



PUC RIO

DANTE JOSÉ ALEXANDRE CID

MINERAÇÃO DE DADOS COM TÉCNICAS DE ROUGH SETS

Dissertação de Mestrado

Departamento de Engenharia Elétrica

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO
Rio de Janeiro, 18 de dezembro de 2000

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Rua Marquês de São Vicente, 225 - Gávea
CEP 22453-900 Rio de Janeiro RJ Brasil
<http://www.puc-rio.br>

N.Cham. 621.3 C568 TESE UC

Autor Cid, Dante José Alexandre

Título Mineração de dados com técnicas de Rough Sets



Ex.2 PUC-Rio - PUCB

00170679

DANTE JOSÉ ALEXANDRE CID

index

MINERAÇÃO DE DADOS COM TÉCNICAS DE ROUGH SETS

Dissertação apresentada ao Departamento de Engenharia Elétrica da PUC/Rio como parte dos requisitos para a obtenção do título de Mestre em Engenharia Elétrica : Sistemas de Computação.

Orientadores: Emmanuel Piseces Lopes Passos
Marley M. B. R. Vellasco

Departamento de Engenharia Elétrica

Pontifícia Universidade Católica do Rio de Janeiro

Dezembro de 2000

109893



621.3
C568
TESE UC
v. 2

Dedicatória

à minha família

Agradecimentos

- À minha esposa, minha filha e meus pais, pelo amor, apoio e compreensão;
- À CAPES, pelo apoio financeiro;
- Aos Profs. Emmanuel Passos, Marley Vellasco e Marco Aurélio Pacheco, pelas valiosas orientações e contribuições para o desenvolvimento deste e de tantos outros trabalhos;
- Aos colegas do ICA e da oficina de manutenção;

RESUMO

Esta dissertação investiga a utilização de Rough Sets no processo de descoberta de conhecimento em Bancos de Dados (KDD - Knowledge Discovery in Databases). O objetivo do trabalho foi avaliar o desempenho da técnica de Rough Sets na tarefa de Classificação de Dados. Classificação é a tarefa da fase de Mineração de Dados que consiste na descoberta de regras de decisão, ou regras de inferência, que melhor representem um grupo de registros do banco de dados. O trabalho consistiu de cinco etapas principais: estudo sobre o processo de KDD; estudo sobre as técnicas de Rough Sets aplicadas à mineração de dados; análise de ferramentas de mineração de dados do mercado; evolução do projeto Bramining; e a realização de alguns estudos de caso para avaliar o Bramining.

O estudo sobre o processo de KDD abrangeu todas as suas fases: transformação, limpeza, seleção, mineração de dados e pós-processamento. O resultado obtido serviu de base para o aprimoramento do projeto Bramining.

O estudo sobre as técnicas de Rough Sets envolveu a pesquisa de seus conceitos e sua aplicabilidade no contexto de KDD. A teoria de Rough Sets foi apresentada por Zdzislaw Pawlak no início dos anos 80 como uma abordagem matemática para a análise de dados vagos e imprecisos. Este estudo permitiu sua aplicação na ferramenta de mineração de dados desenvolvida.

A análise de ferramentas de mineração de dados do mercado abrangeu o estudo e testes de aplicativos baseados em diferentes técnicas, enriquecendo a base de comparação utilizada na avaliação da pesquisa.

A evolução do projeto Bramining consistiu no aprimoramento do ambiente de KDD desenvolvido em estudos anteriores, passando a incluir a técnica de Rough Sets em seu escopo.

Os estudos de caso foram conduzidos paralelamente com o uso do Bramining e de outras ferramentas existentes, para efeito de comparação.

Os índices apresentados pelo Bramining nos estudos de caso foram considerados, de forma geral, equivalentes aos do software comercial, tendo ambos obtido regras de boa qualidade na maioria dos casos. O Bramining, entretanto, mostrou-se mais completo para o processo de KDD, graças às diversas opções nele disponíveis para preparação dos dados antes da fase de mineração.

Os resultados obtidos comprovaram, através da aplicação desenvolvida, a adequação dos conceitos de Rough Sets à tarefa de classificação de dados. Alguns pontos frágeis da técnica foram identificados, como a necessidade de um mecanismo de apoio para a redução de atributos e a dificuldade em trabalhar com atributos de domínio contínuo. Porém, ao se inserir a técnica em um ambiente mais completo de KDD, como o Bramining, estas deficiências foram sanadas. As opções de preparação da base que o Bramining disponibiliza ao usuário para executar, em particular, a redução e a codificação de atributos permitem deixar os dados em estado adequado à aplicação de Rough Sets.

A mineração de dados é uma questão bastante relevante nos dias atuais, e muitos métodos têm sido propostos para as diversas tarefas que dizem respeito a esta questão. A teoria de Rough Sets não mostrou significativas vantagens ou desvantagens em relação a outras técnicas já consagradas, mas foi de grande valia comprovar que há caminhos alternativos para o processo de descoberta de conhecimento.

ABSTRACT

This dissertation investigates the application of Rough Sets to the process of KDD – Knowledge Discovery in Databases. The main goal of the work was to evaluate the performance of Rough Sets techniques in solving the classification problem. Classification is a task of the Data Mining step in KDD process that performs the discovery of decision rules that best represent a group of registers in a database. The work had five major steps: study of the KDD process; study of Rough Sets techniques applied to data mining; evaluation of existing data mining tools; development of Bramining project; and execution of some case studies to evaluate Bramining.

The study of KDD process included all its steps: transformation, cleaning, selection, data mining and post-processing. The results obtained served as a basis to the enhancement of Bramining.

The study of Rough Sets techniques included the research of theory's concepts and its applicability at KDD context. The Rough Sets theory has been introduced by Zdzislaw Pawlak in the early 80's as a mathematical approach to the analysis of vague and uncertain data. This research made possible the implementation of the technique under the environment of the developed tool.

The analysis of existing data mining tools included studying and testing of software based on different techniques, enriching the background used in the evaluation of the research.

The evolution of Bramining Project consisted in the enhancement of the KDD environment developed in previous works, including the addition of Rough Sets techniques.

The case studies were performed simultaneously with Bramining and a commercial mining tool, for comparison reasons.

The quality of the knowledge generated by Bramining was considered equivalent to the results of the commercial tool, both providing good decision rules for most of the cases. Nevertheless, Bramining proved to be more adapted to the complete KDD process, thanks to the many available features to prepare data to data mining step.

The results achieved through the developed application proved the suitability of Rough Sets concepts to the data classification task. Some weaknesses of the technique were identified, like the need of a previous attribute reduction and the inability to deal with continuous domain data. But as the technique has been inserted in a more complete KDD environment like the Bramining Project, those weaknesses ceased to exist. The features of data preparation available in Bramining environment, particularly the reduction and attribute codification options, enable the user to have the database fairly adapted to the use of Rough Sets algorithms.

Data mining is a very relevant issue in present days and many methods have been proposed to the different tasks involved in it. Compared to other techniques, Rough Sets Theory did not bring significant advantages or disadvantages to the process, but it has been of great value to show there are alternate ways to knowledge discovery.

ÍNDICE

ÍNDICE	III
LISTA DE ILUSTRAÇÕES	VI
LISTA DE TABELAS	VII
1 INTRODUÇÃO.....	1
1.1 MOTIVAÇÃO	1
1.2 OBJETIVOS DO TRABALHO	2
1.3 DESCRIÇÃO DO TRABALHO.....	3
1.4 ORGANIZAÇÃO DA DISSERTAÇÃO	3
2 KDD - KNOWLEDGE DISCOVERY IN DATABASES	5
2.1 CONCEITUAÇÃO DE MINERAÇÃO DE DADOS.....	5
2.2 AS FASES DO PROCESSO DE KDD.....	8
2.2.1 A preparação dos dados.....	13
2.2.1.1 Transformação dos dados.....	14
2.2.1.2 Limpeza dos dados.....	16
2.2.1.3 Redução de atributos.....	19
2.2.1.4 Pré-processamento dos dados	22
2.2.2 O processo de mineração dos dados ("Data Mining")	25
2.2.2.1 Objetivos primários da mineração de dados	25
2.2.2.2 Tarefas primárias da mineração de dados	26
2.2.2.2.1 <i>Classificação</i>	26
2.2.2.2.2 <i>Regressão</i>	27
2.2.2.2.3 <i>Clusterização</i>	28
2.2.2.3 Seleção da tarefa de mineração de dados	30
2.2.2.4 Técnicas de inferência.....	32
2.2.2.5 Métodos de mineração de dados	34
2.2.2.5.1 <i>Métodos estatísticos</i>	34
2.2.2.5.2 <i>Métodos lineares</i>	35
2.2.2.5.3 <i>Método das árvores de decisão e regras</i>	35
2.2.2.5.4 <i>Métodos de aprendizagem relacional</i>	37
2.2.2.5.5 <i>Método das redes neurais</i>	37
2.2.2.5.6 <i>Método dos algoritmos genéticos</i>	38
2.2.2.5.7 <i>Considerações finais</i>	40
2.3 CLASSIFICAÇÃO EM DETALHES.....	41
3 ROUGH SETS	43
3.1 BASE DE DADOS E CONCEITOS INICIAIS.....	43
3.2 APROXIMAÇÕES	45
3.3 DEPENDÊNCIA DE ATRIBUTOS.....	47
3.4 REDUÇÃO DE ATRIBUTOS.....	48

3.5	SIGNIFICÂNCIA DE ATRIBUTOS	48
3.6	REGRAS DE DECISÃO	49
4	AVALIAÇÃO DE FERRAMENTAS	51
4.1	WIZRULE.....	51
4.1.1	Relatórios.....	52
4.1.1.1	Relatório de Regras.....	52
4.1.1.2	Relatório de Pronúncia.....	53
4.1.1.3	Relatório de Desvios.....	53
4.1.2	Tipos de regras.....	54
4.1.3	Facilidade de uso.....	55
4.1.4	Informações sobre o banco de dados.....	55
4.1.5	Conclusões.....	56
4.2	WIZWHY E WIZWHY PREDICTOR.....	56
4.2.1	Relatórios.....	56
4.2.1.1	Relatório de Regras.....	56
4.2.1.2	Relatório de Predições.....	57
4.2.2	Facilidade de uso.....	58
4.2.3	Informações sobre o banco de dados.....	58
4.2.4	Conclusões.....	58
4.3	XPERTRULE MINER.....	59
4.3.1	Aspectos funcionais do software.....	59
4.4	BUSINESSMINER.....	61
4.4.1	Módulos do software.....	61
4.5	POLYANALYST.....	62
4.5.1	Facilidade de uso.....	62
4.5.2	Informações sobre o banco de dados.....	63
4.5.3	Conclusões.....	68
4.6	CONCLUSÕES SOBRE OS SOFTWARES AVALIADOS	69
5	PROJETO BRAMINING.....	70
5.1	INTRODUÇÃO.....	70
5.2	A IMPLEMENTAÇÃO.....	71
5.3	UTILIZANDO A FERRAMENTA.....	72
5.3.1	Seleção de dados.....	74
5.3.2	Pré-processamento dos dados.....	77
5.3.3	Codificação.....	85
5.3.4	Mineração de dados.....	89
5.3.5	Relatório.....	93
6	ESTUDOS DE CASO	95
6.1	INTRODUÇÃO.....	95
6.2	VESTIBULAR PUC.....	96
6.2.1	FORMA DE APRESENTAÇÃO DOS RESULTADOS.....	99
6.2.1.1	Tabelas de Regras.....	99
6.2.1.2	Tabelas de coeficientes.....	100
6.2.2	Resultados do WizRule.....	101

6.2.3	Resultados do Bramining	103
6.3	DADOS DE ANIMAIS.....	105
6.3.1	Resultados do WizRule	106
6.3.2	Resultados do Bramining	107
6.4	USUÁRIOS DE COMPANHIA TELEFÔNICA.....	108
6.4.1	Resultados do WizRule	110
6.4.2	Resultados do Bramining	111
6.5	CONSIDERAÇÕES SOBRE OS RESULTADOS.....	113
7	CONCLUSÕES E TRABALHOS FUTUROS	116
7.1	CONCLUSÕES.....	116
7.2	TRABALHOS FUTUROS.....	117
	APÊNDICE A - WIZRULE PASSO A PASSO	118
	APÊNDICE B - WIZWHY PASSO A PASSO.....	122
	APÊNDICE C - XPERTRULE MINER PASSO A PASSO.....	126
	APÊNDICE D - BUSINESSMINER PASSO A PASSO	129
	REFERÊNCIAS BIBLIOGRÁFICAS	140

LISTA DE ILUSTRAÇÕES

Figura 2.1 - Tentativa de classificação do conjunto de dados de empréstimos.....	7
Figura 2.2 - Uma visão geral dos passos que compõem o processo de KDD.....	9
Figura 2.3 - Classificação linear limitada pelo conjunto de dados de empréstimo.....	27
Figura 2.4 - Regressão para o conjunto de dados de empréstimos.....	28
Figura 2.5 - Divisão do conjunto de dados de empréstimos em 3 grupos.....	29
Figura 2.6 - Previsão de série temporal.....	32
Figura 3.1 - As aproximações em forma gráfica.....	46
Figura 4.1 - Exemplo da visualização do conteúdo de um dataset.....	63
Figura 4.2 - Estrutura do dataset.....	64
Figura 4.3 - Estatísticas da base de dados.....	64
Figura 4.4 - Análise de dependência entre atributos.....	66
Figura 4.5 - Resultado da Regressão Linear.....	67
Figura 4.6 - Módulo de geração de gráficos.....	68
Figura 5.1 - DFD do ambiente de KDD desenvolvido.....	72
Figura 5.2 - Tela inicial da ferramenta desenvolvida para KDD.....	73
Figura 5.3 - Tela indicando a seleção de dados.....	75
Figura 5.4 - Representação da base de dados selecionada na Figura 5.3.....	76
Figura 5.5 - Estatísticas acerca da base de dados.....	78
Figura 5.6 - Redução dos dados e geração de uma nova base.....	79
Figura 5.7 - Estatísticas acerca da nova base de dados.....	80
Figura 5.8 - Primeira rodada de substituições e deleções de atributos da base de dados.....	81
Figura 5.9 - Indicação das substituições do valor Com por C.....	81
Figura 5.10 - Segunda rodada de substituições e deleções da base de dados.....	82
Figura 5.11 - Indicação da substituição do valor Res por R.....	82
Figura 5.12 - Exclusão do valor nulo do atributo Faixa_Tarifação.....	83
Figura 5.13 - Indicação do número de registros deletados.....	84
Figura 5.14 - Exclusão do valor 12 do atributo Tipo_Ligação.....	84
Figura 5.15 - Indicação do número de registros deletados.....	84
Figura 5.16 - Exclusão do valor 15 do atributo Tipo_Ligação.....	85
Figura 5.17 - Indicação do número de registros deletados.....	85
Figura 5.18 - Codificação do atributo Minutos.....	86
Figura 5.19 - Codificação do atributo Fim_de_semana.....	87
Figura 5.20 - Estatísticas da base de dados após a Preparação.....	88
Figura 5.21 - Menu de escolha dos softwares de mineração de dados.....	90
Figura 5.22 - Tela para definição de parâmetros da fase de mineração de dados.....	91
Figura 5.23 - Tela para definição de parâmetros complementares de Rough Sets.....	92
Figura 5.24 - Tabelas para acompanhamento da execução do algoritmo.....	93
Figura 5.25 - Resultados obtidos pela mineração de dados, na forma de relatório.....	94

LISTA DE TABELAS

Tabela 2.1 - Dados de empréstimos.	6
Tabela 2.2 - Dados de empréstimos em estado original.....	12
Tabela 2.3 - Dados de empréstimos após seleção de atributos.	15
Tabela 2.4 - Dados de empréstimos após composição de atributos.....	16
Tabela 2.5 - Dados de empréstimos após limpeza.	19
Tabela 2.6 - Dados de empréstimos após redução de atributos.....	22
Tabela 2.7 - Dados de empréstimos após codificação.....	24
Tabela 2.8 - Tarefas da mineração de dados, seus algoritmos e aplicações.....	30
Tabela 2.9 - Dados de cesta de mercado.	31
Tabela 2.10 - Dados de empréstimos.	31
Tabela 2.11 - Dados de empréstimos clusterizados.	32
Tabela 2.12 - Informações de lojas de um bairro.....	41
Tabela 5.1 - Atributos contidos na base de dados de telecomunicações.....	76
Tabela 6.1 - Descrição da base Vestibular da PUC.....	97
Tabela 6.2 - Exemplos da base de vestibulandos da PUC.....	98
Tabela 6.3 - Exemplo de regras geradas pelo WizRule.....	99
Tabela 6.4 - Exemplo de regras geradas pelo Bramining.....	100
Tabela 6.5 - Exemplo de tabela de coeficientes de regras.....	100
Tabela 6.6 - Regras do WizRule. Base: Vestibular da PUC.....	102
Tabela 6.7 - Coeficientes das regras do WizRule. Base: Vestibular da PUC.....	103
Tabela 6.8 - Regras do Bramining. Base: Vestibular da PUC.....	104
Tabela 6.9 - Coeficientes das regras do Bramining. Base: Vestibular da PUC.....	104
Tabela 6.10 - Descrição da base de dados de animais.....	105
Tabela 6.11 - Amostra da base de dados de animais.....	106
Tabela 6.12 - Regras do WizRule. Base: Animais.....	107
Tabela 6.13 - Coeficientes das regras do WizRule. Base: Animais.....	107
Tabela 6.14 - Regras do Bramining. Base: Animais.....	108
Tabela 6.15 - Coeficientes das regras do Bramining. Base: Animais.....	108
Tabela 6.16 - Descrição da base de dados de tráfego telefônico.....	109
Tabela 6.17 - Amostra da base de dados de tráfego telefônico.....	110
Tabela 6.18 - Regras do WizRule. Base: Tráfego telefônico.....	111
Tabela 6.19 - Coeficientes das regras do WizRule. Base: Tráfego telefônico.....	111
Tabela 6.20 - Regras do Bramining. Base: Tráfego telefônico.....	112
Tabela 6.21 - Coeficientes das regras do Bramining. Base: Tráfego telefônico.....	112
Tabela 6.22 - Exemplo de comparação de qualidade de regras.....	113
Tabela 6.23 - Tabela comparativa de resultados.....	115

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

O grande acúmulo de dados em todas as áreas de atividade humana criou a necessidade urgente de uma nova geração de ferramentas e técnicas automáticas e inteligentes para analisar, resumir e extrair conhecimento a partir de dados brutos. O processo de extração não trivial de informações úteis dos dados que estejam implícitas e sejam potencialmente úteis aos seus usuários é conhecido como KDD (Knowledge Discovery in Databases).

KDD não é uma nova técnica, mas um campo multidisciplinar de pesquisa que abrange tradicionalmente o uso de técnicas como: estatística, machine learning, redes neurais, algoritmos genéticos, tecnologias de banco de dados, sistemas especialistas e visualização de dados.

O processo de KDD consiste basicamente dos seguintes estágios:

- Seleção dos dados;
- Pré-processamento;
- Mineração de dados (Data Mining);
- Pós-processamento.

Merece particular destaque a fase de Mineração de Dados, que pode ser executada na forma de uma dentre diversas tarefas, como: geração de regras de classificação, árvores de decisão, regressão e clusterização, entre outras.

Atualmente, há uma diversidade de técnicas utilizadas para a mineração de dados, algumas delas tradicionais, como os métodos estatísticos, e outras que se inserem na área de Inteligência Computacional, como Algoritmos Genéticos e Redes Neurais. A proposta do

presente trabalho é estudar e implementar a técnica de Rough Sets na fase de Mineração de Dados. A parte de implementação será feita dentro do Projeto Bramining, compondo a terceira tese de construção de um ambiente de descoberta de conhecimento desenvolvido no Brasil [18] [21]. Este ambiente abrange diversas fases do processo de KDD e disponibiliza variadas técnicas para a fase de mineração propriamente dita.

A teoria de Rough Sets foi apresentada por Zdzislaw Pawlak no início dos anos 80 como uma abordagem matemática para a análise de dados vagos e imprecisos. O ponto de partida desta teoria é a constatação de que objetos podem ser indiscerníveis (no sentido de similares ou indistinguíveis), devido à limitada informação disponível sobre eles. Neste contexto, objetos que não podem ser especificados através dos dados disponíveis são caracterizados pela teoria de Rough Sets através de dois conceitos precisos: a aproximação inferior e a aproximação superior.

A fundamentação matemática desta teoria permite a descoberta de padrões escondidos na base de dados. Sua utilidade no campo de Mineração de Dados pode ser comprovada pelo crescente número de aplicações e publicações científicas divulgadas com este conteúdo.

1.2 OBJETIVOS DO TRABALHO

Os principais objetivos desta pesquisa incluem:

- Identificar um modelo de aplicação de Rough Sets em Mineração de Dados
- Desenvolver um sistema de descoberta de conhecimento utilizando esta técnica
- Investigar o desempenho da técnica de Rough Sets em Mineração de Dados
- Comparar os resultados obtidos com os de outras técnicas existentes.

1.3 DESCRIÇÃO DO TRABALHO

Esta pesquisa foi elaborada nas seguintes etapas:

- Estudo do processo de KDD;
- Estudo sobre as técnicas de Rough Sets aplicadas à mineração de dados;
- Análise de ferramentas de mineração de dados do mercado;
- Evolução do projeto de um ambiente de descoberta de conhecimento com Rough Sets;
- Realização de alguns estudos de caso para avaliar o Bramining
- Redação da Dissertação

1.4 ORGANIZAÇÃO DA DISSERTAÇÃO

Esta dissertação está dividida em seis capítulos adicionais, descritos a seguir:

O capítulo 2 descreve o processo de descoberta de conhecimento, especificando as tarefas principais e as fases envolvidas em KDD.

O capítulo 3 realiza um estudo sobre as técnicas de Rough Sets, que envolveu a pesquisa de seus conceitos e sua aplicabilidade no contexto de KDD.

O capítulo 4 engloba a análise de ferramentas de mineração de dados do mercado, que abrangeu o estudo e testes de aplicativos baseados em diferentes técnicas.

O capítulo 5 ilustra o projeto Bramining e seu aprimoramento, passando a incluir a técnica de Rough Sets em seu escopo.

O capítulo 6 mostra estudos de caso que foram conduzidos paralelamente com o uso do Bramining e de outras ferramentas existentes, para efeito de comparação.

O capítulo 7 aborda as conclusões e propostas de trabalhos futuros.

2 KDD - KNOWLEDGE DISCOVERY IN DATABASES

2.1 CONCEITUAÇÃO DE MINERAÇÃO DE DADOS

Descoberta de conhecimento em bases de dados - KDD - é um processo não trivial de identificação válida de padrões nos dados. É um processo novo, potencialmente útil e fundamentalmente compreensível. [10]

Para melhor entendimento da definição acima, cada um dos termos relevantes será examinado mais detalhadamente. Antes de mais nada, dados e padrões devem ser definidos:

- **Dados:** é o conjunto de fatos (instâncias do banco de dados). Utilizando o exemplo apresentado na Tabela 2.1 e, graficamente, na Figura 2.1, **F** é a coleção de 14 amostras com quatro atributos que contêm: a identificação do cliente, o valor do seu salário, o débito e o estado do empréstimo;

CLIENTE	SALÁRIO	DÉBITO	NEGLIGENTE
1	1200,00	800,00	SIM
2	1300,00	650,00	SIM
3	1350,00	450,00	SIM
4	1350,00	820,00	SIM
5	1400,00	600,00	NÃO
6	1600,00	200,00	NÃO
7	1600,00	800,00	SIM
8	1650,00	500,00	SIM
9	1850,00	650,00	SIM

10	1900,00	500,00	NÃO
11	2200,00	400,00	NÃO
12	2200,00	550,00	NÃO
13	2250,00	700,00	NÃO
14	2300,00	520,00	NÃO

Tabela 2.1 - Dados de empréstimos.

- Padrão: É um subconjunto utilizado como termo de comparação numa tomada de decisão. Considerando o exemplo da Tabela 2.1, o padrão “*Se salário > t, então o mutuário é bom pagador*” poderá proporcionar uma escolha apropriada de t ;
- Processo: O processo de KDD é não trivial e multi-step. Não trivial porque se refere a processos com algum grau de busca autônoma. Multi-step porque se refere aos diversos passos que compõem o processo de KDD. Os passos estão divididos em duas grandes fases: preparação de dados e mineração de dados. Todos estes passos permitem a iteração após sua execução. No exemplo apresentado na Tabela 2.1, somente quando o “significado” do salário das pessoas produzir um resultado útil será qualificado como descoberta;
- Válida (Validação): Uma vez descobertos os padrões existentes em um banco de dados, estes serão válidos para os novos dados com um determinado grau de certeza. Voltando ao exemplo, se o limite do padrão mostrado na Figura 2.1 é movido para a direita, sua medida de certeza irá aumentar desde que o número de não negligentes apontados pelo padrão seja garantido;
- Novo: Uma vez que a novidade pode ser medida com relação às mudanças nos dados ou no conhecimento, pode-se considerar que os padrões são recentes. Esta mudança é dada

KDD é a descoberta de novos conhecimentos, sejam padrões, tendências, associações, probabilidades ou fatos que não são óbvios ou de fácil identificação.

O processo de KDD é o processo de utilização dos métodos de mineração de dados para extrair conhecimentos de acordo com a especificação de medidas e limites, aplicados a bases de dados submetidas a um pré-processamento.

2.2 AS FASES DO PROCESSO DE KDD

Como já foi dito anteriormente, o processo de KDD envolve duas grandes fases, a saber: preparação de dados e mineração de dados. Estas fases possuem diversos passos, que envolvem um grande número de decisões a serem tomadas pelo usuário, ou seja, é um processo interativo. É também um processo iterativo, porque ao longo do processo de KDD, um passo será repetido tantas vezes quanto necessário para que se chegue a um resultado satisfatório.

