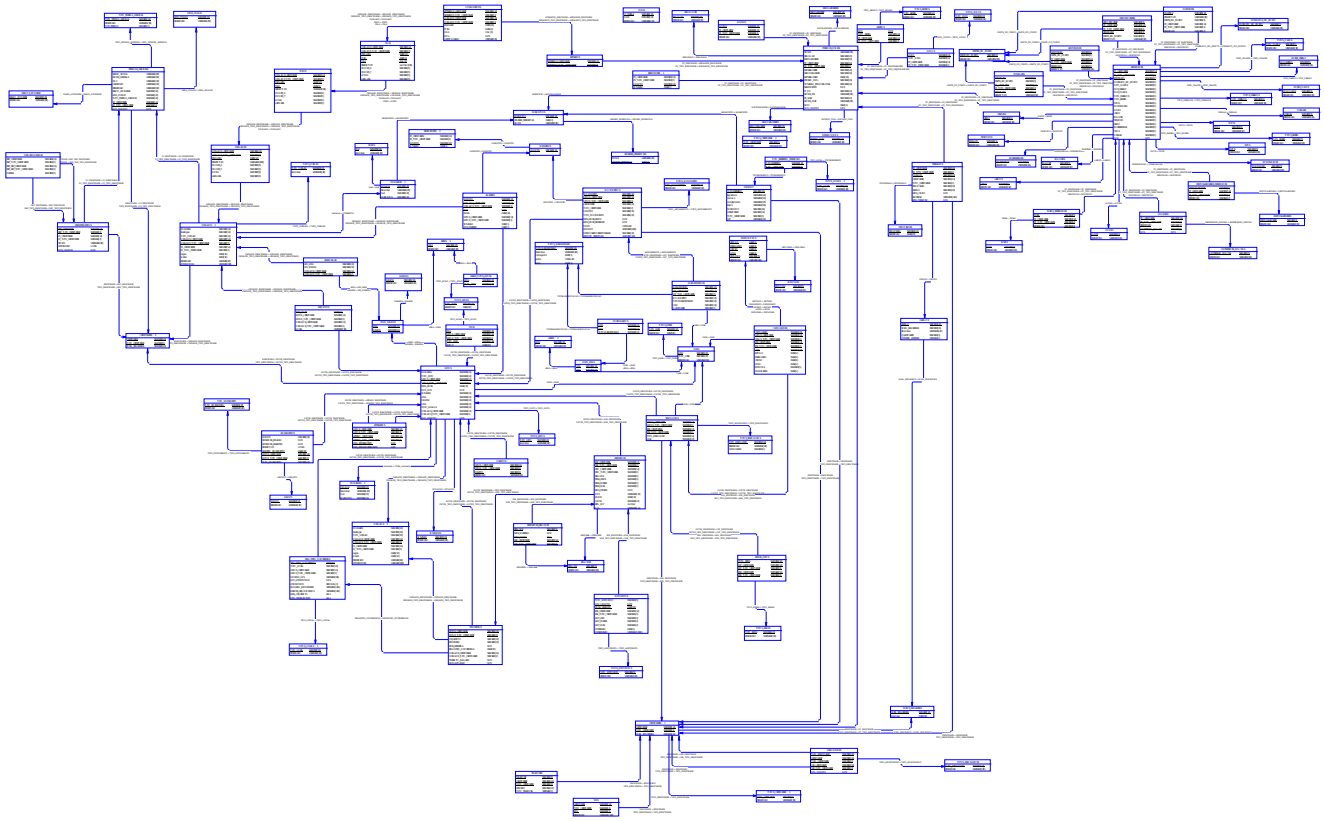
Relacionamentos do CIPO

acompanha	cartaguia.codmureg=infrator.codmureg and cartaguia.reuencript=infrator.reuencript
chamadopor	chamadopor.codmureg=infrator.codmureg and chamadopor.reuencript=infrator.reuencript
comprova	testemunha.codmureg=feitos.codmureg
credita	movimento.cadpromotor=tabpromotor.cadastro
encrimina	qualificacao.reuencript=infrator.reuencript and qualificacao.codmureg=infrator.codmureg
fazpartequad	fazpartequad.codmureg=infrator.codmureg and fazpartequad.reuencript=infrator.reuencript
feianda	feitos.andamento=tabandamento.codigo
gera	feitos.codmureg=movimento.codmureg
indicia	feitos.codmureg=infrator.codmureg
instaura	feitos.delegacia=tabdelegacia.codigo
lotado	tabpromotoria.codigo=tabpromotor.codpromotoria
mencioninstru	feitos.codmureg=mencioninstru.codmureg
mora	infrator.codmunicipio=tabmunicipio.codigo
movdelega	movimento.entidade=tabdelegacia.codigo
movjui	movimento.entidade=tabjuizo.codigo
movprt	movimento.entidade=tabpromotoria.codigo
recebe	tabpromotoria.codigo=movimento.entidade
recebedel	movimento.entidade=tabdelegacia.codigo
recebejuizo	movimento.entidade=tabjuizo.codigo
relata	feitos.codmureg=vitima.codmureg
indiciado	infrator.codmureg=dispoleg.codmureg and infrator.reuencript=dispoleg.reuencript
denunciado	infrator.codmureg=respromoto.codmureg and infrator.reuencript=respromoto.reuencript
sentenciado	infrator.codmureg=resjuizo.codmureg and infrator.reuencript=resjuizo.reuencript
conclusao	resjuizo.codigoconcl=tabresjuizo.codigo
originado	feitos.codmunicipio=tabmunicipio.codigo
mov_porque	movimento.codmotivo=tabmotivo.codmotivo
houve	movimento.recebimento <= saida.distribuicao and movimento.codmureg = saida.codmureg
ficou	tabpromotor.cadastro=nocartorio.cad_promotor

ANEXO 4

Find qualificacao wh diploma=”CPB” or diploma=”LEI”? → set_qualificacao
Find feitos wh codmunicípio=”001” and datafato >= 19950101 and datafato<= 19991231 → setfeitos
Find setfeitos indicia infrator → setinfrator
Find setinfrator encrimina qualificacao wh (qualificacao.diploma=”CPB” and (qualificacao.artigo=”121” or qualificacao.artigo=”213”)) or qualificacao.diploma=”LEI 6368/76” → setqualificacao
Find setqualificacao keep feitos → set_feitos
Find setqualificacao keep infrator → set_infrator