Os passos principais do processo de KDD são apresentados na Figura 2.2 e são descritas sumariamente a seguir:

1. O primeiro passo é a definição do objetivo do problema, que consiste no conhecimento desejado pelo usuário final, ou seja, é definido o tipo de conhecimento que se deseja extrair do banco de dados. Nesta fase é feito um reconhecimento da aplicação e a verificação do conhecimento anterior;

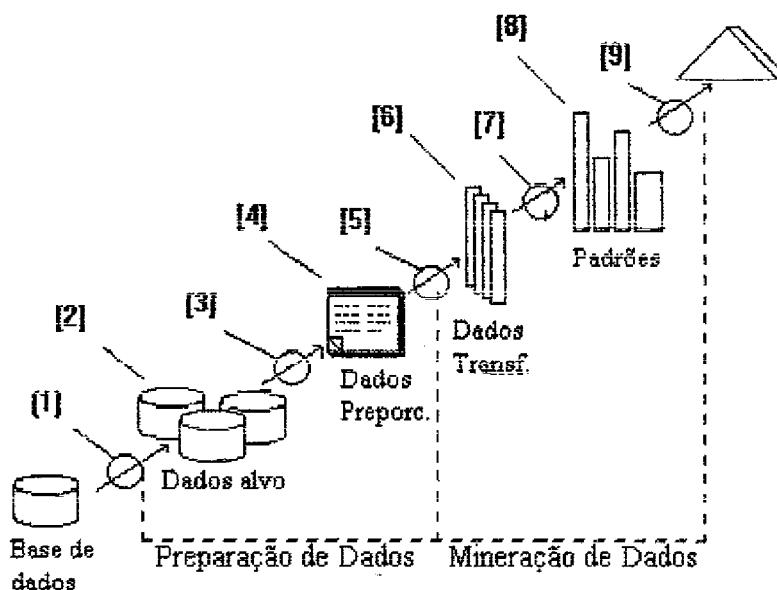


Figura 2.2 - Uma visão geral dos passos que compõem o processo de KDD.

2. O segundo passo é a criação de um conjunto de dados-alvo. Nesta fase seleciona-se um conjunto de dados, ou focaliza-se um subconjunto de atributos e instâncias de dados, onde a descoberta deverá ser efetuada. Muitas vezes o sucesso deste processo depende da correta escolha dos dados que formam o conjunto de dados-alvo. Para isto são usadas técnicas, linguagens, ferramentas e comandos convencionais de bancos de dados, como SQL;
3. O terceiro passo é a limpeza dos dados. Neste caso deve ser feita a limpeza dos dados de maneira que os incorretos ou incompletos sejam corrigidos ou desprezados. Com isto é feita uma purificação dos dados utilizando operações básicas, como as de eliminação do ruído. Nela são coletadas as informações necessárias para modelagem e correção do ruído e para estratégias de manipulação de campos de dados perdidos, considerando as seqüências de informações de tempo e mudanças de conhecimento;

4. O quarto passo é a redução e projeção de dados. Consiste em encontrar as características úteis que representam as dependências dos dados no objetivo do processo. Muitas vezes pode não ser necessário representar todas as faixas de valores de um determinado problema. Assim, pode-se reagrupar estes valores em faixas mais abrangentes, de modo a diminuir o número de faixas de valores e conseqüentemente a complexidade do problema;
5. O quinto passo é a escolha das tarefas de mineração de dados. Neste passo decide-se qual o objetivo do processo de mineração de dados. Os objetivos podem ser os mais diversos tais como: classificação, regressão, clusterização, etc.;
6. O sexto passo é a escolha dos algoritmos de mineração de dados. Nele são selecionados os métodos para serem usados na busca de padrões dos dados. Isto inclui decidir que modelos e parâmetros são mais apropriados para a aquisição do tipo de conhecimento desejado. Através da submissão dos dados aos algoritmos de mineração de dados selecionados, chega-se ao conhecimento. Neste caso pode-se escolher algoritmos como redes neurais, algoritmos genéticos, etc. Estes passos, se utilizados corretamente, serão de grande ajuda para a etapa seguinte;
7. O sétimo passo é a mineração de dados. Caracterizado pela busca de padrões de interesse numa forma representativa em particular ou um conjunto destas representações. Como exemplos pode-se citar: regras de classificação, árvores de decisão, regressão, clusterização e outros. Neste passo é realizada a extração de informações dos dados até então processados;

8. O oitavo passo é a interpretação de padrões da mineração. Os dados de saída definidos no passo anterior são analisados e interpretados pelos especialistas do domínio. Caso seja necessário, pode-se repetir qualquer um dos 7 passos anteriores para se obter a correta interpretação dos padrões;
9. O nono passo é a consolidação do conhecimento descoberto. É a incorporação deste conhecimento no desempenho do sistema, a documentação do conhecimento e o relatório para as partes interessadas. Neste passo também é feita a verificação e a resolução de conflitos potenciais com o conhecimento extraído previamente.

O processo de KDD pode envolver interações significativas e pode retornar a qualquer dos passos, independentemente da fase em que se encontre. Apesar da seqüência apresentada pela Figura 2.2 ser a mais comum, esta pode ser mudada. O maior trabalho está situado nos passos de 1 a 4 (entre 50 e 80 % do trabalho), quando os dados estão sendo preparados para o processo de mineração, embora as demais sejam de igual importância para o sucesso da aplicação como um todo.

Com relação à eficiência deste processo, ela não deve ser medida em termos da rapidez do processamento das conclusões. Para um processo de KDD ser considerado eficiente, o valor da informação encontrada deverá exceder o alto custo da sua implementação e do processamento dos dados. Pode-se dizer que a eficiência do processo reside em uma relação custo-benefício elevada. Geralmente os resultados de um processo de KDD têm alto valor estratégico.

A seguir, serão descritas em detalhes as fases do processo. Para ilustrar as ações, a Tabela 2.2 exibe uma base de dados que será refinada a cada passo do processo de KDD, para ao final atingir o formato da Tabela 2.1.

CLIENTE	SALÁRIO BRUTO	ENCARGOS PADRÃO (R\$)	DÉBITO (R\$)	SALÁRIO ANUAL	NOME	IDENTIDADE	LOGRADOURO	SITUAÇÃO PAGAMENTO
1	R\$1.500	300	800	R\$19.500	Juliane Dornelas	46456	Rua Bartholomeu Ferrare 749 ap 53 B	ATRASO5P
2	R\$1.650	350	650	R\$21.450	Jussara Fontes	5663656	R. Sto Inácio N.130	ATRASO1P
3	R\$1.750	400	450	R\$22.750	Jussara Mattos	5646336	R Santa Catarina, 189	ATRASO2P
4	R\$1.850	500	820	R\$24.050	Jussara Noronha	9698	QSF 04 Casa 423	ATRASO3P
5	US\$900	400	600	US\$11.700	Karina Macedo	225778567	R 26 de dezembro, 26	EM DIA
6	US\$1.050	500	200	US\$13.650	Kátia Castilho	746754	R Barão de Jaguarí 279 Ap. 202 Irajá 7563850 R212 trab	EM DIA
7	R\$2.000	400	800	R\$26.000	Kátia de Matos	675674	QI-18 Conjunto-U Casa 115	ATRASO2P
8	R\$2.300	650	500	R\$29.900	Kátia Moraes	435635	Rua Felipe Camarão, 674	ATRASO1P
9	R\$2.300	450	650	R\$29.900	Kátia Andrade	67547764	R Maestro Inácio Stábile, 756	ATRASO4P
10	US\$1.300	700	500	US\$16.900	Kátia da Costa	56747	Rua Afonso Pena, 82/801	EM DIA
11	R\$2.900	700	400	R\$37.700	Kátia da Silva	5647457	Rua Monte Libano, 58 apto. 101	EM DIA
12	R\$3.000	800	550	R\$39.000	Kamila Pereira	457676	Rua A, 57/1405	EM DIA
13	US\$1.500	750	700	US\$19.500	Leticia Costa	5674747	SQS 412 Bloco R Ap 302	EM DIA
14	R\$3.000	700	520	R\$39.000	Leticia Gomes	56747	r. José Anibal Colleone, 320	EM DIA

Tabela 2.2 - Dados de empréstimos em estado original.

2.2.1 A preparação dos dados

A fase de preparação dos dados absorve uma boa parte do tempo do processo de KDD, consumindo em média 70% do tempo total, além de ser uma fase de grande importância. Nela são identificados os dados relevantes para a solução satisfatória do problema. Na verdade, apenas ter os dados não é suficiente; é necessário que eles estejam suficientemente corretos, adequados e tenham sido corretamente selecionados para que preencham todas as características desejadas. Mesmo assim, sempre existirá a pergunta: os dados existentes preenchem estas características?

Outro fator importante é o tipo de armazenamento de dados utilizado. No caso de processos informatizados o armazenamento deve ser feito em arquivos ou bancos de dados; no caso de processos não informatizados, em fichas ou anotações. Independentemente do tipo de armazenamento de dados empregado, a sua importância continua sendo a mesma. Para que o processo de KDD obtenha sucesso, é necessário que os dados estejam disponíveis para o processamento e em condições de serem utilizados.

A quantidade de informações disponíveis deverá, sempre que possível, estar em excesso e não em falta, pois uma solução satisfatória exige todas as informações importantes para a solução de um determinado problema. A falta de informação pode dificultar e, em alguns casos, até mesmo impedir que se chegue a um resultado confiável. [18]

Somente a posse dos dados permite avaliar de forma confiável aqueles que são realmente relevantes para a solução do problema em questão. Por isso, sempre que houver dúvida com relação à relevância de uma variável, é aconselhável colocá-la na massa de dados.

A quantidade de dados será definida de acordo com os métodos a serem utilizados durante o processo de KDD.

Outro ponto a ser observado é com relação à quantidade de dados para teste. Numa massa de dados deve-se separar um determinado percentual para ser utilizado como dados de teste. Estes dados devem ser cuidadosamente escolhidos de tal forma que representem todas as situações possíveis de serem encontradas. Em alguns casos também deve ser separado um terceiro conjunto de dados para validação. Estes são utilizados para a realização de um teste preliminar de verificação do desempenho do método de mineração de dados antes de submetê-lo ao teste final.

2.2.1.1 Transformação dos dados

O objetivo central da preparação dos dados para a mineração de dados é a transformação dos dados em um formato padrão. Algumas vezes, os dados presentes em um warehouse já estão no formato padrão, ou o método de predição trabalha diretamente com formatos específicos de dados. Geralmente, entretanto, duas tarefas adicionais estão associadas com a produção de uma planilha padrão:

- Seleção de atributo: alguns atributos redundantes e não-preditivos (colunas da tabela) são apagados. Na Tabela 2.2 nota-se que o atributo "SALÁRIO ANUAL" traz uma informação redundante com o atributo "SALÁRIO", provavelmente sem trazer qualquer benefício particular à predição. Portanto, este atributo será descartado e a base de dados passa a ter o formato da Tabela 2.3.

CLIENTE	SALÁRIO BRUTO	ENCARGOS PADRÃO (R\$)	DÉBITO (R\$)	NOME	IDENTIDADE	LOGRADOURO	SITUAÇÃO PAGAMENTO
1	R\$1.500	300	800	Juliana Dorneias	46456	Rua Bartholomeu Ferrare 749	ATRASO5P

						ap 53 B	
2	R\$1.650	350	650	Jussara Fontes	5663656	R. Sto Inácio N.130	ATRASO1P
3	R\$1.750	400	450	Jussara Mattos	5646336	R Santa Catarina, 189	ATRASO2P
4	R\$1.850	500	820	Jussara Noronha	9698	QSF 04 Casa 423	ATRASO3P
5	US\$900	400	600	Karina Macedo	225778567	R 26 de dezembro, 26	EM DIA
6	US\$1.050	500	200	Kátia Castilho	746754	R Barão de Jaguari 279 Ap 202 Irajá 7563850 R212 trab	EM DIA
7	R\$2.000	400	800	Kátia de Matos	675674	QI-18 Conjunto-U Casa 115	ATRASO2P
8	R\$2.300	650	500	Katia Moraes	435635	Rua Felipe Camarão, 674	ATRASO1P
9	R\$2.300	450	650	Kátia Andrade	67547764	R Maestro Inácio Stábile, 756	ATRASO4P
10	US\$1.300	700	500	Kátia da Costa	56747	Rua Afonso Pena, 82/801	EM DIA
11	R\$2.900	700	400	Katia da Silva	5647457	Rua Monte Líbano, 58 apto. 101	EM DIA
12	R\$3.000	800	550	Kamila Pereira	457676	Rua A, 57/1405	EM DIA
13	US\$1.500	750	700	Leticia Costa	5674747	SQS 412 Bloco R Ap 302	EM DIA
14	R\$3.000	700	520	Leticia Gomes	56747	r. José Anibal Colleone, 320	EM DIA

Tabela 2.3 - Dados de empréstimos após seleção de atributos.

- Composição de atributo: fator determinante para a qualidade dos resultados e, muitas vezes, dependente do conhecimento da aplicação. Para o exemplo da análise de empréstimos, um dado importante para a análise da capacidade de pagamento do cliente é a parcela de seus vencimentos que ele efetivamente tem disponível para gastos extras. Isto pode ser obtido criando-se o atributo "SALÁRIO", derivado da diferença entre "SALÁRIO BRUTO" e "ENCARGOS PADRÃO". O resultado é ilustrado na Tabela 2.4.

CLIENTE	SALÁRIO	DÉBITO (RS)	NOME	IDENTIDADE	LOGRADOURO	SITUAÇÃO PAGAMENTO
1	R\$1.200	800	Juliana Dornelas	46456	Rua Bartholomeu Ferrare 749 ap 53 B	ATRASO5P
2	R\$1.300	650	Jussara Fontes	5663656	R. Sto Inácio N.130	ATRASO1P
3	R\$1.350	450	Jussara Mattos	5646336	R Santa Catarina, 189	ATRASO2P
4	R\$1.350	820	Jussara Noronha	9698	QSF 04 Casa 423	ATRASO3P
5	US\$700	600	Karina Macedo	225778567	R 26 de dezembro, 26	EM DIA
6	US\$800	200	Kátia Castilho	746754	R Barão de Jaguarí 279 Ap 202 Irajá 7563850 R212 trab	EM DIA
7	R\$1.600	800	Kátia de Matos	675674	QI-18 Conjunto-U Casa 115	ATRASO2P
8	R\$1.650	500	Katia Moraes	435635	Rua Felipe Camarão, 674	ATRASO1P
9	R\$1.850	650	Kátia Andrade	67547764	R Maestro Inácio Stábile, 756	ATRASO4P
10	US\$950	500	Kátia da Costa	56747	Rua Afonso Pena, 82/801	EM DIA
11	R\$2.200	400	Katia da Silva	5647457	Rua Monte Líbano, 58 apto. 101	EM DIA
12	R\$2.200	550	Kamila Pereira	457676	Rua A, 57/1405	EM DIA
13	US\$1.125	700	Leticia Costa	5674747	SQS 412 Bloco R Ap 302	EM DIA
14	R\$2.300	520	Leticia Gomes	56747	r. José Anibal Colleone, 320	EM DIA

Tabela 2.4 - Dados de empréstimos após composição de atributos.

2.2.1.2 Limpeza dos dados

Os dados a serem usados como matéria prima para o processo de KDD devem ser tão puros e corretos quanto possível. A existência ou não de ruído nos dados de entrada deve ser cuidadosamente verificada, pois o ruído provoca dois problemas: primeiro quando o conjunto

de treinamento com ruído gera descrições; segundo quando objetos com ruído são classificados usando estas descrições, ou seja, o ruído pode levar o processo a chegar a conclusões mais generalizadas que aquelas possíveis quando se utiliza dados mais puros.

As técnicas mais usadas para limpeza dos dados são:

- Técnicas baseadas em regras [28] ;
- Visualização [2] ;
- Informações estatísticas [2] .

Na realidade, os dados fornecidos pelo cliente sempre têm problemas. Uma vez que a procedência dos dados não é muito confiável, trazendo campos não preenchidos em registros, erros de entrada de dados, entre outros, o processo de KDD pode não ter sucesso sem um esforço para purificá-los. Portanto, o trabalho necessário para colocar estes dados de forma a serem utilizados deve ser considerado.

A partir de pesquisas feitas com analistas de dados, observa-se que o método mais usado para a verificação da acuidade dos dados é a extração do mesmo elemento de dado de múltiplas fontes (quando possível) e a posterior comparação dos resultados. A posse dos critérios usados pelo analista é crucial no resultado da limpeza dos dados a partir de comparações. Em algumas situações é possível também implementar procedimentos de edição automática de dados que fazem a limpeza dos dados antes de carregá-los na base de dados.

A limpeza de dados na realidade é uma espada de dois gumes. Devido à baixa qualidade dos dados, cuidado especial deve ser tomado para não confundir um fenômeno interessante do domínio com uma anomalia ocasional. Em outras palavras, o que parece ser uma anomalia ocasional a ser dispensada, pode vir a ser a chave dos pontos a serem focalizados.

Por exemplo, enquanto os pagamentos das parcelas do empréstimo são analisados, pode-se eliminar a loja recebedora que não recebeu nada, porque as outras lojas receberam muito, desde que se atribua a figura zero ao problema da qualidade dos dados. Por outro lado, não se pode esquecer que em algumas instâncias a loja com recebimento zero pode conduzir a um conhecimento sobre condições que a torne não funcional, embora isto possa ser uma limpeza correta dos dados. Pode-se remover todas as transações vazias de uma base de dados, entretanto registros vazios são cruciais para medir a produtividade do encarregado do caixa, e, algumas vezes, ajudam a detectar as fraudes.

A limpeza de dados é também um problema de sintomas recorrentes, se alguns processos básicos da arrecadação forem falhos. No caso do uso de dados estatísticos não atualizados, e de utilizar um processo único de limpeza, certamente serão encontrados problemas continuamente na qualidade dos dados.

Sendo assim, pode-se dizer que, o que parece um simples exercício de mineração de dados na superfície, pode ser realmente o estímulo para uma revisão organizacional e da infraestrutura na produção de dados, que poderão ser coletados e analisados com confiança.

No estado em que a base de empréstimos se encontra, ilustrado na Tabela 2.4, o atributo "SALÁRIO" apresenta um típico caso de necessidade de limpeza: a heterogeneidade de unidades. No caso, entre os registros com salário em Reais há registros com salários cotados em dólar americano, provavelmente devido a uma regra de negócio qualquer implementada no sistema que forneceu os dados. Para otimizar o trabalho de KDD, entretanto, é fundamental que se faça a padronização dos valores, para evitar comparações indevidas ou a necessidade de incluir fórmulas que atrasarão o processo. Os valores de "SALÁRIO" serão todos convertidos para Reais, resultando na Tabela 2.5.

CLIENTE	SALÁRIO (RS)	DÉBITO (RS)	NOME	IDENTIDADE	LOGRADOURO	SITUAÇÃO PAGAMENTO
1	1.200	800	Juliana Dornelas	46456	Rua Bartholomeu Ferrare 749 ap 53 B	ATRASO5P
2	1.300	650	Jussara Fontes	5663656	R. Sto Inácio N.130	ATRASO1P
3	1.350	450	Jussara Mattos	5646336	R Santa Catarina, 189	ATRASO2P
4	1.350	820	Jussara Noronha	9698	QSF 04 Casa 423	ATRASO3P
5	1.400	600	Karina Macedo	225778567	R 26 de dezembro, 26	EM DIA
6	1600	200	Kátia Castilho	746754	R Barão de Jaguari 279 Ap 202 Irajá 7563850 R212 trab	EM DIA
7	1.600	800	Kátia de Matos	675674	QI-18 Conjunto-U Casa 115	ATRASO2P
8	1.650	500	Katia Moraes	435635	Rua Felipe Camarão, 674	ATRASO1P
9	1.850	650	Kátia Andrade	67547764	R Maestro Inácio Stábile, 756	ATRASO4P
10	1.900	500	Kátia da Costa	56747	Rua Afonso Pena, 82/801	EM DIA
11	2.200	400	Katia da Silva	5647457	Rua Monte Líbano, 58 apto. 101	EM DIA
12	2.200	550	Kamila Pereira	457676	Rua A, 57/1405	EM DIA
13	2.250	700	Letícia Costa	5674747	SQS 412 Bloco R Ap 302	EM DIA
14	2.300	520	Letícia Gomes	56747	r. José Anibal Colleone, 320	EM DIA

Tabela 2.5 - Dados de empréstimos após limpeza.

2.2.1.3 Redução de atributos

Uma vez preparados e transformados os dados para um formato padrão, a expectativa para a mineração de dados é bastante grande. Para uma quantidade de dados moderada, a tabela já está pronta para se fazer a mineração, mas para uma grande amostra, existe um passo

intermediário, a redução de atributos, que deve ser seguido antes da aplicação de programas de predição.

Muitas vezes, a base de dados utilizada tem 100 campos, mas apenas 10 destes são usados para uma decisão. O processo de seleção dos dados a serem utilizados no processo KDD deve ser feito de forma que somente os dados não relevantes sejam retirados. É muito importante a correta identificação dos dados relevantes. No caso de uma incorreção no decorrer do processo, os dados descartados podem se fazer necessários para o sucesso deste [2].

O objetivo da redução é encontrar um subconjunto de atributos com desempenho preditivo comparável ao conjunto original. Dado um conjunto de atributos, o número de subconjuntos a serem analisados é finito e o procedimento que procura encontrar a solução ótima é feito sempre com base no conjunto original. Os resultados são avaliados e o subconjunto de atributos com o melhor desempenho é selecionado. Entretanto, há dificuldades óbvias nesta abordagem:

- para uma grande quantidade de atributos, o número de subconjuntos que pode ser enumerado é algo impossível de ser manipulado;
- o padrão de uma avaliação é o erro encontrado. Para uma amostra vasta, a maioria dos métodos de predição leva muito tempo para encontrar uma solução e estimar o erro.

Simplificações são feitas para produzir resultados práticos em termos de tempo de processamento, aceitáveis. As aproximações para a abordagem ótima que podem ser feitas são as seguintes:

- examinar somente os subconjuntos promissores;
- substituir simples medições computacionais de distância por medições de erro;
- usar as medições de treinamento de desempenho, e não as de teste.

Diversas abordagens têm sido descritas para filtrar ou transformar atributos em um conjunto menor. Os métodos lógicos apresentam uma perspectiva inovadora para a seleção de atributos, constituindo-se num processo dinâmico e coordenado pela busca de soluções. Estes particionam os dados em grupos de casos menores, examinando os valores dos atributos, porém em ordem não-randômica. As duas tarefas essenciais para o desempenho destes métodos são:

- Ordenar os valores;
- Analisar o erro para cada valor.

Com uma quantidade de dados moderada, a ordenação de valores não se trata de um processo tão complexo. Já com o uso de uma amostra maior, a ordenação destes valores torna-se uma tarefa penosa. Os outros métodos, entretanto, são estatísticos, pois o aprendizado é feito separadamente da seleção. Um método inteligente que tem demonstrado grande potencial para grandes espaços de busca é o de Algoritmos Genéticos, e sua utilidade na tarefa de redução merece especial atenção.

Muitas vezes, porém, existem diversos atributos que são visivelmente inexpressivos para o conhecimento que se busca, podendo ser removidos da base a partir de simples visualização. No caso da Tabela 2.5, os atributos "NOME" e "IDENTIDADE" e são claramente irrelevantes para a análise de empréstimos, partindo-se do princípio de que não há

histórico de outros empréstimos e os clientes não se repetem na tabela. O atributo "LOGRADOURO" da mesma forma, neste caso, não possui utilidade. Feita a redução, chega-se ao formato da Tabela 2.6.

CLIENTE	SALÁRIO (RS)	DÉBITO (RS)	SITUAÇÃO PAGAMENTO
1	1.200	800	ATRASOSP
2	1.300	650	ATRASO1P
3	1.350	450	ATRASO2P
4	1.350	820	ATRASO3P
5	1.400	600	EM DIA
6	1600	200	EM DIA
7	1.600	800	ATRASO2P
8	1.650	500	ATRASO1P
9	1.850	650	ATRASO4P
10	1.900	500	EM DIA
11	2.200	400	EM DIA
12	2.200	550	EM DIA
13	2.250	700	EM DIA
14	2.300	520	EM DIA

Tabela 2.6 - Dados de empréstimos após redução de atributos.

2.2.1.4 Pré-processamento dos dados

Nesta fase identifica-se os atributos computados, através do processamento de transações projetadas para lidar com a quantidade mínima de dados requeridos na transação. Atributos computados são aqueles atributos que não estão presentes na base de dados com a qual se trabalha, mas são formados a partir de outros atributos existentes na rede. Usualmente aparecem na relação entre dois valores quaisquer, pode surgir como:

- Somatório;
- Produto;

- Diferença entre dois valores;
- Mapeamento entre valores

As técnicas usadas no pré-processamento dos dados são [2] :

- Gradação: que é o tipo de pré-processamento usado com dados para redes neurais. No caso de valores discretos os dados podem ser codificados com valores 0 e 1 ou podem ser distribuídos nos limites contínuos desejados. No caso de valores esparsos, uma função de regressão linear ou logarítmica pode ser usada. Com os dados sendo submetidos a estas funções eles podem ser reduzidos a valores dentro do limite (“range”) desejado;
- Normalização: usa-se um vetor ou matriz de dados numéricos como um grupo de dados. Para que isto seja possível, eles devem estar normalizados. Os métodos de normalização mais comuns são:
 - Norma Euclidiana;
 - Razão entre os valores e o valor máximo; etc.
- Mapeamento simbólico e taxonomias: neste caso os símbolos são transformados em outros símbolos antes de, opcionalmente, transformá-los em valores numéricos, ou agrupar membros de algumas classes ou grupo em um único símbolo para a sua representação. A transformação de símbolos em números pode ser necessária para transformar símbolos discretos em valores numéricos, para o posterior processamento de determinados algoritmos de mineração de dados que trabalham preferencialmente com

números. Para os algoritmos que trabalham bem com dados discretos ou categóricos, isto não se faz necessário. Neste caso, o mapeamento muitas vezes é chamado de Codificação.

- Função hashing: é um algoritmo que a partir de uma cadeia de caracteres gera um único valor numérico.

No exemplo aqui conduzido, o atributo "Situação Pagamento" será mapeado no atributo "Negligente", do tipo booleano (SIM/NÃO), através da transformação do valor "EM DIA" no valor "NÃO" (cliente não negligente no pagamento) e dos demais valores ("ATRASO1P", "ATRASO2P", "ATRASO3P", etc.) - que indicam o número de parcelas em atraso - no valor "SIM" (cliente negligente no pagamento). O resultado desta codificação é ilustrado na Tabela 2.7, onde os dados assumem seu formato final, equivalente ao da Tabela 2.1.

CLIENTE	SALÁRIO (RS)	DÉBITO (RS)	NEGLIGENTE
1	1.200	800	SIM
2	1.300	650	SIM
3	1.350	450	SIM
4	1.350	820	SIM
5	1.400	600	NÃO
6	1600	200	NÃO
7	1.600	800	SIM
8	1.650	500	SIM
9	1.850	650	SIM
10	1.900	500	NÃO
11	2.200	400	NÃO
12	2.200	550	NÃO
13	2.250	700	NÃO
14	2.300	520	NÃO

Tabela 2.7 - Dados de empréstimos após codificação.

2.2.2 O processo de mineração dos dados (“Data Mining”)

Mineração de dados é o processo de busca de relacionamentos e padrões globais existentes nas bases de dados. Devido à grande quantidade de dados dos sistemas de bancos de dados atuais, estes relacionamentos estão “escondidos”. Como exemplo dos relacionamentos escondidos pode-se citar o relacionamento entre dados de pacientes e seus diagnósticos médicos. Estes relacionamentos representam o valioso conhecimento e seus objetos nos bancos de dados existentes. Os bancos de dados devem representar um espelho sincero do mundo real registrado pelo mesmo.

Como o número de possíveis relacionamentos existentes em um banco de dados é muito grande, a busca daqueles corretos através da simples validação de cada um deles, é proibitivo. Devido ao aumento significativo da capacidade de processamento necessário, isto se torna um dos primeiros problemas da mineração de dados. Para a solução deste tipo de problema, pode-se usar estratégias de buscas inteligentes, que tiveram sua origem na área chamada de aprendizado de máquinas [8].

Outro problema importante é que as informações nos objetos de dados são geralmente corrompidas ou esquecidas. Portanto, técnicas estatísticas devem ser aplicadas para estimar a confiança dos relacionamentos descobertos.

2.2.2.1 Objetivos primários da mineração de dados

Os objetivos primários da mineração de dados na prática são descrição e predição.

Descrição encontra os padrões de interpretação humana a partir da descrição dos dados. Como exemplo de descrição, a produção de determinado cereal na última colheita pode ser apontado a partir de um conjunto de dados armazenados. Neste caso o valor da produção pode ou não ser confirmado de pronto.

Por predição entende-se a utilização de algumas variáveis ou campos na base de dados para predizer o desconhecido ou valores futuros de outras variáveis de interesse. Como exemplo de predição, tem-se a previsão da produção do mesmo cereal na próxima colheita, a partir do conjunto de dados usado na descrição. Como a produção só será determinada no final da colheita, então este dado não pode ser confirmado de pronto.

A importância de cada uma destas definições varia consideravelmente com a aplicação em questão [8].

Levando-se em conta o processo de KDD, a descrição tende a ser mais importante que a predição. Enquanto isto, para aplicações de aprendizado de máquinas, como reconhecimento de fala, a predição é geralmente o objetivo principal [8].

Independentemente dos objetivos primários da mineração de dados, as tarefas primárias são as mesmas.

2.2.2.2 Tarefas primárias da mineração de dados

2.2.2.2.1 Classificação

A classificação é uma função de aprendizado onde um dado é mapeado em uma das diversas classes predefinidas [12] [29] [20].

Como exemplos de métodos de classificação usados no processo de KDD, pode-se citar a classificação de tendências do mercado financeiro e a identificação automática de objetos de interesse em grandes bases de dados de imagem [8].

Voltando ao nosso exemplo, na Figura 2.3, tem-se uma partição simples dos dados em duas regiões distintas de classes. Caso o banco deseje usar a região de classificação para uma decisão automática de futuros empréstimos, decisão linear não é uma perfeita separação das classes.

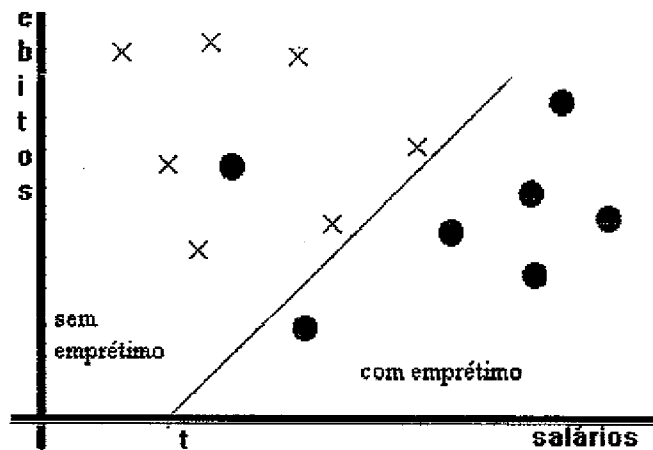


Figura 2.3 - Classificação linear limitada pelo conjunto de dados de empréstimo.

2.2.2.2 Regressão

A regressão é a função de aprendizado que mapeia os dados com predição variável de valores reais. Como exemplos de aplicação de regressão pode-se citar:

- Predizer a soma da biomassa presente em uma floresta fornecida por medidas com sensores remotos de microondas;

- Estimar a probabilidade de um paciente sobreviver dado o resultado de um conjunto de diagnósticos de exames;
- Predizer a demanda do consumo de um novo produto em função da despesa feita;
- Predizer séries temporais onde as variáveis de entrada podem ser versões da variável de predição.

A partir do nosso exemplo, a Figura 2.4 mostra o resultado da regressão linear simples onde o “débito total” é visto como uma função linear da renda, a predição é pobre pois existe uma correlação fraca entre as duas variáveis.

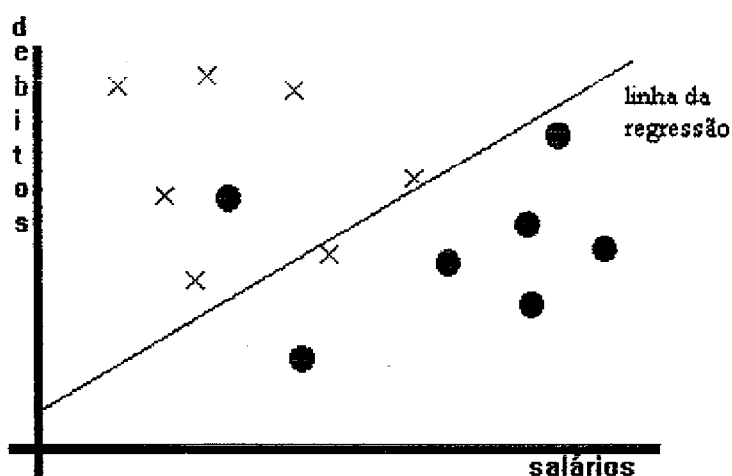


Figura 2.4 - Regressão para o conjunto de dados de empréstimos.

2.2.2.2.3 Clusterização

A Clusterização é a tarefa comum da descrição onde existe um número finito de categorias ou agrupamentos (“clusters”) para descrever os dados. As categorias podem ser mutuamente exclusivas e exaustivas ou consistir numa representação como categorias hierárquicas ou sobrepostas.

Exemplos de aplicações de clusterização no processo de KDD incluem descoberta de sub-populações homogêneas para consumidores do mercado e identificação de subcategorias do espectro de medidas [8] .

A Figura 2.5 mostra a possível clusterização do conjunto de dados anterior em 3 clusters. Neste caso todos os pontos passam a ser representados por x para indicar que os membros das classes não são mais conhecidos. Outro ponto importante na figura é que os clusters se sobrepõem, permitindo que pontos pertençam a mais de um cluster.

O relacionamento do cluster é a tarefa de estimativa da probabilidade, que consiste em técnicas de estimativas de dados da função de probabilidade multi-variada de todas as variáveis/campos do banco de dados [27] .

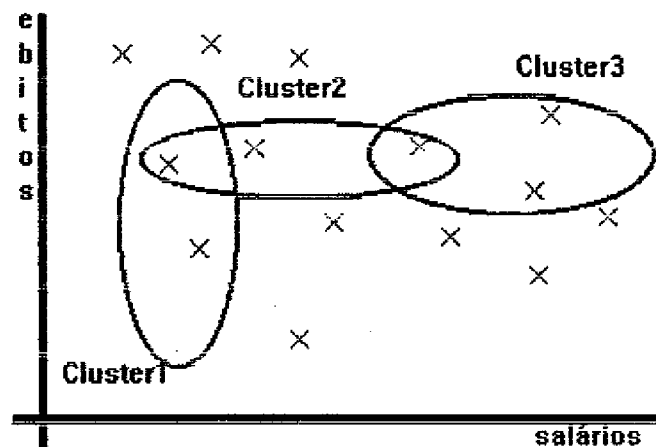


Figura 2.5 - Divisão do conjunto de dados de empréstimos em 3 grupos.

Outras tarefas podem ser citadas:

- Associação: tarefa que busca regras de relacionamento entre ocorrências de itens relevantes ao contexto. Exemplo: Análise de cestas de mercado;

- Predição de séries temporais: tarefa que busca regras ou fórmulas capazes de prever o valor do atributo objetivo em função do tempo. Exemplo: Previsão de colheitas, flutuação sazonal de vendas, etc.

2.2.2.3 Seleção da tarefa de mineração de dados

Para cada tipo de conhecimento que se deseja obter, existe um tipo de aplicação que se adequa melhor. Uma vez definida a aplicação, é feita a seleção da tarefa a ser realizada. Na Tabela 2.8 são apresentadas diversas tarefas, com os algoritmos correspondentes e o tipo de aplicação mais indicada.

Tarefas de Mineração de Dados	Algoritmos associados	Aplicações Típicas	Exemplo
Classificação	Árvores de decisão, redes neurais, algoritmos genéticos	Mercado alvo, controle de qualidade, taxas de riscos	Tabela 2.10
Regressão	Regressão linear e não linear, ajuste de curva, redes neurais	Colocação de clientes, modelos de estimativas de preço, controle de processos	Figura 2.4 e Tabela 2.10
Clusterização	Redes neurais, estatística	Segmentação de mercado, reutilização de projetos	Tabela 2.11
Associação	Estatísticas, teoria de conjuntos	Análise de cestas de mercado	Tabela 2.9
Predição de séries temporais	Modelos estatísticos - ARMA e Box-Jenkins, redes neurais e lógica fuzzy	Predição de vendas, predição de razão do interesse, controle de artigos	Figura 2.6

Tabela 2.8 - Tarefas da mineração de dados, seus algoritmos e aplicações.

Exemplos de dados para as tarefas de mineração:

LEITE	CAFÉ	CERVEJA	PÃO	MANTEIGA	ARROZ	FEIJÃO
NÃO	SIM	NÃO	SIM	SIM	NÃO	NÃO
SIM	NÃO	SIM	SIM	SIM	NÃO	NÃO
NÃO	SIM	NÃO	SIM	SIM	NÃO	NÃO
SIM	SIM	NÃO	SIM	SIM	NÃO	NÃO
NÃO	NÃO	SIM	NÃO	NÃO	NÃO	NÃO
NÃO	NÃO	NÃO	NÃO	SIM	NÃO	NÃO
NÃO	NÃO	NÃO	SIM	NÃO	NÃO	NÃO
NÃO	NÃO	NÃO	NÃO	NÃO	NÃO	SIM
NÃO	NÃO	NÃO	NÃO	NÃO	SIM	SIM
NÃO	NÃO	NÃO	NÃO	NÃO	SIM	NÃO

Tabela 2.9 - Dados de cesta de mercado.

CLIENTE	SALÁRIO	DÉBITO	NEGLIGENTE
1	1200,00	800,00	SIM
2	1300,00	650,00	SIM
3	1350,00	450,00	SIM
4	1350,00	820,00	SIM
5	1400,00	600,00	NÃO

Tabela 2.10 - Dados de empréstimos.

CLIENTE	SALÁRIO	DÉBITO	CLUSTER
1	1200,00	800,00	-
2	1300,00	650,00	1, 2
3	1350,00	450,00	1
4	1350,00	820,00	-
5	1400,00	600,00	2
6	1600,00	200,00	-
7	1600,00	800,00	-
8	1650,00	500,00	-
9	1850,00	650,00	2, 3
10	1900,00	500,00	-

11	2200,00	400,00	-
12	2200,00	550,00	3
13	2250,00	700,00	3
14	2300,00	520,00	-

Tabela 2.11 - Dados de empréstimos clusterizados.

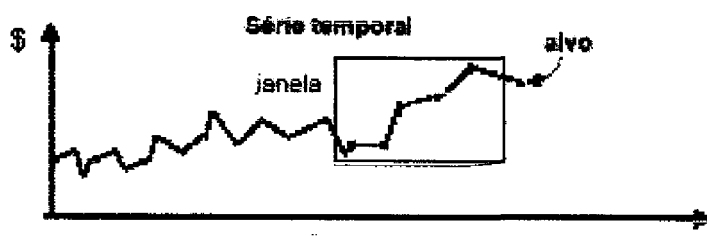


Figura 2.6 - Previsão de série temporal.

2.2.2.4 Técnicas de inferência

O banco de dados é um estoque de informações confiáveis onde a recuperação das informações é feita de maneira eficiente. A informação recuperada não é obrigatoriamente uma cópia exata das informações armazenadas no banco de dados, mas é a informação que pode ser inferida a partir destes dados.

Tendo isto em mente, duas técnicas de inferência se destacam:

- **Dedução:** é a técnica de inferir informação a partir de uma seqüência lógica da informação da base de dados. A maior parte dos sistemas de gerenciamento de base de dados (DBMS), como os DBMS relacionais, oferecem operadores simples para a dedução da informação. Por exemplo, o operador join aplicado em duas tabelas relacionais onde a primeira administra as relações entre solicitantes de empréstimos e agências bancárias e a

segunda as relações entre agências bancárias e gerentes, infere a relação entre solicitantes de empréstimos e gerentes;

- Indução: é uma técnica para inferir informação que é generalizada a partir da informação na base de dados. Por exemplo, a partir das tabelas de solicitantes de empréstimos-agências bancárias e agências bancárias-gerentes, pode ser inferido que cada solicitante de empréstimo tem um gerente responsável.

Este é um conhecimento ou uma informação de alto nível: ou seja, uma declaração geral sobre as propriedades dos objetos. No processo de mineração de dados estas regularidades (combinações de valores para certos atributos, partilhados por fatos dos bancos de dados) são procuradas. Sendo assim, pode-se dizer que a regularidade é o conhecimento ou que são regras usadas para predizer o valor de um atributo em função de outros atributos.

A diferença mais importante entre dedução e indução é que no primeiro caso, os resultados formados por declarações sobre o mundo real são comprovadamente corretas desde que o banco de dados esteja correto, enquanto a indução resulta em declarações que são suportadas pelo banco de dados, mas não são necessariamente verdade no mundo real. Um dos mais importantes aspectos do processo de indução é, portanto, a seleção das regras mais plausíveis e regulares, suportadas pelo banco de dados.

A inferência da informação de um banco de dados está fora da capacidade humana, apenas pela razão do seu tamanho sempre crescente. Então, o processo de inferência pode ser suportado pelo DBMS, porém, nenhum suporta a indução.

2.2.2.5 Métodos de mineração de dados

Existe uma grande variedade de métodos de mineração de dados. Como a idéia deste trabalho é apresentar uma descrição do processo de KDD de forma concisa e de fácil entendimento, um subconjunto de métodos conhecidos será focalizado.

2.2.2.5.1 Métodos estatísticos

Estes métodos são usados em problemas de descoberta de conhecimento, onde o interesse está centrado em uma simples variável de saída y e uma coleção de preditores. Todos os modelos assumem a viabilidade dos dados treinados e têm como objetivo encontrar um modelo para prever o valor y a partir de x , que seja executado e produza bons resultados a partir de novos dados. Este problema tinha uma solução definida antes que avanços da computação tornassem possível relaxar as suposições existentes. Os estatísticos têm, desde então, tentado suprir a ânsia de inventar novos métodos de estimativas (média estimada) e modelos (modelos aditivos) para explorar uma formulação menos restrita.

Outros métodos surgiram na corrida para desenvolver modelos com flexibilidade crescente, talvez encorajadas pelo famoso resultado de Kolmogorov. Segundo este, todas as funções multidimensionais podem ser representadas por uma composição de funções unidimensionais. Apesar de todas as tentativas, os estatísticos não estão satisfeitos com os resultados obtidos, pois as classes de modelos existentes ainda não possuem a flexibilidade desejada para ser útil na prática, porque os dados são finitos e os ruídos prevalecem. Daí existirem modelos que aumentem a quantidade dos dados necessários na proporção de sua complexidade e tamanho, quase enganando a simplicidade da análise.

2.2.2.5.2 Métodos lineares

Os modelos clássicos de predição e classificação são regressões lineares e análise linear de discriminante, respectivamente. O termo “linear” nestes modelos é devido ao fato da superfície de regressão ou classificação ser um plano.

A flexibilidade e a computação direta envolvidas na regressão linear são feitas sem uso de outras técnicas associadas. Por exemplo, funções radiais básicas de redes neurais são meras regressões lineares de um conjunto de características do núcleo (“kernel”). Lowe e Webb [17] empregam uma arquitetura de rede neural para computar dados não lineares que alimentam o estágio final da regressão linear e redes polinomiais, usando regressão linear em todo nó, para combinar previamente as transformações polinomiais dos dados.

A análise linear de discriminante, que apropria pré e pós-processamento, pode ser formulada como um estágio de regressão linear [13]. Ela permite substituir o módulo de regressão linear por um método de estimação não paramétrico e não linear avançado, aumentando os tipos de padrões que podem ser manuseados pelas técnicas de classificação.

2.2.2.5.3 Método das árvores de decisão e regras

Árvores de decisão e regras que usam podas têm uma forma de representação simples, fazendo com que o modelo inferido seja relativamente fácil de ser compreendido pelo usuário. Apesar do fácil entendimento, quando se restringe as informações a uma árvore particular ou as regras de representação pode-se reduzir significativamente a forma funcional do modelo. Quando se fala da forma funcional, está-se referindo ao poder de aproximação do modelo.

Por exemplo, quando na Figura 2.1 se adotou a divisão do limite sobre o salário t , e aplicou-se sobre esta variável o conjunto de dados de empréstimos, foi limitado severamente o tipo de classificação através das fronteiras induzidas.

No momento em que se tornou o modelo mais complexo, através da introdução de hiper-planos variados, com ângulos arbitrários, aumentou-se o espaço do modelo, fazendo crescer a complexidade de seu entendimento, ao mesmo tempo, permitindo que a predição feita pelo modelo fosse mais poderosa.

Uma vez que existem inúmeros estudos sobre árvores de decisão e algoritmos de regras de indução descritos nas máquinas de aprendizado, aplicadas na literatura estatística [3] [25] , pode-se afirmar que na sua maioria estes são baseados em métodos evolutivos de modelos baseados em probabilidade. Estes modelos possuem graus variados de sofisticação, conseqüentemente, quanto maior a sofisticação, maior a sua complexidade e quanto menor a sua sofisticação, menor a sua complexidade.

O principal problema com árvores é que elas destroem dados numa razão exponencial com a profundidade. Deste modo, para cobrir estruturas complexas, extensos conjuntos de dados são requeridos.

Os métodos de busca infinita, que envolvem regras de cultivo e poda e as três estruturas de árvore são tipicamente empregadas para explorar o espaço exponencial dos modelos possíveis. As árvores e as regras são usadas basicamente na modelagem de predição, de classificação e de regressão, embora elas possam ser aplicadas para modelagem descritiva resumida.

2.2.2.5.4 Métodos de aprendizagem relacional

Enquanto árvores de decisão e regras têm a sua representação restrita à lógica proposicional, o aprendizado relacional (também conhecido como programação em lógica indutiva) usa um padrão mais flexível de linguagem de lógica de primeira ordem.

O aprendizado relacional pode facilmente encontrar fórmulas como $\underline{x} = \underline{y}$. Muitas pesquisas em modelos de métodos de avaliação para aprendizado relacional constituem pura lógica.

2.2.2.5.5 Método das redes neurais

Redes Neurais Artificiais (RNAs) são uma classe útil de modelos, constituída de camadas de nós. Cada nó calcula o somatório dos pesos de suas entradas e realiza uma transformação na saída. Geralmente a saída dos nós de uma camada se conectam as camadas subseqüentes através dos nós de entradas. Utilizando o procedimento de aprendizado do algoritmo de retropropagação [33], apresenta-se um padrão de cada vez; os erros são usados para ajustar os pesos dos nós de saída, proporcionalmente às suas contribuições (magnitude). Os pesos são assim ajustados similarmente e o erro final retorna à primeira camada. Os Pesos iniciais são tipicamente randômicos.

O procedimento de ajuste de peso é um método de gradiente local, seqüencial e interminável. É de gradiente local porque esquece a otimização geral, é seqüencial porque permite que os casos iniciais tenham muita influência e é interminável porque lida com um tipo cru de regularização, onde a moderação do tempo de processamento é o modo principal de permitir sobrecarga. Entretanto, estas características permitem o cancelamento de

parâmetros como a lenta busca local e não permite que o excesso de parâmetros sobrecarregue facilmente a rede.

Note que este é o grau certo de liberdade aplicada a uma RNA e é usualmente menor que se imagina. O perigo da sobrecarga pode depender na duração do treinamento, desde que pesos de nós randômicos lidem essencialmente com funções lineares (operação dos nós no meio e nas extremidades da senoíde), que são absorvidas pelas camadas subseqüentes. Como os nós são puxados para a parte curva da sigmóides durante o treinamento, poucos parâmetros se tornam ativos. Isto explica a observação comum de que o desempenho de uma RNA num problema é geralmente suprida em sua robustez com respeito a mudanças na estrutura da rede.

2.2.2.5.6 Método dos algoritmos genéticos

Algoritmos genéticos como o próprio nome diz são algoritmos que simulam o processo de seleção natural proposto por Charles Darwin em 1859. Segundo Darwin, a seleção natural é um processo que privilegia os organismos que melhor se adaptam ao meio ambiente, isto é, quanto mais o organismo esta adaptado ao seu ambiente, maior a chance de sobrevivência e mais características ele irá transmitir para seus sucessores através de seus cromossomos. Com isto a tendência de aprimoramento pode ser verificada nas diversas espécies existentes.

A natureza mantém e nutre muitas inconsistências e contradições para um dado problema. De fato a manutenção da diversidade genética é um ingrediente importante na evolução além de assegurar a habilidade de futuras adaptações à alterações no ambiente.

Os algoritmos genéticos utilizam esta mesma propriedade para desenvolver seus modelos. Vários modelos são estudados, mas apenas aqueles que se mostram mais perto da

solução desejada são desenvolvidos. Pode-se dizer então que os organismos da teoria de Darwin são equivalentes as estruturas de dados, enquanto os cromossomos são equivalentes as cadeias de bits. Daí existir mais de um conjunto de considerações inteiramente diferentes que podem ser usados numa mesma solução do problema. É muito difícil existir uma solução matematicamente ótima para um problema, porém existem soluções muito próximas da ótima.

Com isto é possível obter soluções de problemas sem que estes sejam explicitamente programados. A solução de um problema pelos métodos tradicionais implica na existência de métodos de aprendizado de máquina, tais como: inteligência artificial, sistemas auto-aperfeiçoáveis, redes neurais e indução, que utilizam programas de computadores convencionais. Estes métodos utilizam paradigmas que necessitam de estruturas especializadas para cada tipo de solução desejada. Se o que se deseja é chegar a um computador genérico, o qual possa ser utilizado para resolver problemas genéricos, uma preocupação básica deve ser com os programas. Estes sim representam as estruturas que se deseja encontrar.

As características gerais deste tipo de algoritmo são coincidentes com as características gerais da evolução das espécies, que são:

- A evolução é um processo que ocorre basicamente nos cromossomos;
- O processo de seleção natural codifica as estruturas mais aptas para a reprodução, com mais frequência que aquelas não tão aptas;
- O processo de reprodução se dá em três modos:
 - Mutação;
 - Reprodução;
 - Cruzamento;

- A evolução genética não tem memória.

Holland, antes de 1970, acreditava que incorporando o algoritmo no computador apropriadamente teria uma técnica para resolver problemas difíceis como a natureza faz através da evolução.

Em geral os algoritmos genéticos não são usados na área genética, como o nome nos leva a acreditar, mas na área de otimização.

2.2.2.5.7 Considerações finais

Existem muitas outras técnicas de mineração de dados, particularmente métodos especializados para tipos particulares de dados e domínios que não foram mencionados neste trabalho. Isto não desmerece nenhuma delas, foi feita apenas uma tentativa de apresentar as mais importantes. Uma discussão geral sobre tarefas de mineração de dados e componentes têm relevância para uma variedade de métodos. Embora estes algoritmos e aplicações possam parecer diferentes na superfície, não é incomum descobrir que eles possuem muitos componentes comuns.

Entendendo a mineração de dados e os modelos indutivos no que diz respeito aos seus componentes, a tarefa de algoritmos de mineração de dados fica mais clara e mais fácil para o usuário entender as contribuições e aplicações para o processo de KDD.

2.3 CLASSIFICAÇÃO EM DETALHES

Para melhor entendimento da natureza do problema de classificação de dados e a aplicação da técnica de Rough Sets nesta tarefa, objeto principal deste trabalho, tomemos como exemplo a Tabela 2.12, representando uma pequena porção de uma base de dados armazenada.

LOJA	EXPERIÊNCIA VENDEDORES	QUALIDADE DO PRODUTO	BOA LOCALIZAÇÃO	RETORNO
1	Alta	Boa	não	lucro
2	Média	Boa	não	prejuízo
3	Média	Boa	não	lucro
4	Baixa	Média	não	prejuízo
5	Média	Média	sim	prejuízo
6	Alta	Média	sim	lucro

Tabela 2.12 - Informações de lojas de um bairro

As colunas representam os campos ou atributos da informação, e as linhas representam os registros ou instâncias de dados.

Os dados informam características básicas de lojas de um bairro:

- Nível de experiência dos vendedores
- Qualidade do produto vendido
- Se possui boa localização
- Se é lucrativa ou deficitária

Neste caso, o objetivo é realizar uma classificação dos dados acima, obtendo relacionamentos entre conjuntos de atributos preditivos (experiência, qualidade e localização são candidatos) que resultem em determinada classe de um atributo objetivo (no caso, retornar lucro ou prejuízo).

Após o trabalho de mineração poderia ser obtida uma regra como, por exemplo:

“SE EXPERIÊNCIA = alta E QUALIDADE = boa ENTÃO RETORNO = lucro”

Assim, para uma nova loja da qual se deseja prever o retorno, como a de número 7, dadas suas informações:

LOJA	EXPERIÊNCIA VENDEDORES	QUALIDADE DO PRODUTO	BOA LOCALIZAÇÃO	RETORNO
7	ALTA	BOA	SIM	?

- EXPERIÊNCIA DOS VENDEDORES É ALTA;
- QUALIDADE DO PRODUTO É BOA;
- POSSUI BOA LOCALIZAÇÃO.

Ela poderia ser classificada, com determinada probabilidade de acerto, como lucrativa, pela regra exibida como exemplo.