Find setqualificacao keep qualificacao → set_qualificacao
Find set_feitos relata vitima → setvitima
Find setvitima keep vitima → set_vitima
Find set_feitos comprova testemunha → settestemunha
Find settestemunha keep testemunha → set_testemunha
Find set_feitos mencioninstru mencioninstru → setarmas
Find setarmas keep mencioninstru → set_armas
Find set_feitos instaura tabdelegacia → set_delega
Find set_delega keep tabdelegacia → set_delegacia
Find set_feitos feianda tabandamento → set_anda
Find set_infrator fazpartequad fezpartequad → set_fezparte
Find set_infrator indiciado dispoleg → setdispoleg
Find setdispoleg keep dispoleg → set_dispoleg
Find set_infrator sentenciado resjuizo → setresjuizo
Find setresjuizo keep resjuizo → set_resjuizo
Find set_infrator acompanha cartaguia → setcarta
Find setcarta keep cartaguia → set_cartaguia
Find set_infrator chamadopor chamadopor → setapelido
Find setapelido keep apelido → set_apelido
Find set_infrator sentenciado resjuizo → setresjuizo
Find setresjuizo keep resjuizo → set_resjuizo
Find set_resjuizo conclusao tabresjuizo → settabresjuizo
Find settabresjuizo keep resjuizo → set_tabresjuizo
Find set_infrator denunciado respromoto → setrespromoto
Find setrespromoto keep respromoto → set_respromoto
Find set_infrator mora tabmunicipio → settabmunicipio
Find settabmunicipio keep tabmunicipio → set_tabmunicipio

ANEXO 5



ANEXO 6

FERRAMENTAS DATA MINING

Redes

Adaptive Logic Network

BioNet Simulator

Clementine

DataMining Workstation (DWM) and DWM/Marksman

FCM (Fuzzy Control Manager)

havBpNet++

KnowMan Basic Suite

Matlab: Neural Network Toolbox

Neural Bench

NeuroLution simulation and development system

Owl Neural Network

Propagator

SAS: Neural Network Add-On

Saxon

STATISTICA: Neural Networks

Thinks and Thinks Pro

TNs2Server

Trajan

Viscovery SOMine

WinBrain

Árvore de Decisão

AC2
Alice d'Isoft 6.0
Business Miner
C5.0/See5
CART 4.0 decision-tree software
Cognos Scenario
Decisionhouse
Kernel Miner
KnowledgeSEEKER
PolyAnalyst
SPSS AnswerTree,
XpertRule Miner
C4.5
EC4.5
IND
LMDT
ODBCMINE
OC1
PC4.5

Indução de Regras

AIRA

Datamite

PolyAnalyst

SuperQuery

WizWhy

XpertRule Miner

CBA

Claudien

CN2

DBPredictor,

KINOsuite-PR

RIPPER

GLOSSÁRIO

AID – Um dos primeiros algoritmos de *Data Mining* desenvolvido na Universidade de Michigan.