A solução nos casos reais, infelizmente, não é tão simples ou visual como no exemplo dado acima. Para varrer imensos volumes de dados à busca de padrões ocultos, sem mesmo se conhecer que atributos são relevantes para a conclusão a que se deseja chegar, são necessários métodos automáticos e eficazes de busca e análise de informações.

A seguir, será abordada uma das novas técnicas que vêm sendo utilizadas neste sentido: a teoria de Rough Sets.

3 ROUGH SETS

O termo Rough Sets pode ser aproximadamente traduzido como Conjuntos Imprecisos. Ao longo do texto é adotado o termo original, por não haver ainda uma consolidação para a melhor tradução para o português.

A teoria de Rough Sets pode ser descrita, de forma bem sumária, como a análise de informações baseada na descoberta de grãos de similaridade existentes no universo dos dados, e no estabelecimento de relações entre estes grãos.

Inicialmente, é importante ressaltar que os conceitos de Rough Sets complementam, mas não competem com outros métodos. Eles podem ser usados conjuntamente com outras abordagens como, por exemplo, Lógica Nebulosa, Algoritmos Genéticos, métodos estatísticos, Redes Neurais, etc.

A teoria é de fácil compreensão e aplicação. Diversos sistemas aplicativos baseados em Rough Sets já foram implementados e muitas aplicações em diversas áreas do conhecimento foram relatadas. Mais detalhes sobre estas aplicações podem ser encontradas em [24].

3.1 BASE DE DADOS E CONCEITOS INICIAIS

Para as considerações acerca dos conceitos, será utilizada como exemplo a Tabela 2.12, exibida anteriormente.

Naquela pequena base de dados, cada loja é descrita pelo valor dos atributos Experiência, Qualidade, Localização e Retorno, aos quais chamaremos E, Q, L e R por simplificação.

Cada subconjunto de atributos determina uma partição de todos os objetos (registros) em classes, compostas por registros que têm o mesmo valor para cada um destes atributos. Por exemplo, os atributos Q e L agregam as lojas nas seguintes classes: {1,2,3}, {4}, {5,6}. Assim, cada base de dados possui uma família de padrões de classificação que serão usados como base para futuras considerações.

Uma base de dados será formalmente definida da seguinte maneira: um par $S=(U,A)$, onde U e A são conjuntos finitos, não-vazios, chamados Universo (objetos ou registros) e Atributos, respectivamente.

A cada atributo $a \in A$ é associado um conjunto V_a de seus Valores distintos, chamado Domínio de a. Qualquer subconjunto B de A determina uma relação binária $I(B)$ em U, chamada de Relação de Indiscernibilidade, definida a seguir:

$(x,y) \in I(B)$ se e somente se $b(x) = b(y)$ para todo $b \in B$, onde $b(x)$ é o valor do atributo b no elemento (objeto ou registro) x.

Em [23] é demonstrado que $I(B)$ é uma relação de equivalência. A família de todas as classes de $I(B)$ é a partição determinada por B em U, e é denominada $U/I(B)$, ou simplesmente U/B . Uma classe de equivalência de $I(B)$ (um bloco da partição) contendo x será identificado como $B(x)$.

Se (x,y) pertence a $I(B)$ diremos que x e y são B-indiscerníveis. As classes de equivalência da relação $I(B)$ são referidos como Conjuntos Elementares por B ou grãos-B.

A relação de equivalência não é suficiente para abranger toda a teoria de Rough Sets. Relações de tolerância e de ordenação, entre outras, são propostas, por exemplo, em [29] [30] [35]. Para o escopo deste trabalho, porém, será suficiente como base a **relação de equivalência**.

3.2 APROXIMAÇÕES

Inicialmente seja considerada a seguinte questão: Quais são as características que definem uma loja como dando lucro ou prejuízo, a partir das informações da Tabela 2.12? Pode-se notar que não há uma resposta única para esta pergunta, pois as lojas 2 e 3 têm as mesmas características pelos atributos E, Q e L, mas a loja 2 é lucrativa enquanto a 3, não. Em vista dos mesmos atributos, pode-se dizer com certeza que as lojas 1 e 6 dão lucro e as lojas 4 e 5 dão prejuízo, mas nada se pode afirmar para lojas com as características de 2 e 3. Em outras palavras, do ponto de vista dos atributos E, Q e L, 1 e 6 certamente pertencem ao conjunto das lojas lucrativas $\{1,3,6\}$ (ou conceito lojas lucrativas - X), enquanto 2 e 3 têm uma probabilidade de pertencer a este conjunto. O conjunto de lojas $\{1,6\}$ compõe a chamada Aproximação Inferior ou Região Positiva de $\{1,3,6\}$ pelo conjunto de atributos $B=\{E,Q,L\}$. O conjunto de lojas $\{1,2,3,6\}$ compõe a chamada Aproximação Superior de $\{1,3,6\}$, enquanto sua diferença, o conjunto $\{2,3\}$, representa o que é denominado Região de Fronteira.

As aproximações podem ser definidas formalmente como dois conjuntos, $B_*(X)$ e $B^*(X)$, chamados respectivamente de Aproximação Inferior de X por B e Aproximação Superior de X por B, conforme a definição a seguir:

$$B_*(X) = \bigcup_{x \in U} \{B(x) : B(x) \subseteq X\}$$
$$B^*(X) = \bigcup_{x \in U} \{B(x) : B(x) \cap X \neq \emptyset\}$$

Assim, a Aproximação Inferior por B de um conceito X (lucro, por exemplo) é a união de todos os grãos-B contidos no conceito, enquanto a Aproximação Superior por B é a união

de todos os grãos-B que têm interseção não nula com o conceito. O conjunto denominado Região de Fronteira de X por B é definido como:

$$\text{BNB}(X) = B^*(X) - B_*(X)$$

Se a região de fronteira for o conjunto vazio, isto é, $\text{BNB}(X) = \emptyset$, então X é crisp (preciso) com relação a B. Caso contrário, isto é, $\text{BNB}(X) \neq \emptyset$, então X é dito rough (impreciso) com relação a B.

A Figura 3.1 exibe uma forma gráfica de representar o conceito das aproximações.

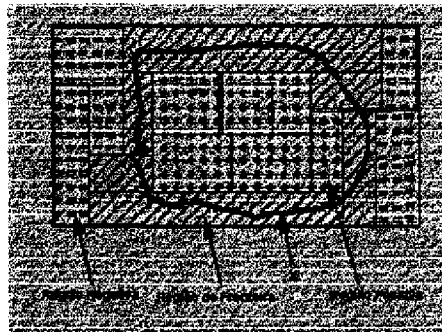


Figura 3.1 - As aproximações em forma gráfica.

O Grau de Imprecisão, ou Roughness, de um conjunto pode ser caracterizado numericamente como:

$$\alpha_B(X) = \left(\frac{\text{card}(B_*(X))}{\text{card}(B^*(X))} \right)$$

Se $\alpha_B(X) = 1$, então X é preciso (crisp) com relação a B. Senão, X é impreciso (rough) com relação a B. Por exemplo, o Grau de Imprecisão de $X = \{1,3,6\}$ é dado por $\text{card}(\{1,6\}) / \text{card}(\{1,2,3,6\}) = 2/4 = 0,5$.

3.3 DEPENDÊNCIA DE ATRIBUTOS

Outra questão importante na análise de dados é descobrir dependências entre atributos. Suponha que um conjunto A de atributos de uma base de dados $S=(U,A)$ seja dividido em dois subconjuntos P e O , chamados atributos preditivos e objetivos, respectivamente, tal que $P \cup O = A$ e $P \cap O = \emptyset$. Tal base de dados é chamada tabela de decisão.

Intuitivamente, um conjunto de atributos O depende totalmente de P ($P \Rightarrow O$), se todos os valores de O forem univocamente determinados por valores de atributos de P . Em outras palavras, O depende totalmente de P se existe uma dependência funcional entre valores de P e O .

Formalmente, a dependência pode ser assim definida:

Sejam P e O subconjuntos de A . O Grau de Dependência de O em relação a P é dado por:

$$\gamma(P,O) = \frac{\text{card}(P.(X))}{\text{card}(X)}$$

Se $\gamma(P,O) = 1$ diz-se que O depende totalmente de P ; se $\gamma(P,O) = 0$, O não depende de P ; se $0 < \gamma(P,O) < 1$, O depende parcialmente de P ;

Por exemplo, na Tabela 2.12 o grau de dependência de $O=\{R\}$ em relação a $P=\{E,Q,L\}$ é dado por $\text{card}(\{1,6\}) / \text{card}(\{1,3,6\}) = 2/3$.

3.4 REDUÇÃO DE ATRIBUTOS

Uma redução é um subconjunto de atributos que preserva o grau de dependência. Em outras palavras, uma redução é um subconjunto de atributos preditivos que permite tomar as mesmas decisões do conjunto completo.

Por exemplo, na Tabela 2.12 existem duas reduções, {E,Q} e {E,L}, dos atributos preditivos {E,Q,L}.

A necessidade de prévia redução de atributos é um ponto fraco da teoria de Rough Sets. O cálculo de reduções é um problema n-p completo, e seu processamento em grandes bases de dados exige grande esforço computacional.

3.5 SIGNIFICÂNCIA DE ATRIBUTOS

O conceito de redução nos leva a desconsiderar alguns atributos de tal forma que as relações básicas no banco de dados sejam preservadas. Alguns atributos, no entanto, têm maior relevância que outros na preservação destas relações. Este conceito é chamado de *Significância do atributo*, e pode ser assim definido:

Sejam P e O conjuntos de atributos preditivos e objetivos, respectivamente, e seja a um atributo preditivo pertencente a P. A significância de a será calculada em função da mudança do Grau de Dependência de O em relação a P com a remoção de a, segundo a fórmula abaixo:

$$\sigma_{(P,O)}(a) = 1 - \left(\frac{\gamma(P - \{a\}, O)}{\gamma(P, O)} \right)$$

Este coeficiente pode ser visto como o erro que ocorre na definição de O por P quando a é removido. Por exemplo, para $X = \{1,3,6\}$, $O = \{R\}$, $P = \{E,Q,L\}$ e $a = L$, o Grau de

Significância do atributo L é dado por $1 - (\gamma(\{E,Q\}) / \gamma(\{E,Q,L\})) = 1 - ((2/3) / (2/3)) = 0$, significando que remover o atributo L de consideração não afetará os resultados. O mesmo resultado seria encontrado ao se remover Q de $\{E,Q,L\}$.

3.6 REGRAS DE DECISÃO

As dependências entre atributos encontradas com as técnicas de Rough Sets descritas anteriormente podem ser expressas na forma de regras de decisão. Este é o ponto de ligação entre a teoria e a aplicação em mineração de dados a que este trabalho se propõe.

Para exemplificar regras geradas pelos métodos abordados, tomemos a redução $P=\{E,Q\}$ da base da Tabela 2.12. Algumas regras de decisão descrevendo a dependência de $\{R\}$ em relação a P poderiam ser:

- Se E = ALTA e Q = BOA ENTÃO R = LUCRO
- Se E = MÉDIA e Q = BOA ENTÃO R = PREJUÍZO
- Se E = MÉDIA e Q = BOA ENTÃO R = LUCRO
- Se E = BAIXA e Q = MÉDIA ENTÃO R = PREJUÍZO
- Se E = ALTA e Q = MÉDIA ENTÃO R = LUCRO

Pode-se notar, sob a ótica dos dados fornecidos, que algumas regras extraídas são 100% válidas, enquanto outras têm um nível de confiabilidade menor.

Um conjunto de regras de decisão normalmente é chamado de base de conhecimento. A cada regra podem ser associados dois coeficientes básicos que denotam sua qualidade: o Fator de Certeza (também chamado de eficácia) e o Fator de Cobertura (também chamado de abrangência), definidos a seguir.

Seja a regra de decisão “SE p ENTÃO q ”. O Fator de Certeza ou Eficácia desta regra é dado por:

$$\pi (p/q) = \left(\frac{\text{card}(p \wedge q)}{\text{card}(p)} \right)$$

O Fator de Certeza indica o quanto a regra acerta a previsão, dadas suas pré-condições p . Tomando como exemplo a regra “Se $E = \text{MÉDIA}$ e $Q = \text{BOA}$ ENTÃO $R = \text{LUCRO}$ ”, tem-se $p = \{E = \text{MÉDIA} \text{ e } Q = \text{BOA}\}$ e $q = \{R = \text{LUCRO}\}$. As lojas 2 e 3 satisfazem p , enquanto somente a loja 3 satisfaz p e q . O Fator de Certeza desta regra é, portanto, $1/2 = 50\%$.

Dada a regra “SE p ENTÃO q ”, seu Fator de Cobertura ou Abrangência é dado por:

$$\pi (q/p) = \left(\frac{\text{card}(p \wedge q)}{\text{card}(q)} \right)$$

O Fator de Cobertura indica o quanto a regra abrange os casos em que o atributo objetivo tem um determinado valor. No exemplo da regra “Se $E = \text{MÉDIA}$ e $Q = \text{BOA}$ ENTÃO $R = \text{LUCRO}$ ”, $p = \{E = \text{MÉDIA} \text{ e } Q = \text{BOA}\}$ e $q = \{R = \text{LUCRO}\}$. As lojas 1, 3 e 6 satisfazem q , enquanto somente a loja 3 satisfaz p e q . O Fator de Cobertura desta regra é, portanto, $1/3 = 33,3\%$.

As técnicas exibidas neste capítulo serão implementadas em uma aplicação de KDD denominada Bramining, a ser abordada no capítulo 5. Antes de apresentar a aplicação, entretanto, serão avaliadas no próximo capítulo algumas ferramentas de mineração de dados do mercado, para melhor balizar o escopo do projeto desenvolvido.

4 AVALIAÇÃO DE FERRAMENTAS

Cinco softwares foram avaliados: WizRule, WizWhy, XpertRule Miner, BusinessMiner e PolyAnalist, e seus resultados serão descritos a seguir:

4.1 WIZRULE

O WizRule é um software de auditoria, descrição e limpeza de dados que, automaticamente, revela todas as regras que modelam a base, e indica os casos de desvio encontrados com relação ao conjunto de regras geradas. Estes desvios podem ser erros nos dados, ou simples variações aceitáveis às regras.

Como dito anteriormente, o WizRule propõe-se a revelar todas as regras que descrevem a base de dados, dentre elas, regras se-então, regras matemáticas, erros na escrita de nomes e valores, bem como, calcular o nível de incerteza de cada desvio e evitar falsos alarmes (casos em que um registro é considerado um desvio à regra, porém não o é).

O WizRule implementa uma abordagem inovadora que, automaticamente, limpa e audita um banco de dados. Esta abordagem se baseia na hipótese de que, na maioria dos casos, os erros são exceções à norma. Por exemplo, se em todas as transações de vendas para um certo cliente, o vendedor se chama João, e existe uma única transação na qual o vendedor se chama Roberto (o qual está relacionado a vendas para outros clientes), isto pode ser considerado um desvio ou uma suspeita de erro.

Para se descobrir tais exceções à norma, precisamos, primeiramente, extrair todas as regras que modelam a base de dados. O WizRule é baseado em um algoritmo matemático capaz de descobrir todas estas regras num curto espaço de tempo. Além disso, também pode

procurar por regras que levem em consideração certos desvios e determinem o nível de incerteza destes.

Ao analisar uma base de dados, o WizRule realiza as seguintes operações:

- Lê a base de dados. Existe a possibilidade de ajustar a análise a ser executada, através da definição de parâmetros tais como: “probabilidade mínima das regras se-então” e “número mínimo de casos de uma regra”. É possível ainda, definir que tipo de regras o WizRule tentará encontrar;
- Revela as regras da base de dados e ainda indica a confiabilidade de cada regra;
- Analisa cada campo em cada registro relativo às regras reveladas e calcula o grau de certeza;
- Lista, por regra, aqueles registros em que ocorrem os mais altos níveis de incerteza, ou seja, possíveis erros. O WizRule gera, então, três relatórios: regras, desvios e erros de escrita.

4.1.1 Relatórios

4.1.1.1 Relatório de Regras

O Relatório de Regras lista sentenças no seguinte formato:

*If cidade is **BRASILIA**
Then
UF is DF
Rule's probability: **0,974**
The rule exists in **111** records.
Significance Level: Error probability is almost 0
Deviations (records' serial numbers):**1414***

Significado das informações:

Regra: "If **informação1** is **valor1** Then **informação2** is **valor2**."

Rule's probability: probabilidade de ocorrência da regra na base de dados (eficácia).

The rule exists in N records: Quantidade de registros em que a regra ocorreu.

Significance Level: avaliação da probabilidade de haver exceções à regra.

Deviations: identificação dos registros em que a regra foi falsa.

4.1.1.2 Relatório de Pronúncia

O Relatório de Pronúncia exibe informações como:

*The value of **Av. Rio Branco** appears 45 times in the **logradouro** field.
There are 2 cases containing similar values: Record Nr 1446*

Esta informação revela que há ocorrências de dados na base que aparentemente são erros de pronúncia ("spelling"). No caso, apesar de "Av. Rio Branco" ocorrer 45 vezes na base, houve uma ocorrência "AL Rio Branco" no registro N° 1446, que pode ser fruto de erro de digitação, por exemplo.

4.1.1.3 Relatório de Desvios

O Relatório de Desvios exibe os registros em que foram verificadas exceções às regras do Relatório de Regras, colocando a seu lado a regra violada.

O WizRule não revela todos os possíveis erros de uma base de dados. O seu sofisticado algoritmo de implementação impossibilita o software de determinar certas

exceções à regra, tratando-as como desvios aceitáveis, ao invés de erros. Isto reduz o número de “falsos alarmes” que outros softwares ou métodos de limpeza de dados podem revelar.

4.1.2 Tipos de regras

As regras analisadas pelo WizRule são, basicamente, de três tipos: fórmulas matemáticas, regras se-então e regras baseadas em escrita.

Um exemplo de regra matemática é:

$$A = B * C$$

Em que:

A = Total

B = Quantidade

C = Preço unitário

Grau de precisão da regra: 0.99

A regra aparece em 1890 registros

Um exemplo de regra se-então é:

If Customer is Summit

and Item is Computer type A

Then

Salesperson = João

Probabilidade da regra: 0.98

A regra aparece em 102 registros

Grau de significância: Probabilidade do erro < 0.1

Um exemplo de regra de escrita é:

O valor de Rio de Janeiro aparece 52 vezes no campo Cidade.

Existem 2 caso(s) contendo valor(es) similar(es).

4.1.3 Facilidade de uso

O WizRule pode ser considerado um software amigável para alguém que detenha conhecimentos básicos de regras de inferência, ou simplesmente lógica matemática. Ele permite a realização de uma análise básica dos dados sem exigir grande interferência do usuário. Conforme o grau de conhecimento de seus mecanismos pelo usuário aumenta, ele permite que sejam feitos “ajustes finos” na análise das informações.

4.1.4 Informações sobre o banco de dados

O software provê de imediato informações gerais sobre os campos da base de dados fornecida. Após a análise, os resultados são exibidos em três categorias de relatórios: Regras, Pronúncia e Desvios.

4.1.5 Conclusões

O WizRule mostra-se uma ferramenta bastante útil na extração de informações ocultas em bases de dados, principalmente quando se trata de grande número de registros e campos contendo dados cujo cruzamento pode resultar em regras para otimização do negócio.

4.2 WIZWHY E WIZWHY PREDICTOR

O WizWhy é um software da mesma família do WizRule, com as mesmas funcionalidades de descrição, limpeza e auditoria da base de dados. Na verdade, o WizWhy e o WizRule trabalham em conjunto, uma vez que o primeiro possui um módulo de predição de casos futuros, com base na descrição de dados realizada pelo último.

A partir das regras geradas pelo WizRule, o WizWhy é capaz de determinar novas saídas para qualquer caso que lhe for apresentado.

4.2.1 Relatórios

Pode-se solicitar que o software obtenha Regras ou Predições.

4.2.1.1 Relatório de Regras

O Relatório de Regras lista sentenças no seguinte formato:

*If cidade is **BRASILIA**
Then*

UF is DF

Rule's probability: 0,974

The rule exists in 111 records.

Significance Level: Error probability is almost 0

Significado das informações:

Regra: "If informação1 is valor1 Then informação2 is valor2.

Rule's probability: probabilidade de ocorrência da regra na base de dados.

The rule exists in N records: Quantidade de registros em que a regra ocorreu.

Significance Level: avaliação da probabilidade de haver exceções à regra.

4.2.1.2 Relatório de Predições

O Relatório de Predições exibe valor2 para determinado valor1 informado pelo usuário.

Ele tem o seguinte formato:

Field to Predict: SPSS.PORTE

Condition Fields:

SPSS.EMPREGADOS = 364,00

Prediction : SPSS.PORTE is 4,00

Relevant rules:

1)If SPSS.EMPREGADOS is $368,50 \pm 156,50$

Then

SPSS.PORTE is 4,00

Rule's probability: 0,899

The rule exists in 213 records.

Significance Level: Error probability $< 0,1$

Significado das informações:

Field to predict: Informa o campo cujo valor será previsto

Condition Fields: Informa o(s) campo(s) sobre os quais foram simulados valores.

Prediction: valor previsto.

Relevant Rules: Regra(s) utilizada(s) para inferir previsão.

4.2.2 Facilidade de uso

Bastante semelhante ao WizRule, o WizWhy também é um software amigável para um usuário com conhecimentos básicos de lógica.

4.2.3 Informações sobre o banco de dados

O software provê de imediato informações gerais sobre os campos da base de dados fornecida. A grande diferença em relação ao WizRule está na necessidade de se destacar um campo da tabela sobre o qual se deseja realizar previsões e estabelecer regras nas quais ele é condição necessária. Ou seja, ele foi feito para gerar regras de classificação, não de associação.

4.2.4 Conclusões

O WizWhy é uma ferramenta útil para a visualização de como se comportariam determinados parâmetros do negócio em condições hipotéticas, característica bastante valiosa em determinados ramos de atividade. O WizWhy Predictor executa exatamente as mesmas funções do módulo de previsões do WizWhy. Sua versão como módulo independente provavelmente dirige-se a determinados segmentos da empresa aos quais somente interessa a realização de projeções de mercado, não o trabalho de construção de regras.

4.3 XPERTRULE MINER

O XpertRule Miner suporta diversas tarefas da fase de mineração de dados, com o uso de uma interface gráfica que permite ao desenvolvedor definir uma seqüência de operações de mineração.

4.3.1 Aspectos funcionais do software

Uma seqüência de operações de mineração é denominada um **script** de mineração. Uma vez desenvolvido, este pode ser executado. O XpertRule Miner só permitirá que sejam executados scripts com interconexões válidas entre as operações. É importante notar que uma operação de mineração não é, como de costume, executada e sim, editada quando acionada.

O XpertRule Miner consiste dos seguintes módulos:

1. Ambiente de desenvolvimento para definição do **script**;
2. Ambiente de execução das operações não-padrão de mineração (ex.: manipulação de campos e tabelas e geração de relatórios);

Dentro do primeiro módulo, existem as operações de Mineração de Dados propriamente ditas, tais como:

<i>ProfileX</i>	Execução de operações da árvore de mineração (ou árvore de decisão);
<i>Case AssociationX</i>	Execução das operações de associação de dados baseados em casos. Entende-se por dados baseados em casos, aqueles dados que possuem um número fixo de campos em cada registro;
<i>AssociationX</i>	Execução de operações de associação em dados transacionais. Por dados de transação, entende-se aqueles dados que













possuem dois campos em cada registro: o identificador da transação (código) e outro, representando um evento relacionado a esta transação.

O XpertRule Miner foi desenvolvido com base nos conceitos da mineração de dados.

Por isso, apresenta inúmeras características:

- Os scripts de mineração são armazenados em arquivos texto e podem ser executados pelo XpertRule Miner Engine sem a necessidade do ambiente de desenvolvimento do XpertRule Miner;
- Os componentes de árvore de mineração podem ser exportados como componentes embutidos;
- As árvores gráficas podem ser copiadas e coladas em qualquer outro aplicativo capaz de ler o formato WMF.

A mineração funciona pela execução do **script** de transformação como uma seqüência de operações. Exemplos de operações individuais são:

-  conexão de dados de entrada;
-  agregação dos dados;
-  procedimento de derivação de novos campos de dados;
-  filtro de dados;
-  junção de tabelas;
-  comparação entre tabelas;
-  extração de uma amostra randômica;
-  ordenação de tabelas;
-  agregação de série temporal;
-  mineração de dados para árvore;
-  mineração de dados para associações transacionais;
-  geração de uma nova saída para a tabela.

4.4 BUSINESSMINER

O Business Miner é um componente do pacote Business Objects que realiza a tarefa de Mineração de Dados sobre o processamento feito pelo módulo de relatórios/visualização do próprio Business Objects. A técnica de Mineração de Dados implementada pelo Business Miner é a árvore de decisão, sendo esta usada para a construção do modelo de mineração, bem como da visualização dos resultados.

Através da árvore construída pelo Business Miner, é possível responder, por exemplo, a questões sobre o comportamento de certos atributos e a influência destes no resultado final. Adicionalmente, é possível fazer previsões de novos casos através de cada regra gerada e, desta forma, descobrir como se comportaria um cliente hipotético quando colocado em uma nova situação.

4.4.1 Módulos do software

O Business Miner está dividido em 5 janelas denominadas itens do projeto. São elas:

- | | |
|----------------------|--|
| <i>Object</i> | Módulo de definição de todos os campos da base de dados, especificando suas faixas de valores, tipos, etc. |
| <i>Record</i> | Módulo de visualização dos dados contidos na base de dados em forma de tabela. |
| <i>Tree</i> | Módulo em que será feita a geração da árvore de decisão |

Rules	Módulo em que as regras de produção de cada atributo de saída especificado serão descritas
Chart Window	Módulo de criação de gráficos para melhor visualização dos resultados gerados.