ALGORITMO – Formula matemática complexa que formaliza a lógica a ser executada para obtenção de um resultado esperado.

ALGORITMO GENÉTICOS – São resultantes da combinação de padrões que vêem sendo descobertos há muito tempo.

ÁRVORE DE DECISÃO – Forma simples de lógica condicional. Particiona uma tabela em tabelas menores baseada em um atributo determinado.

BANCO DE DADOS – Uma coleção de dados que são logicamente associados.

BANCO DE DADOS MULTIDIMENSIONAL – Um banco de dados construído com múltiplas dimensões e preenchido em “cubos” de dados em substituição ao modelo de dados relacional com tabelas de duas dimensões.

BRAIMSTORMING – Reunião participativa na qual os integrantes são motivados a partilhar idéias a respeito de determinado fato.

C4.5 – Algoritmo de *Data Mining* desenvolvido a partir do ID3, ID4 e ID6.

CART – Algoritmo de regressão estatística CHI.

CHAID – Algoritmo híbrido que faz uso de formula CHI dentro do algoritmo AID.

COBERTURA – Quão freqüente a regra se aplica. Corresponde ao número de registros (ou o percentual em relação ao total de registros) do dataset em que a regra se aplica.

DATA MART – Um subconjunto do Banco de Dados, usualmente orientado para um propósito específico.

DATA MINING – O processo da utilização dos resultados da exploração de dados para ajustar ou promover estratégias de negócios. É construído sobre padrões, tendências e exceções encontradas na exploração dos dados para dar suporte ao negócio.

DATA WAREHOUSE – Uma coleção de dados orientado por assunto, integrado, não volátil e com variação de tempo que dá suporte ao processo de tomada de decisão.

DATASET – Um conjunto específico de dados armazenados e obtidos modelados para *Data Mining*.

EXATIDÃO – Quão freqüente a regra está correta. A probabilidade de que se um antecedente é verdadeiro então o conseqüente será verdadeiro.

GINI – Algoritmo de índice para árvore de decisão, opera muito bem com números e textos e possui ótimo desempenho.

ID3 – O primeiro algoritmo desenvolvido para árvore de decisão.

INDUÇÃO DE REGRAS – Método de descoberta através de indução de regras sobre os dados. Geram hipóteses que se transformam em padrões. Podem utilizar lógica condicional, lógica de afinidade ou um híbrido.

METADADO – Dados sobre dados. Exemplos de Metadados incluem : Descrição dos elementos de dados, descrição dos tipos de dados, descrição dos atributos/propriedades, descrição dos processos/métodos.

MICROSOFT SQL SERVER 6.5 - Sistema Gerenciador de Banco de Dados Relacional da MICROSOFT CORPORATION.

OLAP – On line Analytical Processing. Originalmente introduzido em 1994 por E. F. Codd. Ferramenta de análise de dados com visualizações multidimensionais e cúbicas.

OLTP – On line Transaction Processing. Sistemas operacionais com a finalidade de análise de dados de curtos períodos de tempo, não superiores há 90 dias.

ORACLE8i - Sistema Gerenciador de Banco de Dados Objeto Relacional da ORACLE CORPORATION**.**

PREDIÇÃO – Utilizar dados existentes para conhecer como outros fatores se comportarão.

QUERY – Uma complexa sentença SELECT para suporte a decisão.

ROI – Return on Investiment – Uma medida financeira usada para quantificar o desejo de promover um esforço particular. Compara os benefícios obtidos no empreendimento contra o gasto na implementação. Geralmente é expresso em percentagem.

SAD – Sistema de apoio a decisão.

SGBD – Sistema Gerenciador de Banco de Dados que compreende um conjunto de programas que permite aos usuários definir, crear e manter Bancos de Dados.

SUPORTE A DECISÃO – Um conjunto de programas aplicativos que permitem ao usuário pesquisar em grandes bancos de dados informações para auxilio na tomada de decisão.

SYBASE ADAPTIVE SERVER ANYWHERE DATABASE VERSION 6.0 – Sistema Gerenciador de Banco de Dados Relacional da SYBASE INCORPORATION.

TUPLA – Termo utilizado em Bancos de Dados Relacionais para definir um registro, ou uma linha em uma tabela.

ZIM – Gerenciador de arquivos e linguagem de 4ª Geração.