4.5 POLYANALYST

4.5.1 Facilidade de uso

Em contraste com outras ferramentas já avaliadas, como o WizRule e WizWhy, o PolyAnalyst da Megaputer Intelligence requer um estudo mais detalhado do manuseio do software para utilizá-lo mais eficientemente. Ele é oferecido em cinco versões:

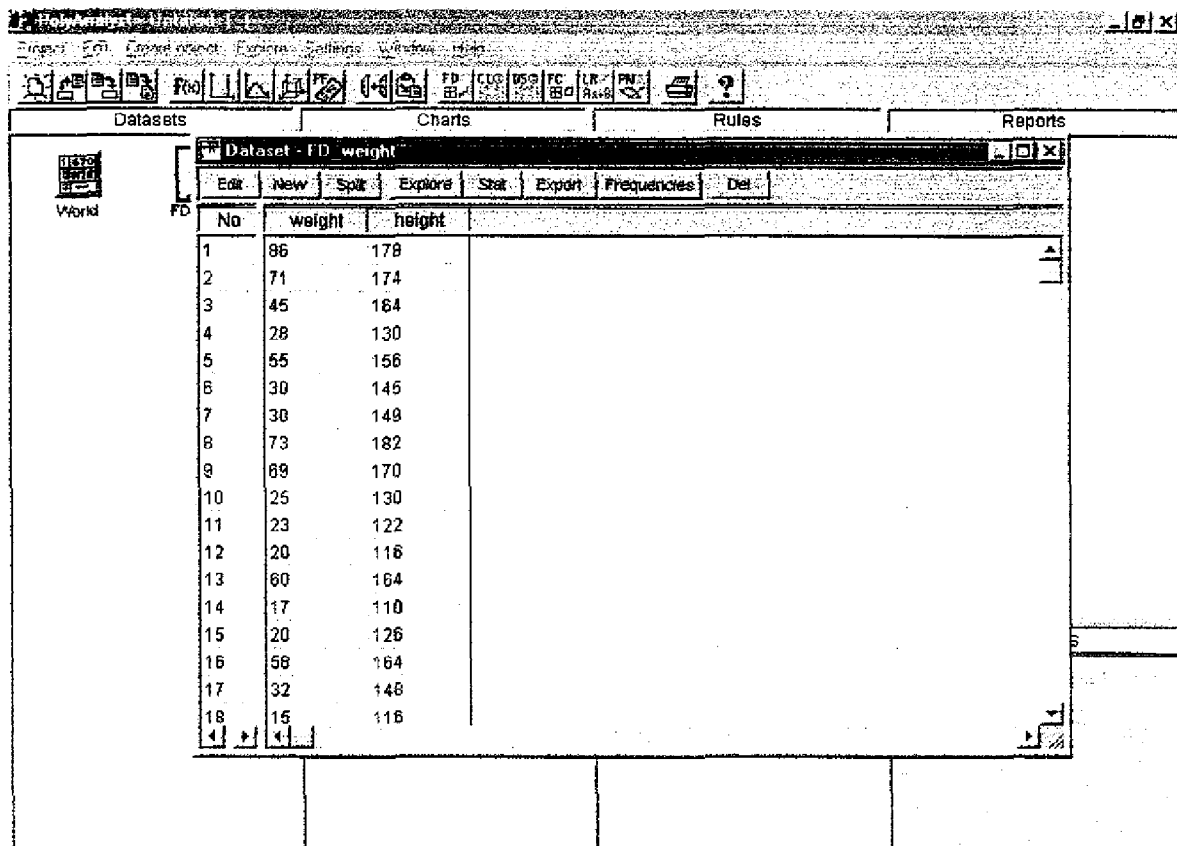
1. PolyAnalyst Lite for Windows 95, Windows 98 and Windows NT
2. PolyAnalyst Professional for Windows NT
3. PolyAnalyst Power for Windows NT
4. PolyAnalyst Power'95 for Windows 95
5. PolyAnalyst Knowledge Server for Windows NT

A versão avaliada foi a Power'95, em "trial-version".

4.5.2 Informações sobre o banco de dados

O software tem recursos para visualização da estrutura e do conteúdo de uma base de dados (“dataset”), que pode ser carregada a partir de arquivos texto no formato “CSV” (Comma-Separated Values), arquivos DBF, planilhas Excel 7 ou 97 ou qualquer banco de dados via ODBC.

Após carregar o dataset, podem ser realizadas análises sob diversos aspectos. Na Figura 4.1 vemos, por exemplo, a primeira análise normalmente realizada: o conteúdo dos registros do dataset.



The screenshot shows a software window titled "Dataset - FD weight". The window has a menu bar with "Edit", "New", "Split", "Explore", "Stat", "Export", "Frequencies", and "Del". Below the menu bar is a table with three columns: "No", "weight", and "height". The table contains 18 rows of data. The software interface also shows a "World" icon on the left and a "FD" label above the table.

No	weight	height
1	86	178
2	71	174
3	45	164
4	28	130
5	55	156
6	30	145
7	30	149
8	73	182
9	69	170
10	25	130
11	23	122
12	20	116
13	60	164
14	17	110
15	20	126
16	58	164
17	32	148
18	15	116

Figura 4.1 - Exemplo da visualização do conteúdo de um dataset

No caso, foi utilizada uma tabela fornecida com o software, que contém dados de peso e altura de uma base de dados para medicina. Ao clicar em “Edit”, surge uma janela com a estrutura do dataset, conforme vemos na Figura 4.2.

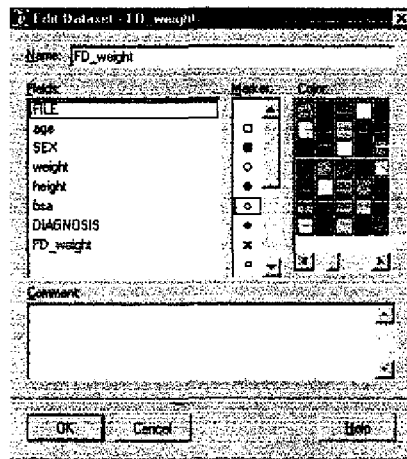


Figura 4.2 - Estrutura do dataset

À esquerda vemos os campos da base de dados (sexo, peso, altura, etc.). Na coluna “Marker” pode-se escolher o tipo de marcador que irá representar o campo em um gráfico, bem como sua cor, à direita.

O botão “Stat” fornece estatísticas sobre o conteúdo do dataset (Figura 4.3).

	Values	Mean	Std.Dev	Min	Max	Range	Median	Mode
weight	81	54.28	23.47	15	100	85	60	
height	81	155.4	22.37	101	184	83	164	

Figura 4.3 - Estatísticas da base de dados

Podemos observar as informações de quantidade de atributos (campos) e de registros, além de estatísticas dos dados em cada campo (quantidade, média, desvio padrão, mínimo, máximo, diferença entre o mínimo e máximo e mediana).

O módulo principal do produto é a análise de dependência entre atributos. Na Figura 4.4 vemos a análise do peso de um conjunto de indivíduos em função de sua altura.

O resultado é apresentado em três janelas, conforme a Figura 4.4, que podemos denominar: Regra Tabular (superior), Análise do Erro (inferior à direita) e Residuais (inferior à esquerda).

Na janela de Regra Tabular é exibida uma tabela em que as colunas são faixas de valores de altura e há cinco linhas de informações:

1. Valor previsto
2. N° de ocorrências da faixa
3. Erro padrão total
4. N° de ocorrências em que a regra é válida
5. Erro padrão nas ocorrências em que a regra é válida

Na janela de Análise do Erro os pontos azuis indicam o valor previsto pela regra e os pontos vermelhos, as ocorrências reais no dataset. Quanto mais próximos os azuis dos vermelhos e, portanto, da diagonal, mais acurada estará a fórmula.

Na janela de Residuais é exibido um gráfico que mostra o percentual de ocorrências para uma determinada margem de erro. Por exemplo, observa-se que em 80% dos casos o erro é menor que 5Kg.

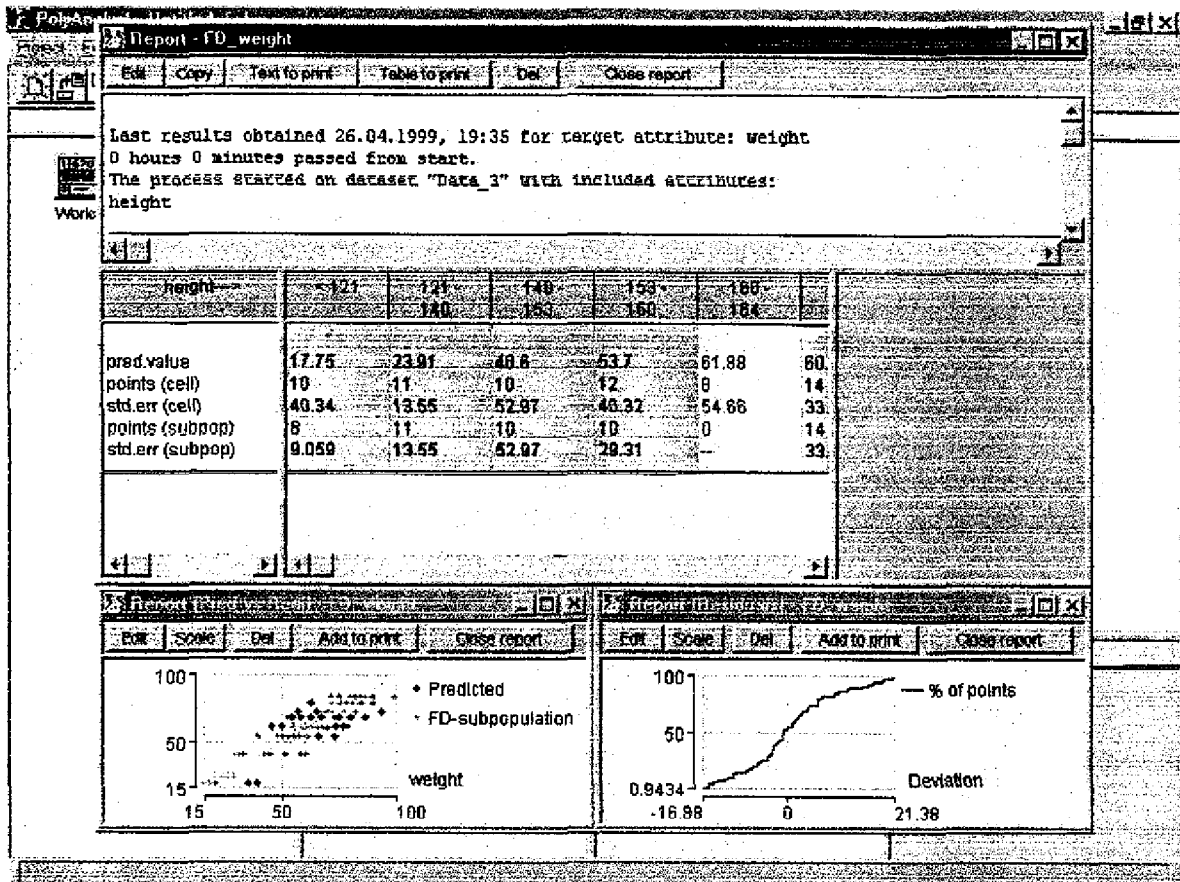


Figura 4.4 - Análise de dependência entre atributos

O módulo de Regressão Linear tenta definir uma fórmula relacionando atributos, conforme o exemplo ilustrado na Figura 4.5 (peso em função da altura).

Isto funciona como um detalhamento do módulo tabular visto anteriormente. Aqui, pode-se avaliar valores contínuos, não apenas faixas de valores como no caso anterior.

Na janela superior são exibidos os dados estatísticos do resultado da fórmula. Abaixo à esquerda, uma visualização dos valores encontrados pela fórmula. De novo, quanto mais próximo à diagonal, mais preciso o resultado. Abaixo à direita, a contribuição de cada termo (atributo) para o resultado. Como foi tomado apenas um atributo (altura), não há outros para comparar.

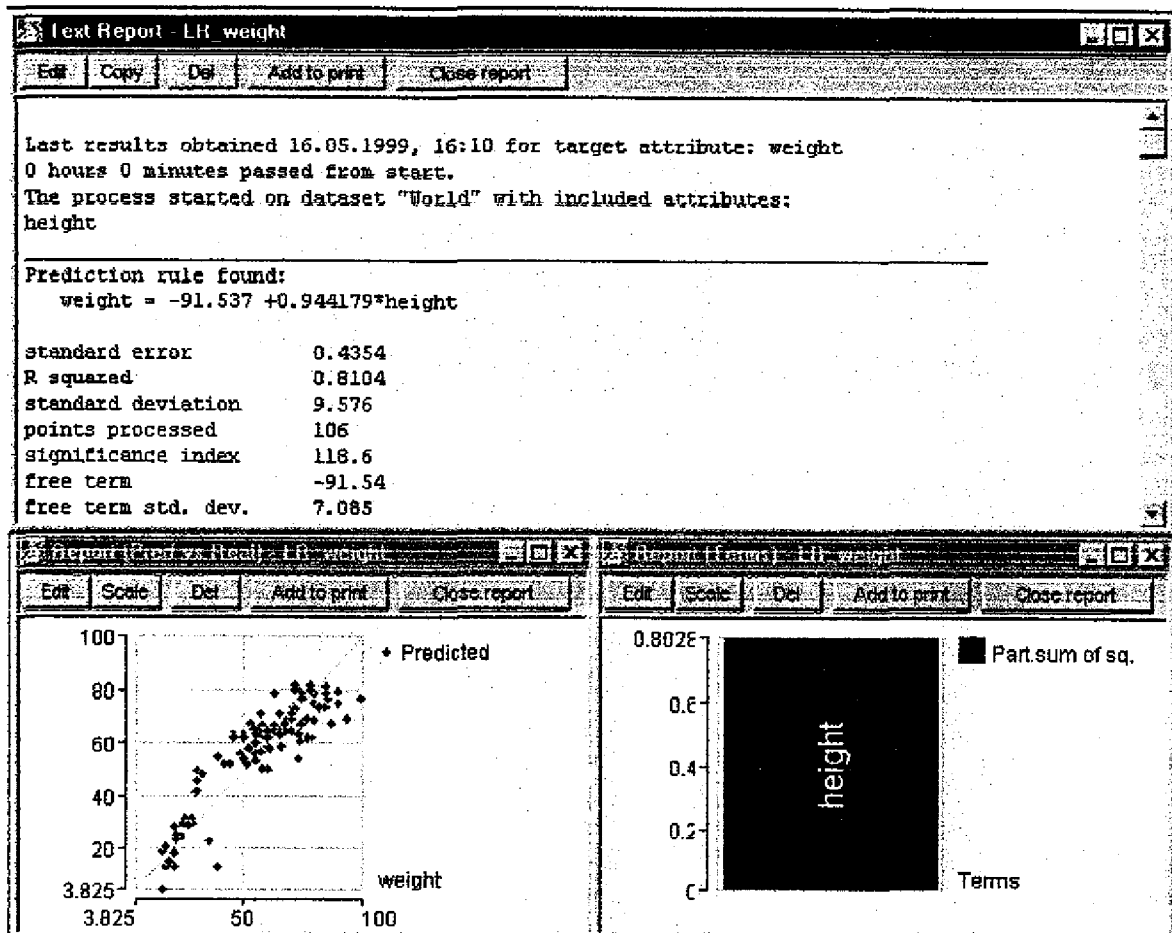


Figura 4.5 - Resultado da Regressão Linear

Há ainda um módulo gerador de gráficos, que também facilita a visualização da variação de um elemento em função de outro, conforme pode ser visto na Figura 4.6.

No caso, vemos as ocorrências de peso no dataset em função da altura.

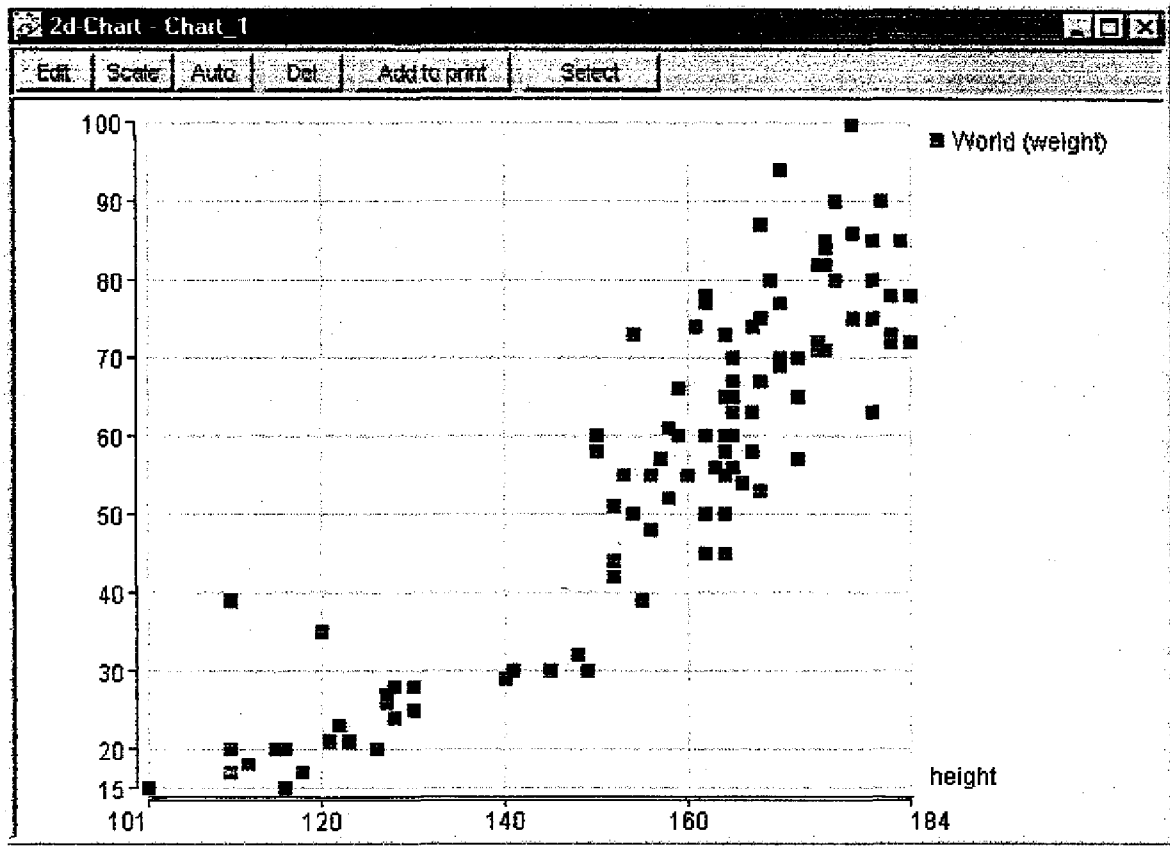


Figura 4.6 - Módulo de geração de gráficos

4.5.3 Conclusões

O PolyAnalyst cumpre bem seu objetivo: criar relacionamentos entre informações de uma base de dados através de fórmulas, diferenciando-o de ferramentas como o WizWhy, que criam estes relacionamentos através de regras de inferência. Seu processo de descoberta é baseado em Programação Genética para estabelecer o formato das fórmulas, o que torna ainda mais impressionante uma de suas qualidades: boa performance.

4.6 CONCLUSÕES SOBRE OS SOFTWARES AVALIADOS

Depois de avaliados os softwares de Mineração de Dados apresentados nos itens 4.1 a 4.5 deste capítulo, ficou claramente evidente a necessidade da tarefa de preparação de dados, descrita pelo processo de KDD. Isto porque, na maioria das vezes, os dados se apresentam em diferentes formatos e com algumas “impurezas” (dados inválidos). Neste caso, é preciso que estes dados sejam corrigidos ou, até mesmo, eliminados da tabela.

Como já visto anteriormente, a fase de preparação de dados engloba cerca de 80% do tempo de todo o processo mas, curiosamente, não é tratada com a importância que deveria, por grande parte das ferramentas disponíveis no mercado. Já a fase de mineração de dados propriamente dita, é realizada de diferentes maneiras em cada um deles e seus resultados pareceram bastante satisfatórios.

Outro aspecto interessante de se analisar, é o fato de que algumas ferramentas se preocupam com a confiabilidade das regras geradas, enquanto outras visam a alta abrangência (população para cada regra) destas.

Desta forma, percebemos que não há como trabalhar com um software individualmente. Cada um possui suas funcionalidades próprias e podem ser bastante úteis se usados em conjunto com outros. O importante é conhecer o problema, estabelecer um objetivo e sair em busca dos resultados, utilizando, para isso, todo e qualquer benefício encontrado.

Uma das propostas do projeto aqui desenvolvido, o BRAMINING, é exatamente proporcionar uma abrangência maior das diversas fases de KDD, permitindo que se trabalhe em seu ambiente praticamente do início ao fim do processo. O BRAMINING é uma interface inteligente, evoluída em diversas teses de mestrado, que comporta-se como um *wizard*, ou assistente, do processo de KDD. No capítulo a seguir o projeto será abordado em detalhes.

5 PROJETO BRAMINING

5.1 INTRODUÇÃO

A maioria dos programas disponíveis no mercado, inclusive aqueles avaliados no Capítulo 4 desta dissertação, assume que a base de dados utilizada apresenta-se limpa, padronizada e, portanto, pronta para a execução da mineração dos dados nela contidos. São ambientes altamente preparados para a geração de regras de produção e criação de árvores de decisão, mas que ignoram o estado dos dados ali presentes. Neste casos, seria preciso que houvesse um tratamento nos dados (redução, limpeza e padronização) feito de forma externa ao programa para que os resultados gerados pelos algoritmos de mineração fossem considerados satisfatórios.

Para utilizar estes programas de higienização e padronização de dados independentes, as empresas têm que se submeter a um custo muito alto, deixando, muitas vezes, de adquiri-los e, conseqüentemente, de preparar seus dados de forma adequada.

Para amenizar este problema, foi desenvolvido um ambiente de preparação e tratamento da base de dados, além do ambiente de mineração, mas que também pode ser utilizado de forma independente, como base para os demais softwares de mineração de dados. A este projeto foi dado o nome de BRAMINING.

A ferramenta desenvolvida, apresentada neste capítulo, abrange as duas etapas principais do processo de KDD – preparação e mineração de dados. O objetivo desta ferramenta é apoiar o analista de dados nas etapas em que seu conhecimento sobre o negócio avaliado é indispensável, porém pode ser automatizado. É importante notar que cada módulo da ferramenta é independente dos demais, de forma a possibilitar que cada alteração na base

de dados utilizada sirva como preparação para a mineração dos dados nesta ou em qualquer outra ferramenta.

5.2 A IMPLEMENTAÇÃO

Para a implementação do projeto foi escolhida a linguagem Borland Delphi na versão 4, por suas características de estabilidade e pela boa estruturação de suas classes de objetos de acesso a bancos de dados. O programa possui cerca de 1500 linhas de código.

A ferramenta lê arquivos no formato "dbf" para entrada dos dados e gera saídas (relatórios de regras) em arquivos padrão texto puro.

Na Figura 5.1 é exibido o DFD 0 do sistema. Nele pode ser vista forma com que os dados entram no sistema, são tratados e armazenados para as etapas futuras.

A Figura 5.1 mostra que o analista de dados é uma entidade externa ao processamento dos dados e que a interação do usuário com o analista é dada fora do foco do diagrama. Como prova da iteratividade do processo, vale destacar que o analista de dados pode submeter a mesma massa de dados à ferramenta, quantas vezes forem necessárias, até que os resultados aproximem-se de um determinado objetivo.

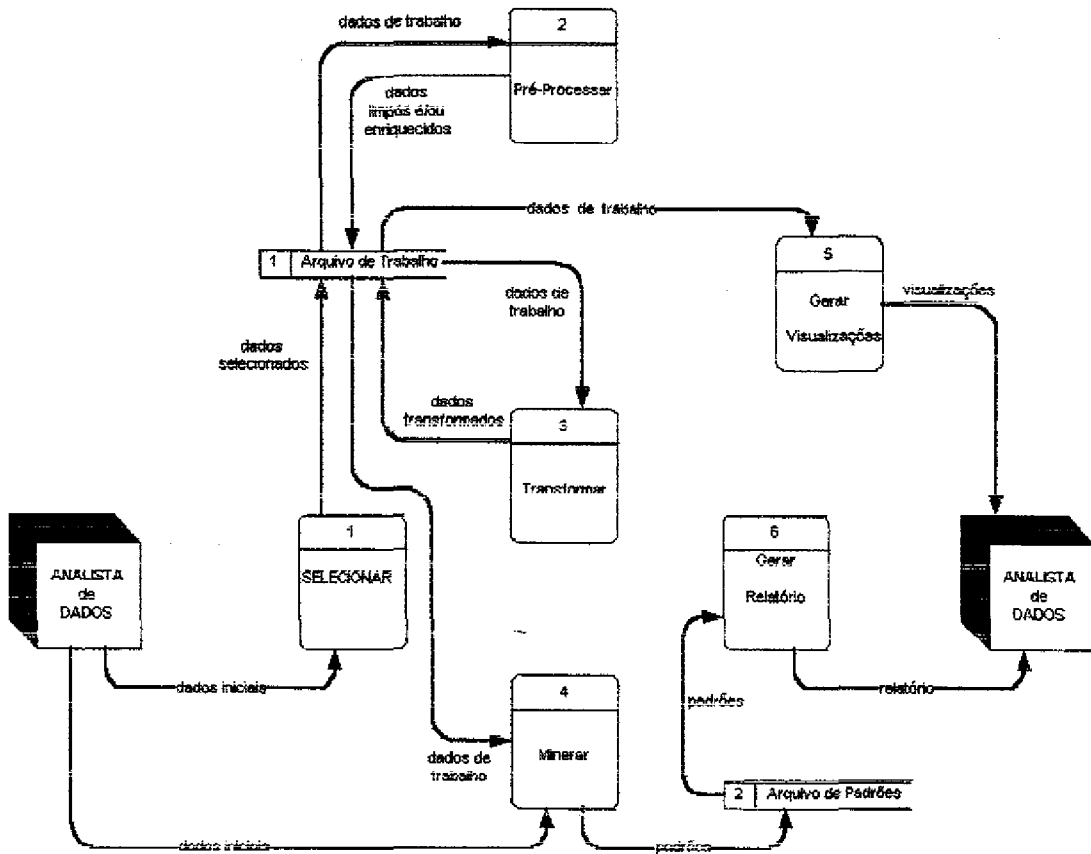


Figura 5.1 - DFD do ambiente de KDD desenvolvido.

5.3 UTILIZANDO A FERRAMENTA

A Figura 5.2 exibe a visão inicial da ferramenta ao ser iniciada. Uma seqüência de janelas (identificadas como KDD, Seleção, Pré-Processamento, Codificação, Mineração e Relatório) permite selecionar uma das etapas do processo de KDD. A primeira delas - KDD - exibe uma figura explicativa deste processo e um resumo do que pode ser realizado em cada uma de suas fases.

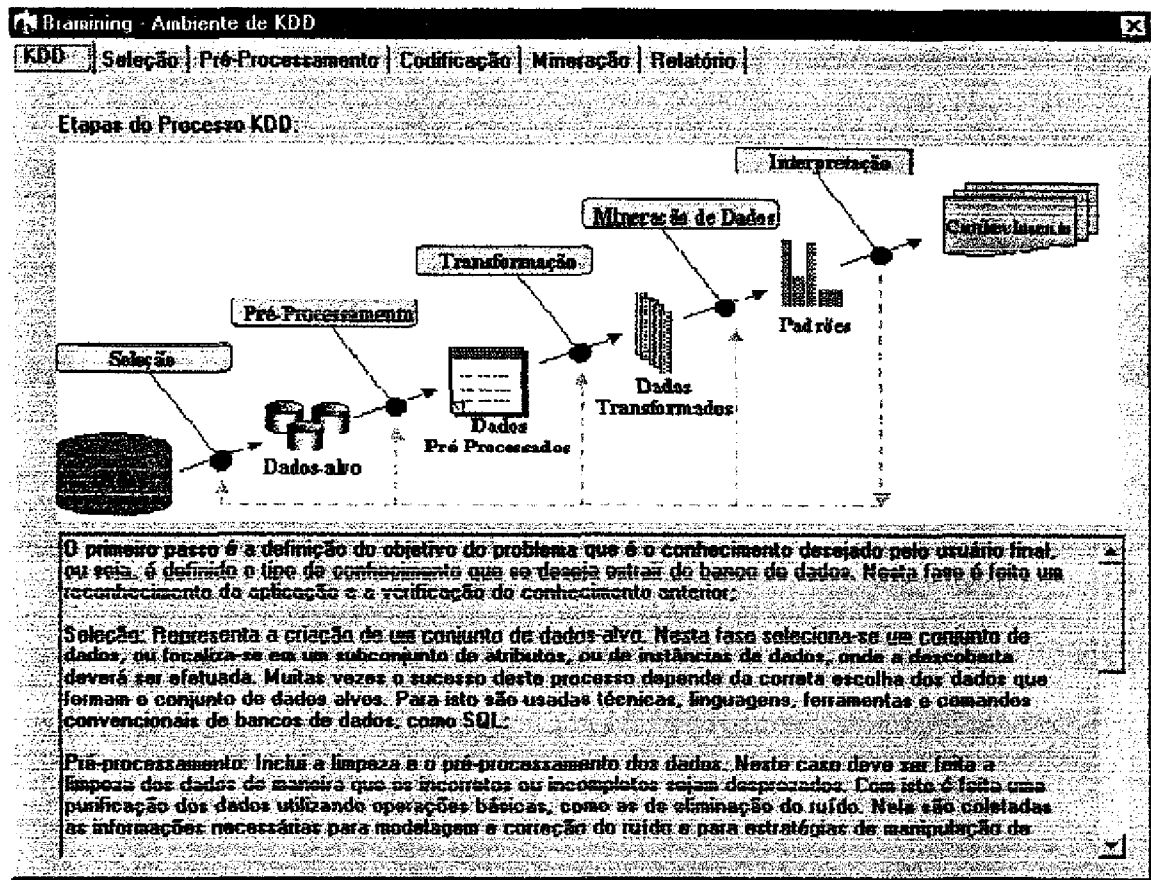


Figura 5.2 - Tela inicial da ferramenta desenvolvida para KDD.

As etapas indicadas na Figura 5.2 e descritas a seguir foram detalhadas no Capítulo 2:

- 1) **Seleção**: escolha da base de dados a ser trabalhada;
- 2) **Pré-Processamento**: redução, limpeza e padronização da base de dados selecionada;
- 3) **Transformação**: preparação dos dados em um formato padrão;
- 4) **Mineração de Dados**: obtenção dos padrões desejados;
- 5) **Avaliação e Interpretação**: crítica dos resultados obtidos.

Existe uma certa limitação quanto ao percurso a ser adotado, apesar da manipulação de dados ser iterativa. Isto porque é preciso que alguns parâmetros sejam especificados para que a mineração de dados seja realizada com sucesso. Desta forma, sugere-se que:

- Uma base de dados seja selecionada;
- A base selecionada seja tratada (reduzida e padronizada);
- Seja feita a codificação dos valores, quando necessário;
- A fase de mineração só seja iniciada após os dados terem sido efetivamente preparados para a tarefa a que serão submetidos;

5.3.1 Seleção de dados

Na janela Seleção, além da escolha da base de dados a ser trabalhada, é permitida uma visualização destes dados, de modo a identificar todos os atributos nela contidos, bem como a forma com que os dados estão apresentados. Isso equivale a entender o problema em questão e definir que atributos são relevantes para a compreensão do mesmo.

A Figura 5.3 e a Figura 5.4 apresentam, respectivamente, a seleção e a representação da base de dados utilizada como exemplo.

A base de dados em estudo contém informações do setor de telecomunicações acerca dos hábitos de ligação de cada cliente. A base possui cerca de 325.000 registros e 9 atributos que permitem caracterizar um determinado cliente com um certo perfil.

O objetivo deste estudo é identificar o perfil residencial ou comercial nas ligações de cada cliente. Este perfil é traçado com base em parâmetros (frequência das ligações, horário, dia, por exemplo) que nos permitem atribuir um grau de certeza para a caracterização gerada.

Os dados contidos na base em questão são apresentados na Tabela 5.1.

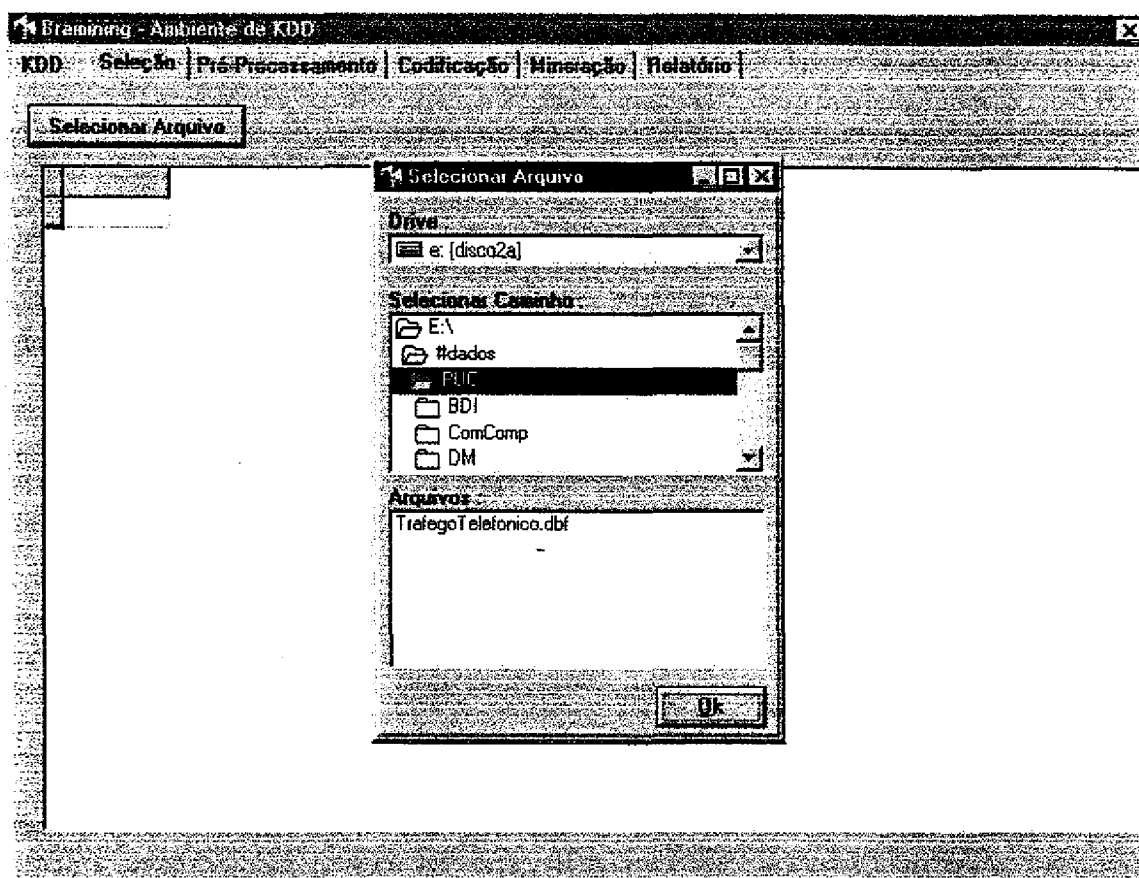


Figura 5.3 - Tela indicando a seleção de dados.

CAMPO	DESCRIÇÃO
Assinante	Assinante de origem, aquele do qual partiram as ligações
Tipo_ligação	1 – DDD, 2– Local, 3 – A cobrar, 4 – 0800, 5 – Celular
Faixa_tarifação	1 – Normal, 2 – Reduzido, 3 – Super Reduzido, 4 – Misto, 5 – Diferenciado
Distância	Distância física entre a central telefônica do assinante de origem e de destino
Fim_semana	S – Sábado e Domingo, N – Dia de semana
Minutos	Duração total das ligações, em minutos

Receita	Valor gasto com as ligações
Num_ligações	Quantidade de ligações feitas no mês
Tipo_cliente	R – Residencial, C – Comercial

Tabela 5.1 - Atributos contidos na base de dados de telecomunicações

É preciso um grande conhecimento da base de dados em estudo para que possam ser identificados os atributos de maior relevância para a geração de um bom resultado final. Para o caso em questão, os atributos assinante e distância serão descartados por não agregarem qualquer tipo de informação ao problema proposto.

ASSINANTE	TIPO_CLIEN	TIPO_LIGAC	FAIXA_TARI	DISTANCIA	FIM_SEMANA	MINUTOS	RECEITA	LI
258	C	4	5	3	N	1	0,27	
259	C	1	5	1	N	12	1,7	
260	C	1	5	2	N	9	1,69	
261	C	1	5	3	N	3	0,86	
262	C	1	5	5	N	4	0,2	
263	C	3	1	1	N	0	0,05	
264	C	3	1	4	N	1	0,18	
265	C	3	5	1	N	4	0,59	
266	C	3	1	1	N	1	0,08	
267	C	4	5	4	N	17	6,3	
268	C	1	5	1	N	81	10,45	
269	C	1	5	1	N	724	79,75	
270	C	1	3	1	N	16	0,26	
271	C	1	2	5	S	4	0,05	
272	C	1	2	4	S	27	2,51	
273	C	1	2	4	N	19	1,95	
274	C	1	2	3	N	39	2,61	
275	C	1	1	1	N	3	0,22	
276	C	1	3	4	N	9	0,42	
277	C	1	1	5	N	9	0,24	

Figura 5.4 - Representação da base de dados selecionada na Figura 5.3.

5.3.2 Pré-processamento dos dados

Esta etapa pode compreender duas tarefas distintas: redução e padronização. Mas é necessário que o analista saiba identificar quais atributos podem ser descartados, por não agregarem informação valiosa ao processamento. Além disso, também é preciso que sejam escolhidas as formas de apresentação dos dados. Por exemplo, no campo Tipo_cliente, encontramos valores distintos mas que possuem o mesmo significado: Com, C. Neste caso, é possível substituir todas as ocorrências de Com por C, ou vice-versa.

Para que o usuário chegue a este tipo de conclusão, é gerada uma tela, Figura 5.5, contendo algumas estatísticas acerca da base de dados selecionada. Informações do tipo quantidade de registros em branco e uma lista dos valores distintos para cada atributo da tela, ajudam demasiadamente na definição dos padrões a serem utilizados.

Por convenção, os atributos com até 10 valores distintos na base de dados, terão seus valores mostrados um a um; porém, aqueles cuja quantidade de valores distintos ultrapassar o limite estipulado, serão considerados contínuos e representados por uma faixa de valores, indicando o menor e o maior deles.

A Figura 5.5 mostra isso claramente: os atributos Assinante, Receita, Minutos e Num_ligações, têm apenas seus menores e maiores valores especificados, enquanto que todos os demais, são apresentados distintamente.

Como indicações para limpeza da base de dados, a Figura 5.5 apresenta duas opções: Redução e Padronização de dados. Clicando no botão Redução de dados, uma nova tela, Figura 5.6, é mostrada, e nela devem ser escolhidos os atributos relevantes ao restante do processo. Aqueles atributos marcados com S, são considerados importantes e serão utilizados

nas etapas seguintes. Já os marcados com N, serão descartados no momento em que o analista definir um novo nome para esta base de dados, agora, possivelmente reduzida.

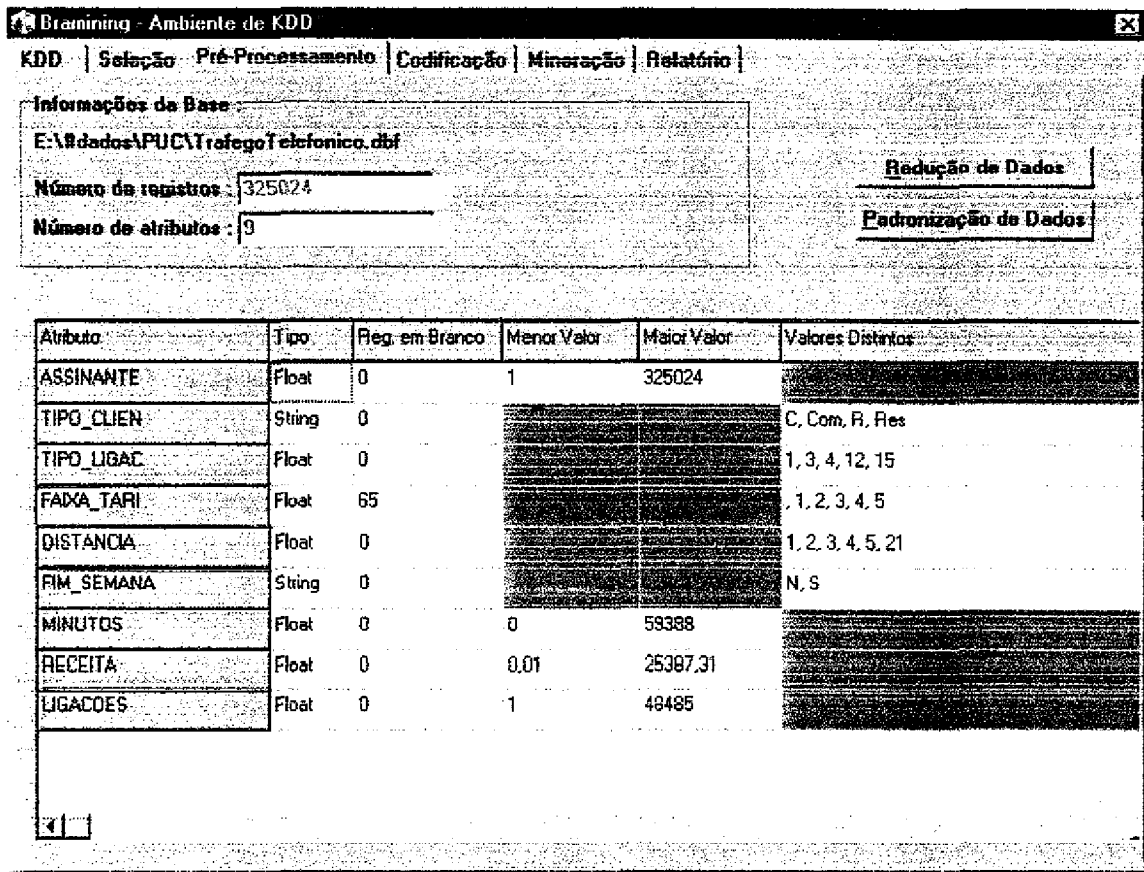


Figura 5.5 - Estatísticas acerca da base de dados.

É imprescindível que um novo nome seja especificado para a base em questão, uma vez que alterações poderão ser feitas e, portanto, será preciso manter a base original.

Como exemplo da redução de dados vemos, na Figura 5.6, que os atributos Assinante e Distância foram descartados por não terem relevância no resultado final do processamento. Para selecionar os atributos que serão excluídos do processo, é preciso clicar duas vezes no campo Salvar ou clicar uma vez e teclar Enter, para que a opção S (Sim) seja trocada para N (Não). Feito isto, basta especificar um novo nome para a tabela a ser gerada.

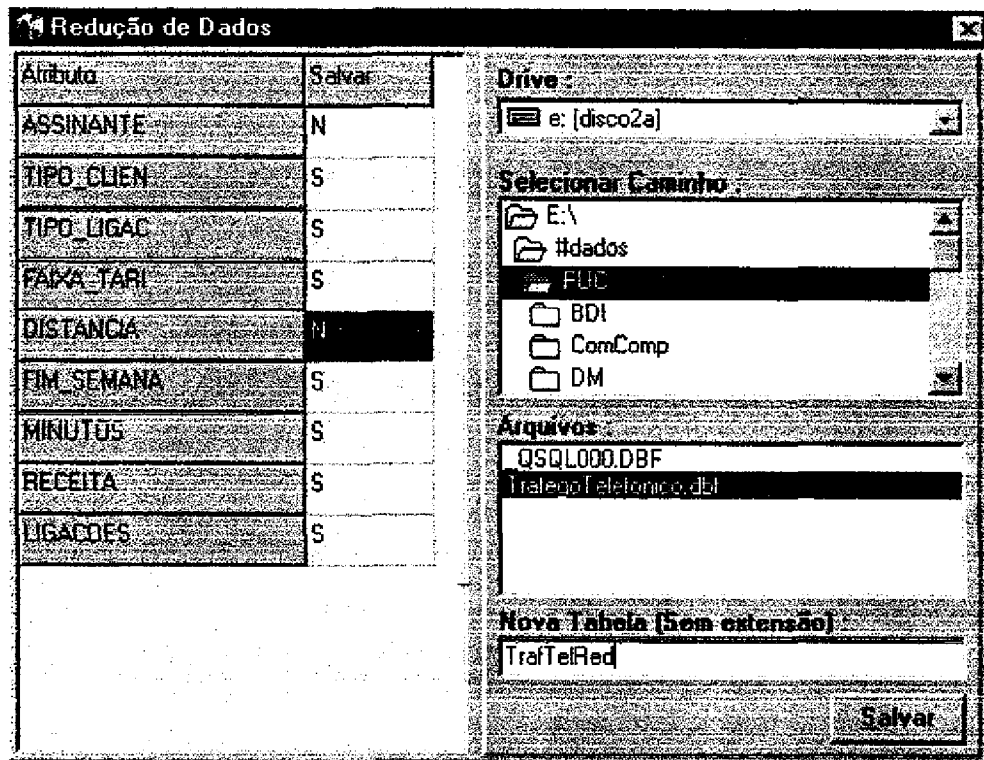


Figura 5.6 - Redução dos dados e geração de uma nova base.

Veja que uma nova estatística é gerada, na Figura 5.7, referente à base agora reduzida.

Terminada a redução dos dados, uma nova opção é apresentada. Clicando no botão Padronização de dados, os atributos selecionados com S na Figura 5.6 serão mostrados. Para cada um deles, serão permitidas substituições ou deleções. Os tipos de substituição oferecidos são: pela média dos valores daquele atributo, pelo valor mais freqüente do mesmo atributo ou por um novo valor, definido pelo analista de dados. Já as deleções, podem ser feitas especificando-se o atributo, a operação desejada e o valor.

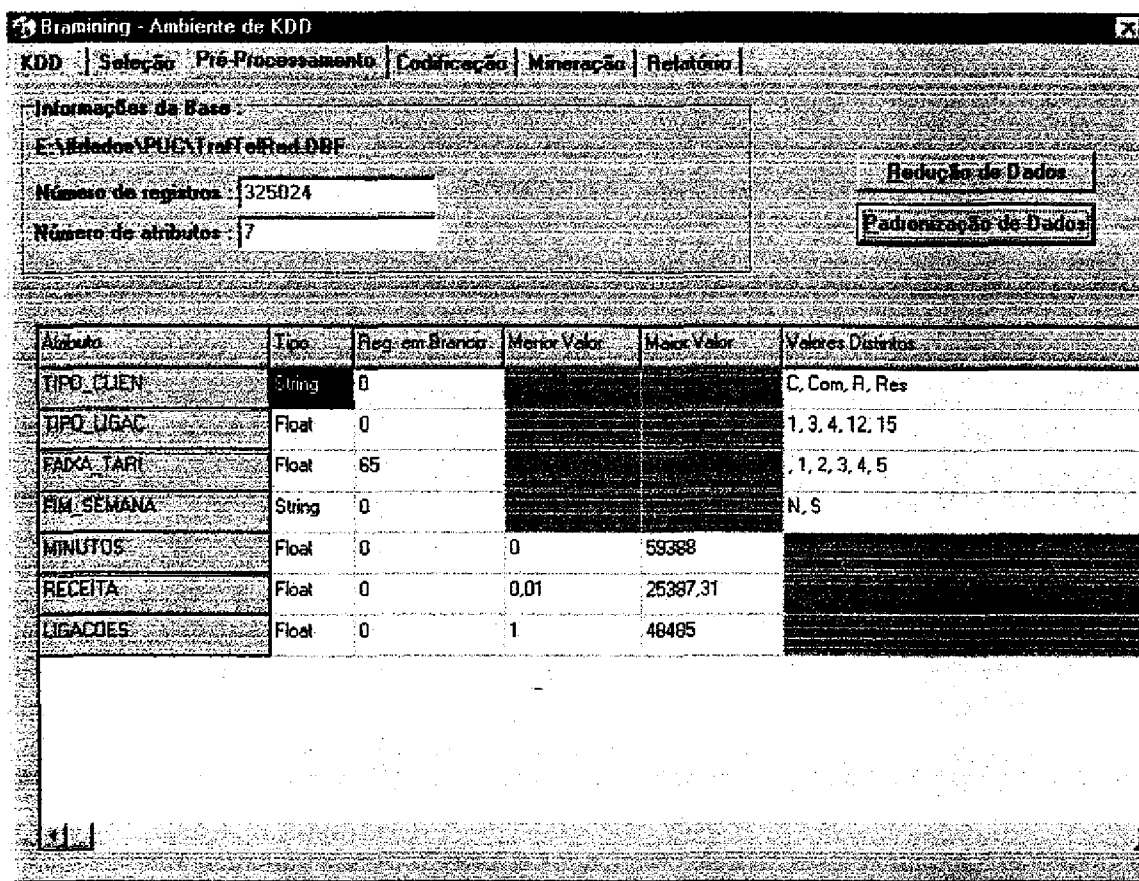


Figura 5.7 - Estatísticas acerca da nova base de dados.

Sugere-se que a primeira tarefa a ser realizada seja a substituição e, a seguir, a exclusão. No caso apresentado na Figura 5.8 e na Figura 5.9, por exemplo, as substituições vão sendo feitas gradativamente. O primeiro atributo a sofrer alterações é o Tipo_cliente, em que todas as ocorrências do valor Com serão substituídas por um valor a ser definido pelo analista de dados, neste caso, C; já que é sabido pelo analista de dados que C e Com significam comercial e R e Res, significam residencial. Outro campo que merece atenção é Tipo_ligação, pois como apresentado na Figura 5.5, há valores considerados inválidos, ou pelo menos, não especificados no item 5.3.1. Desta forma, os valores 12 e 15 do atributo em questão, deverão ser substituídos pelo valor mais frequente deste campo ou, se preferido, simplesmente excluídos da base de dados.

Especificadas as alterações a serem feitas, basta clicar no botão Padronizar e uma caixa de diálogo indicará o número de substituições efetuadas. No caso de uma substituição

pela média do atributo ou pelo valor mais freqüente, uma caixa de diálogo aparecerá para informar o valor calculado (média ou mais freqüente).

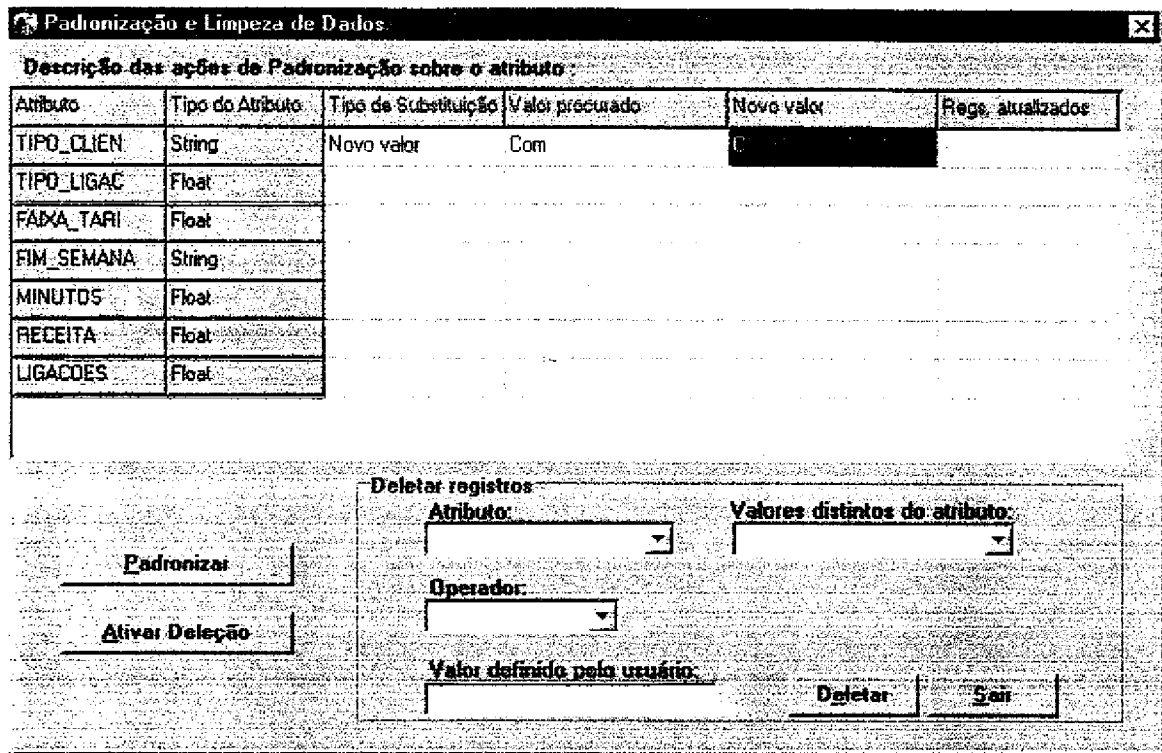


Figura 5.8 - Primeira rodada de substituições e deleções de atributos da base de dados.

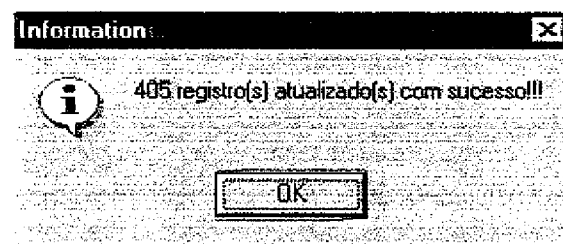


Figura 5.9 - Indicação das substituições do valor Com por C.

A segunda etapa das substituições, apresentada na Figura 5.10, dá continuidade ao processo iniciado na Figura 5.8. O atributo Tipo_cliente, desta vez, terá todas as ocorrências do valor Res alteradas para R;

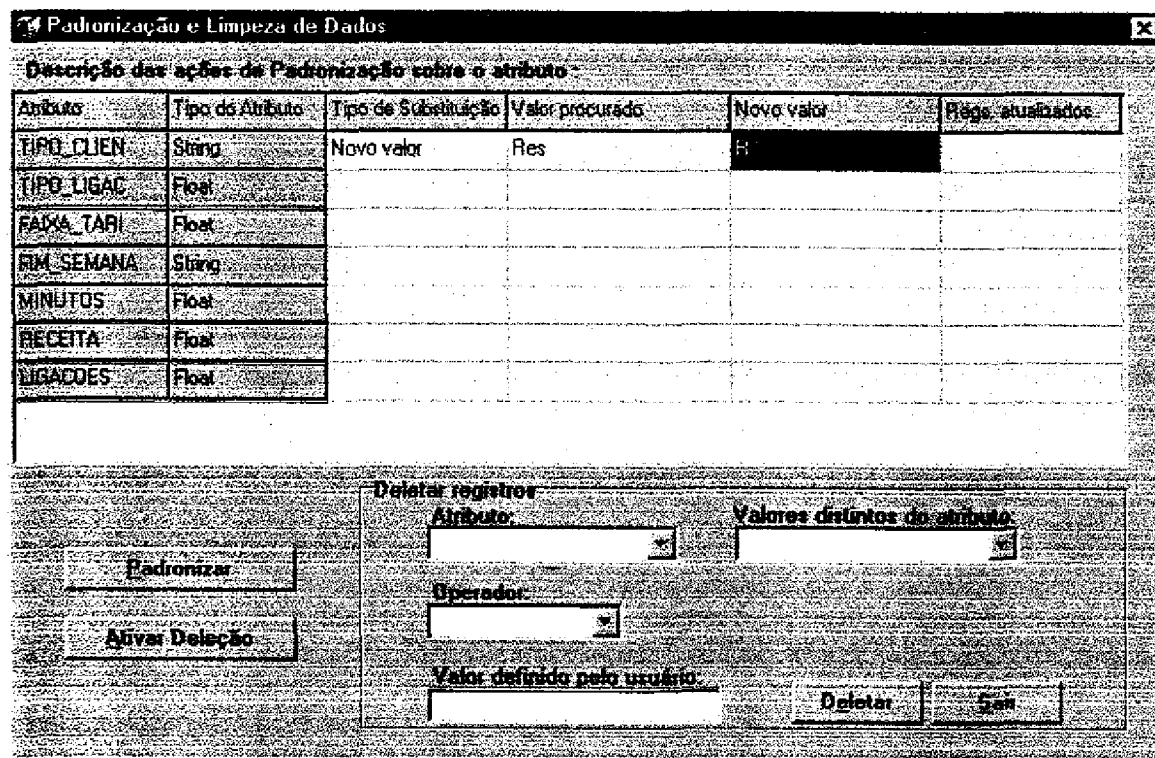


Figura 5.10 - Segunda rodada de substituições e deleções da base de dados.

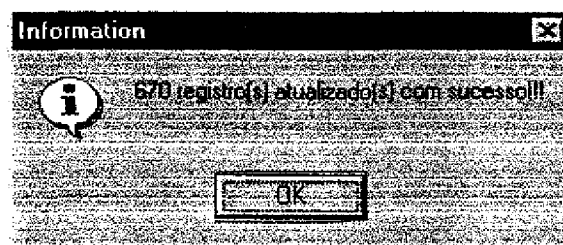


Figura 5.11 - Indicação da substituição do valor Res por R.

Sempre que os tipos de substituição Média ou Valor + freqüente forem selecionados, a célula Novo valor ficará desabilitada. Isto porque não será necessário especificar um novo valor, pois o cálculo do valor médio ou do mais freqüente, é realizado pelo próprio programa. Neste caso, só será preciso definir o valor a ser substituído, ou seja, o Valor procurado.

As alterações feitas na Figura 5.8 e Figura 5.10 serão salvas na tabela definida na Figura 5.6, de forma a não modificar a tabela original selecionada na Figura 5.4.

A fase de exclusão de atributos é iniciada assim que as substituições são concluídas. Se houver ocorrências de valor nulo para algum dos atributos da tabela, é adequado apagar todas as linhas em que isto ocorre. Neste caso, selecionamos o atributo de interesse na caixa Atributo, e o operador IS NULL. Mas caso haja um valor específico a ser excluído de algum campo - como é apresentado na Figura 5.5 referente às estatísticas da base de dados, em que os valores 12 e 15 aparecem para o atributo Tipo_ligação - é preciso escolher o atributo, o valor a ser excluído na caixa Valores distintos do atributo e um dos operadores de exclusão, =, <, >, <=, >=, like, is null ou is not null na caixa Operador. Feito isto, clicando no botão Deletar, as exclusões serão realizadas. Isto é ilustrado da Figura 5.12 à Figura 5.17.

The screenshot shows a window titled "Padronização e Limpeza de Dados" with a sub-header "Descrição das ações de Padronização sobre o atributo:". Below this is a table with the following data:

Atributo	Tipo do Atributo	Tipo de Substituição	Valor procurado	Novo valor	Regs. atualizados
TIPO_CLIEN	String				
TIPO_LIGAC	Float				
FAIXA_TARI	Float				
FIM_SEMANA	String				
MINUTOS	Float				
RECEITA	Float				
LIGACDES	Float				

Below the table, there are two buttons: "Padronizar" and "Ativar Deleção". To the right, there is a "Deletar registros" section with the following fields:

- Atributo:** FAIXA_TARI
- Operador:** IS NULL
- Valores distintos do atributo:** (empty)
- Valor definido pelo usuário:** (empty)

At the bottom right of this section are two buttons: "Deletar" and "Sair".

Figura 5.12 - Exclusão do valor nulo do atributo Faixa_Tarifação.

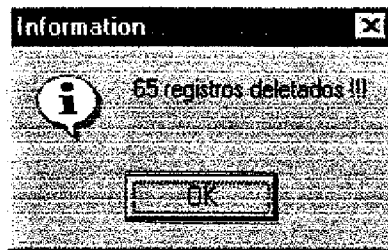


Figura 5.13 - Indicação do número de registros deletados.

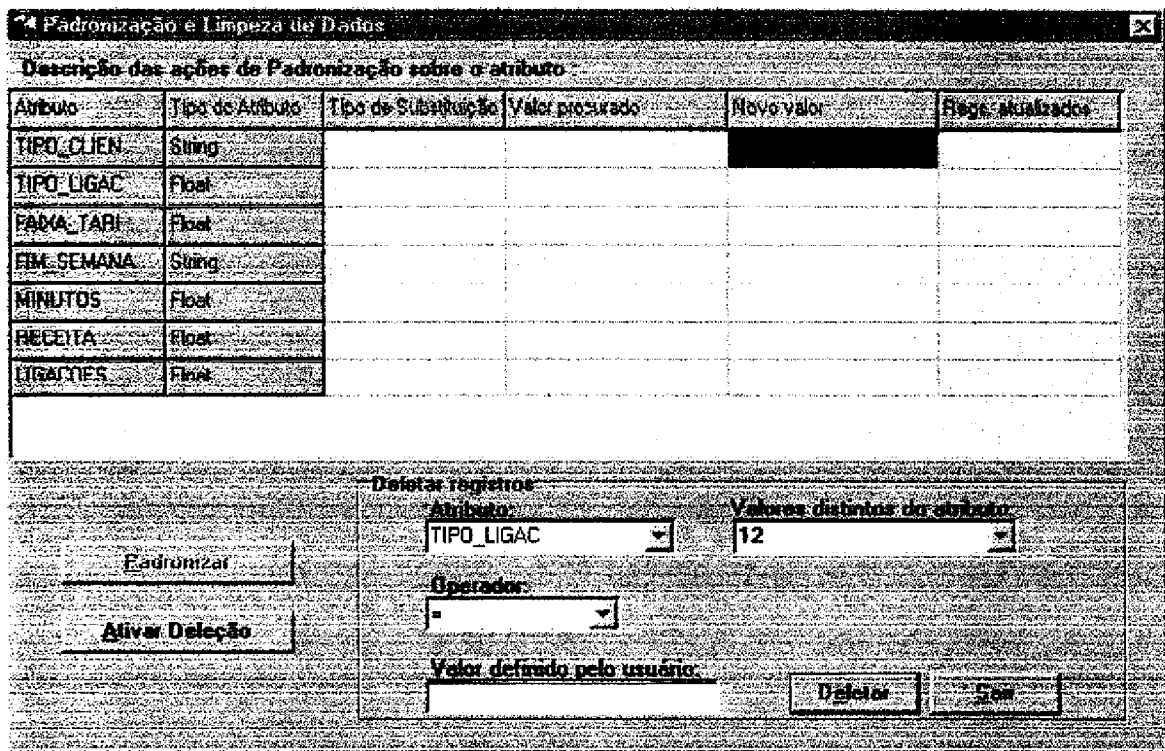


Figura 5.14 - Exclusão do valor 12 do atributo Tipo_Ligação.

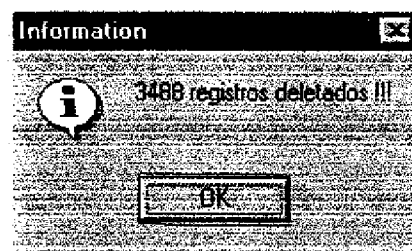


Figura 5.15 - Indicação do número de registros deletados.

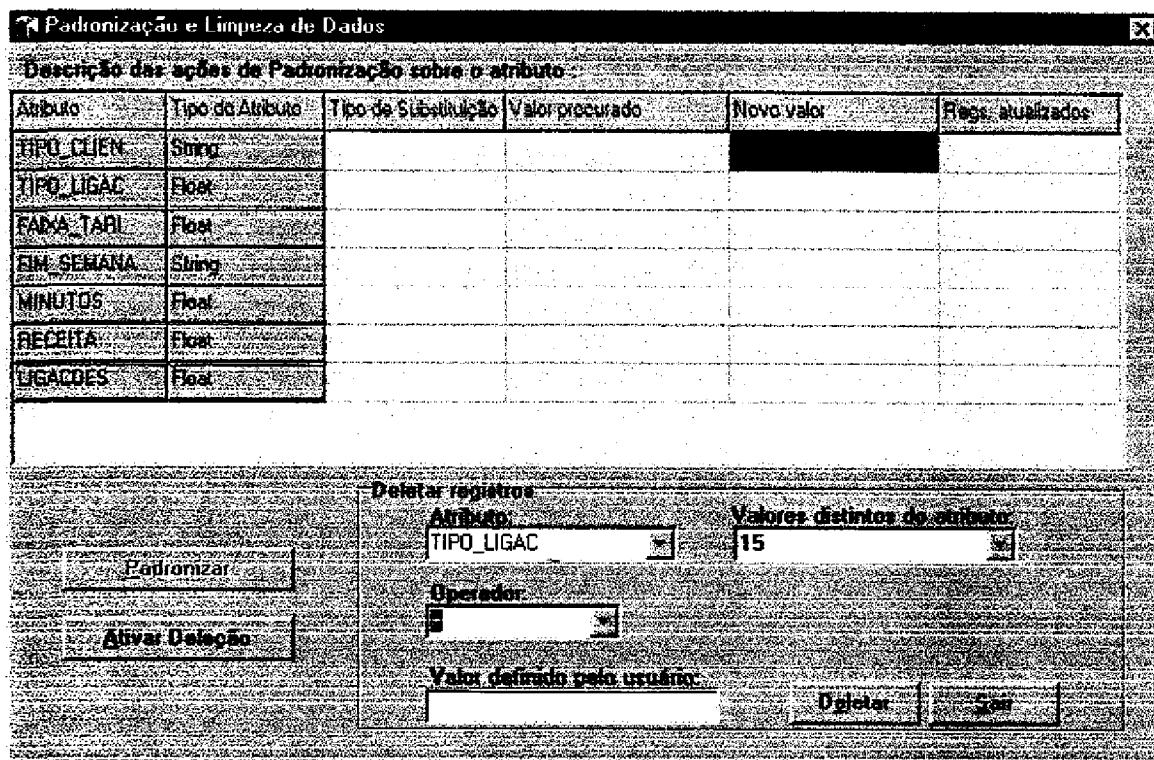


Figura 5.16 - Exclusão do valor 15 do atributo Tipo_Ligação.

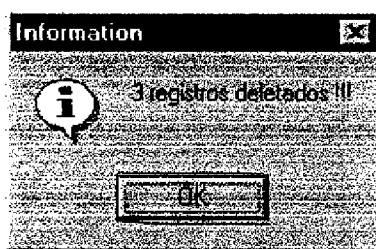


Figura 5.17 - Indicação do número de registros deletados.

5.3.3 Codificação

É a forma através da qual se diminui a grande variação encontrada nos valores dos atributos. É necessária como fase de preparação para a mineração de dados em redes neurais onde só se admite atributos numéricos como entrada no processamento.

Como exemplo de codificação, temos os casos da Figura 5.18. Há dois tipos de codificação, aquele em que um atributo apresenta uma quantidade muito grande de valores distintos e, portanto, não pode ser utilizado como entrada de uma rede neural e outro, em que a quantidade de valores distintos para cada atributo não é tão grande assim. No primeiro caso, é preciso que faixas de valores sejam definidas. No segundo, é atribuído um único número para cada valor.

Para o estudo em questão, não há necessidade de codificação dos atributos, uma vez que o método de mineração de dados implementado na janela Mineração de Dados, não é uma rede neural.

Porém, a título de ilustração, o atributo Minutos será codificado com faixas de valores, enquanto Fim_de_semana será codificado sem faixas.

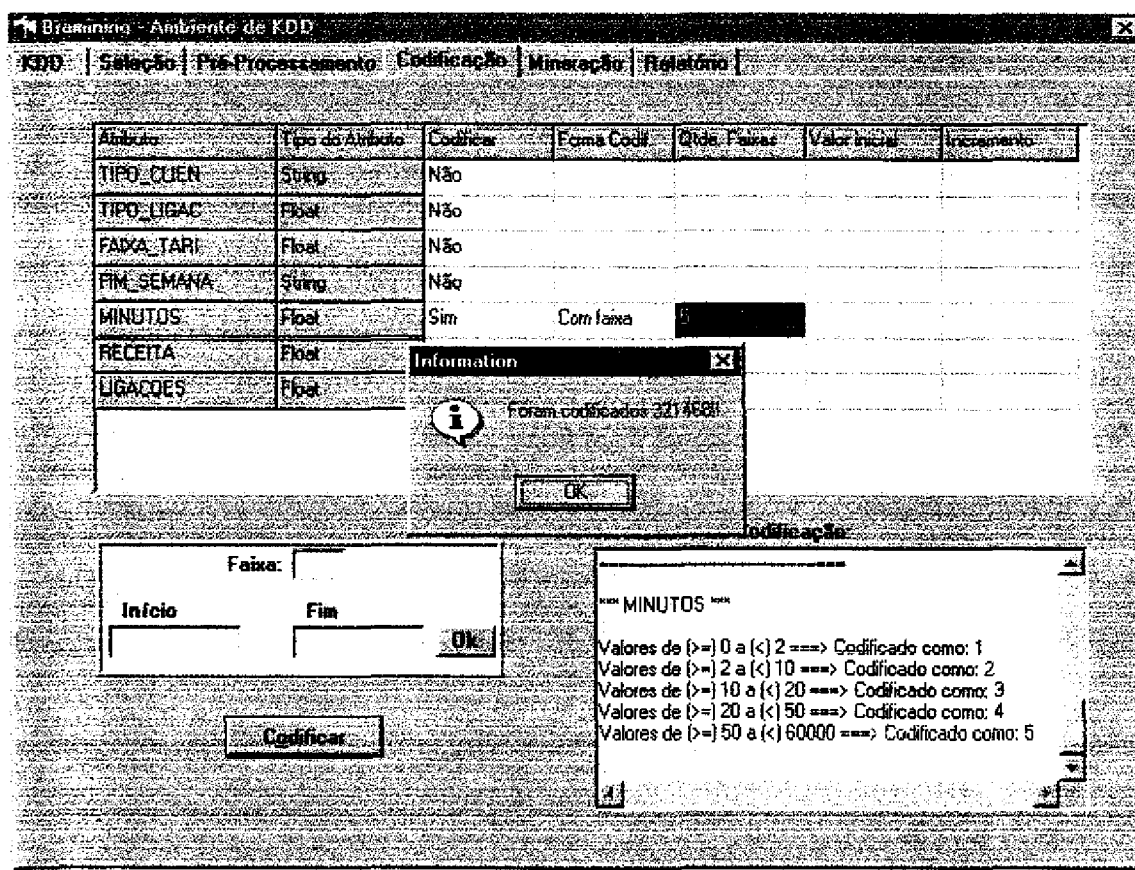


Figura 5.18 - Codificação do atributo Minutos.

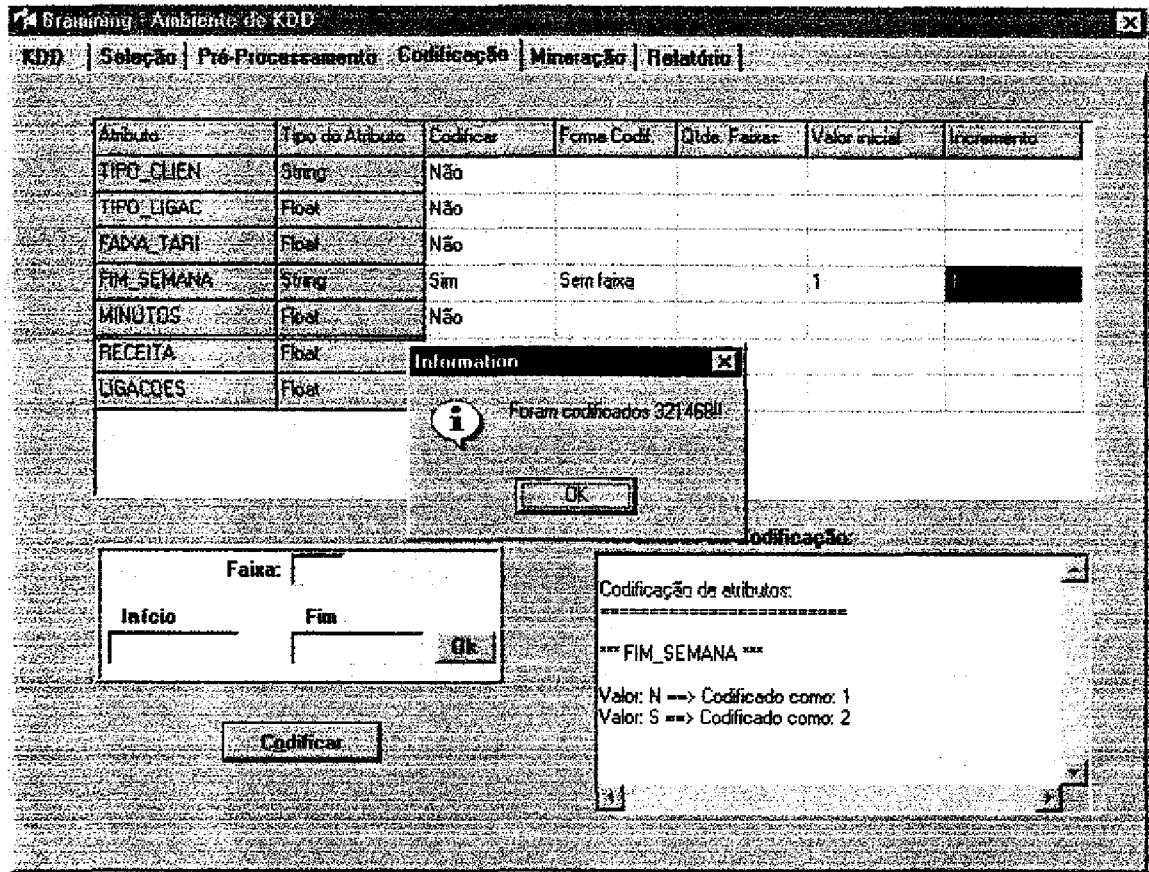


Figura 5.19 - Codificação do atributo Fim_de_semana.

A ferramenta desenvolvida disponibiliza a possibilidade de codificação de valores, em se tratando de atributos do tipo caracter (string) e em faixas de valores no caso de atributos do tipo numérico (float, integer, etc.).

A divisão em faixas não deve ser aleatória e sim representativa. No exemplo apresentado na Figura 5.18 foi feita uma normalização para o atributo Minutos a partir da distribuição de valores dentro da base de dados, ou seja, foi notado que existiam padrões dentro dos intervalos descritos nas faixas: 0 a 2 minutos, 2 a 10 minutos, 10 a 20 minutos, 20 a 50 minutos e > 50 minutos. Já o atributo Fim_de_semana foi codificado de forma mais simples, tendo 1 atribuído ao valor N e 2 ao valor S, conforme a Figura 5.19.

A codificação tanto pode ser empreendida pelo usuário (como na Figura 5.18 e na Figura 5.19) quanto através do uso de programas de normalização de dados utilizando cálculos com matrizes ou outros.

A Figura 5.20 mostra as estatísticas da base de dados após a execução das tarefas de Preparação.

Atributo	Tipo	Reg. em Branco	Menor Valor	Maior Valor	Valores Distintos
TIPO_CLIEN	String	0			C, R
TIPO_LIGAC	Float	0			1, 3, 4
FAIXA_TARI	Float	0			1, 2, 3, 4, 5
FIM_SEMANA	String	0			1, 2
MINUTOS	Float	0			1, 2, 3, 4, 5
RECEITA	Float	0	0,01	25387,31	
LIGACOES	Float	0	1	48485	

Figura 5.20 - Estatísticas da base de dados após a Preparação.

5.3.4 Mineração de dados

É a fase mais importante do processo de KDD. Nesta fase (se as anteriores tiverem sido corretamente executadas) serão descobertos os padrões escondidos na base de dados.

Para tanto, uma série de métodos podem ser utilizados:

- Estatística;
- Redes neurais;
- Algoritmos genéticos;
- Regras de produção e Árvores de decisão.
- Rough Sets

Além de Rough Sets, o outro método implementado nesta ferramenta é baseado no algoritmo apresentado pelo programa C4.5, desenvolvido por J. Ross Quinlan em [25]. Este algoritmo é um aprofundamento daquele existente no software ID3 e tem como objetivos, apresentar o resultado da mineração de dados na forma de árvores de decisão e regras de produção.

Além destes métodos de mineração de dados, a ferramenta tem ligação com outros softwares (WizRule, WizWhy, PolyAnalyst, XpertRule Miner, Predict, DBMiner e BusinessMiner). Isto permite que o analista de dados rode estes programas sem sair do ambiente de mineração de dados.

A tela inicial de mineração, Figura 5.21, apresenta um menu com opções de escolha para os softwares citados acima ou para os métodos implementados.

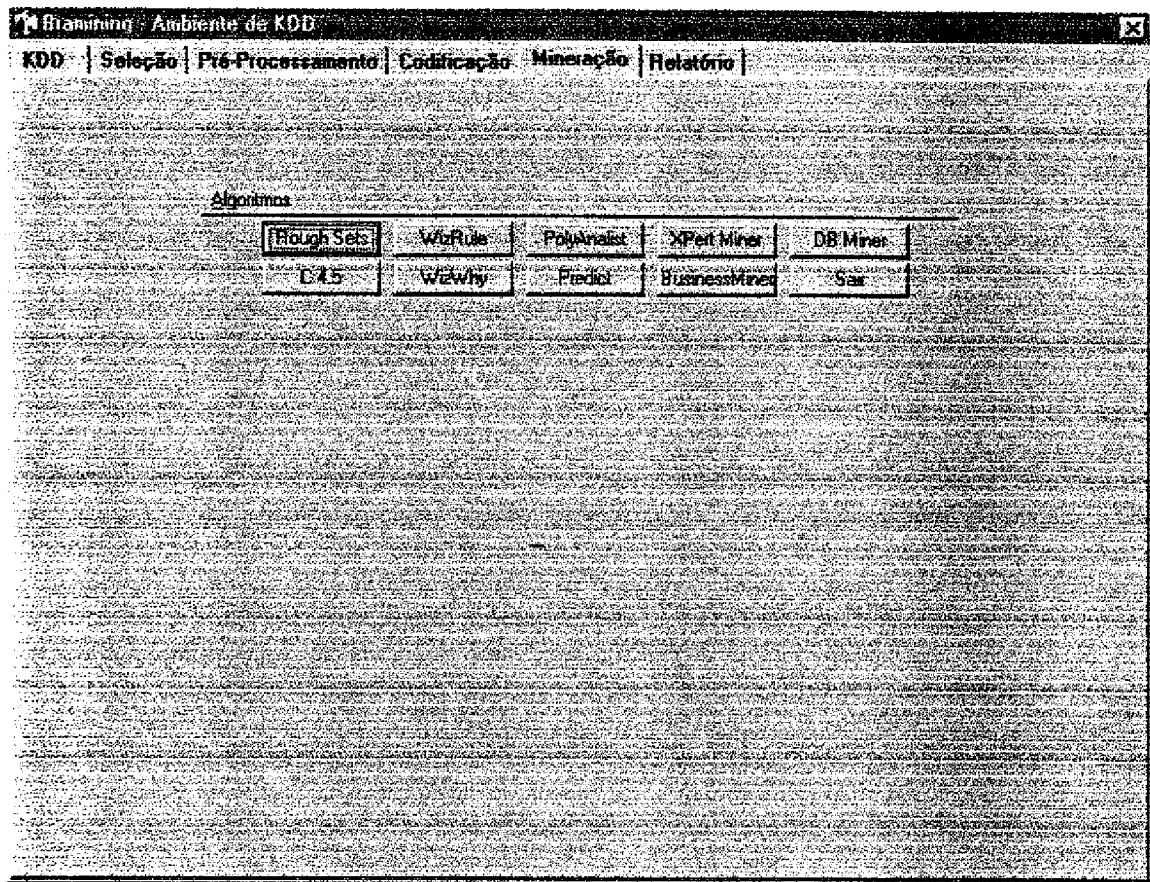


Figura 5.21 - Menu de escolha dos softwares de mineração de dados.

Após a escolha do método, a Figura 5.22 é apresentada para definição de parâmetros como escolha do atributo de saída e nome do projeto (no exemplo, foi escolhido Rough Sets).

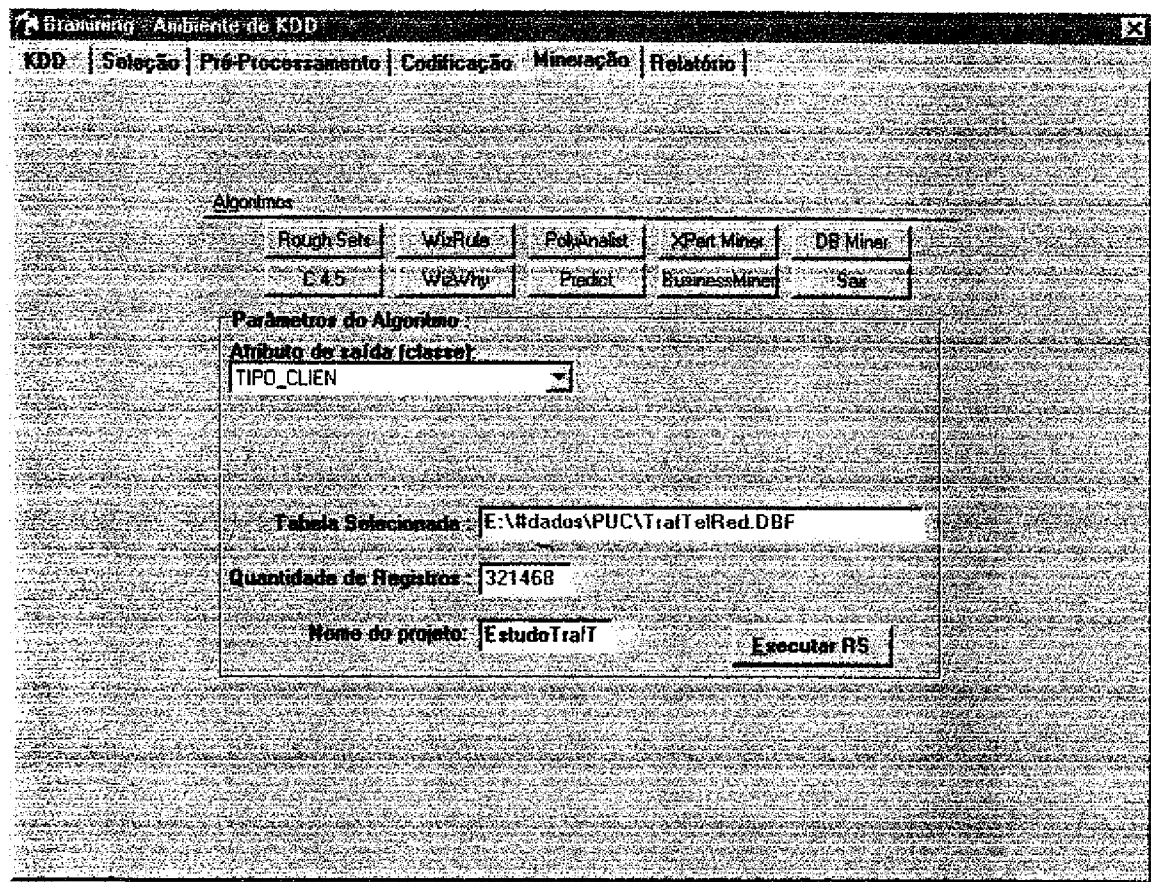


Figura 5.22 - Tela para definição de parâmetros da fase de mineração de dados.

Após o preenchimento destes dados, clicando no botão Executar é apresentada a tela da Figura 5.23, onde devem ser definidos alguns parâmetros complementares específicos para o algoritmo de Rough Sets.

À direita podem ser definidos percentuais mínimos de eficácia e abrangência para as regras a serem geradas. Estes coeficientes são descritos no Capítulo 3.

À esquerda da tela há dois campos que são usados para balizar a variedade de combinações de atributos a serem exploradas. Quanto maior a diferença entre os dois valores, maior a amplitude de combinações mas, conseqüentemente, a performance é proporcionalmente onerada.

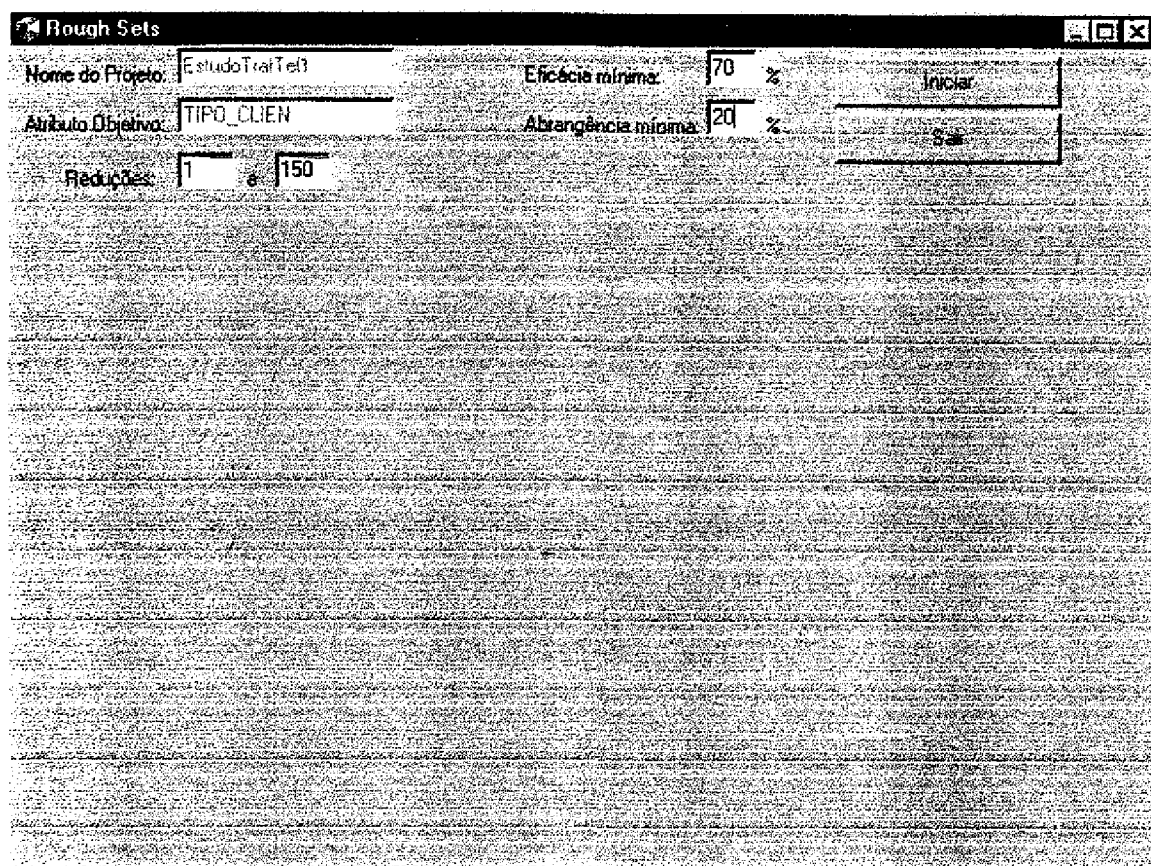


Figura 5.23 - Tela para definição de parâmetros complementares de Rough Sets.

Após clicar no botão Iniciar, são apresentadas algumas tabelas para acompanhamento da execução do algoritmo, conforme mostrado na Figura 5.24. A tabela de Classes contém os valores do atributo objetivo e a quantidade de registros para cada um; a tabela Dados de Origem exibe a base de dados sobre a qual foi feito o trabalho de mineração; a tabela Regras mostra as regras encontradas pelo algoritmo até o momento, juntamente com seus coeficientes. Abaixo e à direita, existe uma indicação do percentual concluído da execução.

The screenshot shows the 'Rough Sets' application window. At the top, there are input fields for 'Nome do Projeto' (EstudoTraTel), 'Eficácia mínima' (70%), and 'Abrangência mínima' (20%). Below these are 'Reduções' (1 a 150) and buttons for 'Iniciar' and 'Sair'. The main area is divided into three sections: 'Classes', 'Dados de origem', and 'Regras'. The 'Classes' table has columns 'DEL' and 'CARD'. The 'Dados de origem' table has columns 'TIPO_CLIEN', 'TIPO_LIGA', 'FAIXA_TARI', 'FIM_SEMANA', 'MINUTOS', and 'RECEITA'. The 'Regras' section lists several rules based on 'TIPO_LIGA' values.

DEL	CARD	TIPO_CLIEN	TIPO_LIGA	FAIXA_TARI	FIM_SEMANA	MINUTOS	RECEITA
C		C	3	5:1		2	0
R		C	3	5:1		3	4
		C	3	1:1		2	0
		C	4	1:1		2	0
		C	3	1:1		2	0

Regras

- SE TIPO_LIGA_ = 3 ENTÃO TIPO_CLIEN = C
- SE TIPO_LIGA_ = 4 ENTÃO TIPO_CLIEN = C
- SE TIPO_LIGA_ = 12 ENTÃO TIPO_CLIEN = C
- SE TIPO_LIGA_ = 3 ENTÃO TIPO_CLIEN = R
- SE TIPO_LIGA_ = 4 ENTÃO TIPO_CLIEN = R
- SE TIPO_LIGA_ = 12 ENTÃO TIPO_CLIEN = R

Concluído: 100%

Figura 5.24 - Tabelas para acompanhamento da execução do algoritmo.

Ao final da execução o usuário deve clicar em Sair para verificar os resultados, conforme descrito a seguir.

5.3.5 Relatório

Terminada a fase de mineração de dados, a janela Relatório ficará ativada para que o analista de dados possa visualizar o resultado obtido. Este relatório é mostrado na forma de regras de produção e seu objetivo é permitir que a interpretação do conhecimento gerado seja

feito de maneira bastante clara. Vejamos, na Figura 5.25, o resultado da mineração de dados tendo como atributo de saída, Tipo_cliente.

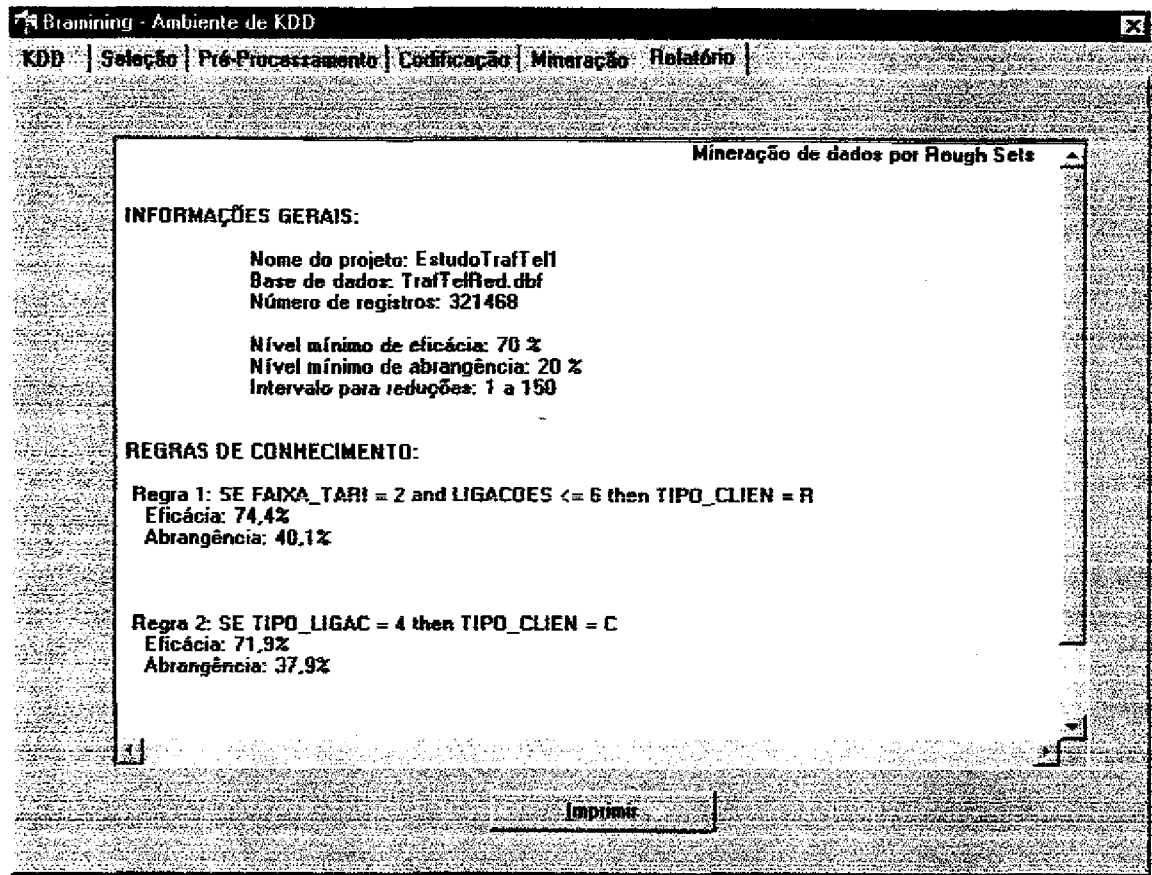


Figura 5.25 - Resultados obtidos pela mineração de dados, na forma de relatório.

Cada regra é apresentada seguida de sua confiabilidade, o que permite ao analista de dados selecionar aquelas mais representativas, de acordo com o objetivo proposto no início do processamento.

Para melhor ilustrar e avaliar a utilização do BRAMINING, no próximo capítulo serão feitos alguns estudos de caso de mineração de dados.

6 ESTUDOS DE CASO

6.1 INTRODUÇÃO

O principal objetivo deste capítulo é submeter a ferramenta desenvolvida a testes em condições reais de uso de um software de KDD. Através destes testes e da comparação dos resultados obtidos com um software de mercado, será possível avaliar melhor o desempenho e utilidade da ferramenta no processo de descoberta de conhecimento e, conseqüentemente, também dos algoritmos nela implementados.

Três bases de dados foram selecionadas para o estudo: dados de vestibulandos da PUC, para os quais se pretende encontrar um padrão para os candidatos que se classificam e não deixam de se matricular; uma base com dados de animais, na qual se deseja encontrar regras para classificá-los corretamente nas categorias previamente definidas e, finalmente, dados de tráfego telefônico de uma empresa de telecomunicações em que se deseja obter regras para classificar os clientes em duas categorias de uso do telefone: comercial ou residencial.

Cada uma das bases foram submetidas, além do Bramining, ao software WizRule, da WizSoft. Ele é um dos mais conhecidos comercialmente em mineração de dados e busca todas as regras que descrevem a base, dentre elas, regras se-então e regras matemáticas, bem como calcula o nível de certeza de cada regra. Ele é apresentado em detalhes no Capítulo 4 e no APÊNDICE A - WIZRULE PASSO A PASSO.

6.2 VESTIBULAR PUC

Neste estudo serão utilizados registros com dados de vestibulandos da PUC, descritos na Tabela 6.1 e exemplificados na Tabela 6.2.

CAMPO	DESCRIÇÃO/VALORES
SEXO	Sexo: FEMININO, MASCULINO
NATURAL	Naturalidade: RIO, OUTRA
IDADE	Idade: ATE16ANOS, ENTRE17E18, MAISDE19AN
BAIRRO	Região onde reside: ZONANORTE, ZONAOESTE, ZONASUL, CENTRO, OUTRO
CIDADE	Cidade onde reside: RIO, ESTADORIO, OUTRAS
OPCAO1	Área em que fez a 1ª opção no vestibular: ADMINETPD - Administração ou tecnologia de processamento de dados, EXATAS - Exatas, HUMANAS - Humanas
NUMPONTOS	Número de pontos obtidos no vestibular: ATE7000, ENTRE7E8MIL, MAISDE8000
EST_CIVIL	Estado civil: CASADO, SOLTEIRO, OUTRO
RELIGIAO	Religião: CATOLPRATI - Católico praticante, CATOLNAOPR - Católico não praticante, ATEU, CRISNAOCAT - Evangélico, JUDEU, OUTRACRIST - Outras religiões cristãs, S/RELIC/FE - Sem religião
TIPO_ESC	Tipo de escola de origem: MPPARTICUL - Meio período particular, MPPUBLICA - Meio período pública, TPARTICULA - Período integral particular, TPUBLICA - Período integral pública
TURNNO	Turno para o qual se classificou: MPDIURNO - Meio período diurno, MPNOTURNO - Meio período noturno, TDIURNO - Período integral diurno, TNOTURNO - Período integral noturno

JAFECURSUP	Se já fez curso superior: SIMABANDON - Sim mas abandonou, SIMCONCLUI - Sim e concluiu, SIMCURSAND - Está cursando, CONCL1ABAN - Concluiu um e abandonou outro, CONCL1CURS - Concluiu um e está cursando outro
INSCRVESTI	Inscreeu-se em outro vestibular: SONAPUC - Inscreeu-se somente na PUC, OUTOUT1OPC - Sim e em área diferente da 1ª opção da PUC, OUTMES1OPC - Sim e na mesma área da 1ª opção na PUC
RENDA	Renda familiar: MENOSDE1SAL - Menos de 1 salário mínimo, DE1A2SAL - de 1 a 2 salários mínimos, DE2A3SAL, DE3A4SAL, DE4A5SAL, DE5A7SAL, DE7A10SAL, DE10A20SAL, DE20A30SAL, DE30A50SAL, MAISDE50SAL
TRABALHA	O candidato trabalha: SIMINTEGRA - Sim em período integral, SIMPARCILA - Em período parcial, SIMEVENTUA - Eventualmente, NÃO
RESPOSTA	Se o candidato matriculou-se: CLASMATRIC - Classificado e não matriculado, CLASNAOMAT - Classificado e não matriculado
POSICAO	Colocação no vestibular

Tabela 6.1 - Descrição da base Vestibular da PUC

SE	NA	IDA	BAIR	CIDA	OPC	NUM	EST_	RELI	TIPO	TUR	JAFE	INSC	REN	TRA	RESP	PO
XO	TU	DE	RO	DE	AO1	PON	CIVI	GIAO	_ESC	NO	CUR	RVES	DA	BAL	OST	SIC
	RAL					TOS	L				SUP	TI		HA	A	AO
FEMI	RIO	ATE	ZON	RIO	HUM	ENTR	SOLT	JUDE	TPAR	TDIU	NAO	OUT	DE20	NAO	CLAS	81
NIN		16A	ASUL		ANA	E7E8	EIRO	U	TICU	RNO		MES1	A30S		NAO	
O		NOS			S	MIL			LA			OPC	AL		MAT	
FEMI	RIO	ATE	ZON	RIO	EXAT	ENTR	SOLT	CAT	TPAR	TDIU	NAO	SON	MAIS	NAO	CLAS	
NIN		16A	ASUL		AS	E7E8	EIRO	OLN	TICU	RNO		APUC	DE50		MAT	
O		NOS				MIL		AOPR	LA				SAL		RIC	
FEMI	RIO	ATE	ZON	RIO	EXAT	ATE7	SOLT	CAT	TPAR	TDIU	NAO	SON	DE30	NAO	CLAS	186

NIN	16A	ASUL	AS	000	EIRO	OLPR	TICU	RNO		APUC	A50S		MAT			
O	NOS					ATI	LA				AL		RIC			
MAS	RIO	ATE	ZON	RIO	ADMI	MAIS	SOLT	ATEU	TPAR	TDIU	NAO	OUT	DE30	NAO	CLAS	23
CULI	16A	ASUL			NETP	DE80	EIRO		TICU	RNO		MES1	A50S		MAT	
NO	NOS				D	00			LA			OPC	AL		RIC	
MAS	RIO	ATE	ZON	RIO	EXAT	ATE7	SOLT	ATEU	TPAR	TDIU	NAO	OUT	DE10	NAO	CLAS	23
CULI	16A	ASUL			AS	000	EIRO		TICU	RNO		MES1	A20S		NAO	
NO	NOS								LA			OPC	AL		MAT	
FEMI	RIO	ATE	ZON	RIO	EXAT	ATE7	SOLT	JUDE	TPAR	TDIU	NAO	SON	DE20	NAO	CLAS	976
NIN	16A	ASUL			AS	000	EIRO	U	TICU	RNO		APUC	A30S		MAT	
O	NOS								LA				AL		RIC	
FEMI	RIO	ATE	ZON	RIO	EXAT	ENTR	SOLT	CAT	TPAR	TDIU	NAO	OUT	MAIS	NAO	CLAS	13
NIN	16A	ANO			AS	E7E8	EIRO	OLN	TICU	RNO		MES1	DE50		NAO	
O	NOS	RTE				MIL		AOPR	LA			OPC	SAL		MAT	
FEMI	RIO	ATE	ZON	RIO	EXAT	ENTR	SOLT	JUDE	TPAR	TDIU	NAO	SON	DE20	NAO	CLAS	157
NIN	16A	ASUL			AS	E7E8	EIRO	U	TICU	RNO		APUC	A30S		MAT	
O	NOS					MIL			LA				AL		RIC	
MAS	RIO	ATE	ZON	RIO	EXAT	ENTR	SOLT	JUDE	TPAR	TDIU	NAO	OUT	DE4A	NAO	CLAS	
CULI	16A	ASUL			AS	E7E8	EIRO	U	TICU	RNO		MES1	5SAL		NAO	
NO	NOS					MIL			LA			OPC			MAT	
FEMI	RIO	ATE	ZON	RIO	EXAT	ENTR	SOLT	CAT	TPAR	TDIU	NAO	OUT	MAIS	NAO	CLAS	88
NIN	16A	ASUL			AS	E7E8	EIRO	OLN	TICU	RNO		MES1	DE50		MAT	
O	NOS					MIL		AOPR	LA			OPC	SAL		RIC	

Tabela 6.2 - Exemplos da base de vestibulandos da PUC

O objetivo proposto é descobrir regras de classificação que descrevam o perfil do vestibulando que realmente vem a se matricular na PUC (RESPOSTA = CLASMATRIC).

Os dados, que totalizam 2520 registros, foram submetidos ao BRAMINING e ao WizRule.

6.2.1 FORMA DE APRESENTAÇÃO DOS RESULTADOS

Em todos os estudos de caso a serem exibidos, os resultados de cada software são apresentados em duas tabelas: uma de regras e uma de coeficientes das regras, conforme será explanado nos itens a seguir.

6.2.1.1 Tabelas de Regras

As regras obtidas em cada teste são apresentadas com seu texto original retirado de cada software, conforme os exemplos das Tabela 6.3 e Tabela 6.4.

Nº REGRA	TEXTO DA REGRA
1	<i>If</i> SEXO <i>is</i> <u>FEMININO</u> and IDADE <i>is</i> <u>ATE16ANOS</u> and CIDADE <i>is</i> <u>RIO</u> <i>Then</i> RESPOSTA <i>is</i> <u>CLASNAOMAT</u>
2
...

Tabela 6.3 - Exemplo de regras geradas pelo WizRule

Nº REGRA	TEXTO DA REGRA
...
2	SE BAIRRO = ZONANORTE E OPCAO1 = HUMANAS

	E INSCRVESTI = OUTOUTIOPC ENTÃO RESPOSTA = CLASNAOMAT
...

Tabela 6.4 - Exemplo de regras geradas pelo Bramining

Uma regra gerada pelo WizRule tem a forma aproximada da regra 1 da Tabela 6.3. Ela deve ser interpretada da seguinte forma: "Se o candidato classificado é de sexo feminino, tem até 16 anos e reside na cidade do Rio de Janeiro, então ele não se matriculará."

Analogamente, uma regra gerada pelo Bramining tem a forma geral da regra 2 da Tabela 6.4. Ela deve ser assim interpretada: "Se o candidato classificado reside na Zona Norte, sua 1ª opção na PUC é na área de Humanas e se inscreveu em outro vestibular para área diferente da 1ª opção da PUC, então ele não se matriculará ".

Estas regras de conhecimento têm determinados coeficientes a ela associados que definem seu nível de qualidade, conforme será visto no próximo item.

6.2.1.2 Tabelas de coeficientes

REGRA	CLASSE	EFICÁCIA (%)	ABRANGÊNCIA (%)	REGISTROS NA CLASSE
1	NÃO MATRICULADO	75	4,1	879
2

Tabela 6.5 - Exemplo de tabela de coeficientes de regras

Para cada tabela de regras obtida é apresentada em seguida uma tabela com os coeficientes destas regras, conforme o exemplo da Tabela 6.5, referente às regras da Tabela 6.3. Nela são colocadas as informações de: Valor (classe) do atributo objetivo contido na

regra, seu grau de certeza ou confiança (eficácia), grau de suporte (abrangência) e o número de registros daquela classe.

Basicamente, a eficácia indica a probabilidade de um registro da base que se enquadra em todos os critérios do lado esquerdo da regra (o "SE ...") pertencer à classe indicada no lado direito da regra ("ENTÃO ..."). Já a abrangência vem indicar que percentual de registros da classe é coberto pela regra. Estes coeficientes são explicados em detalhes no Capítulo 3.

6.2.2 Resultados do WizRule

A Tabela 6.6 mostra o texto das principais regras geradas pelo WizRule para a base da PUC e a Tabela 6.7 exhibe os coeficientes obtidos por estas regras.

Nº REGRA	TEXTO DA REGRA
1	<i>If</i> SEXO <i>is</i> <u>FEMININO</u> and IDADE <i>is</i> <u>ATE16ANOS</u> and CIDADE <i>is</i> <u>RIO</u> <i>Then</i> RESPOSTA <i>is</i> <u>CLASNAOMAT</u>
2	<i>If</i> SEXO <i>is</i> <u>FEMININO</u> and BAIRRO <i>is</i> <u>ZONANORTE</u> and NUMPONTOS <i>is</i> <u>MAISDE8000</u> <i>Then</i> RESPOSTA <i>is</i> <u>CLASNAOMAT</u>
3	<i>If</i> BAIRRO <i>is</i> <u>ZONANORTE</u> and NUMPONTOS <i>is</i> <u>MAISDE8000</u> and RELIGIAO <i>is</i> <u>CATOLNAOPR</u>

	<p><i>Then</i></p> <p>RESPOSTA is <u>CLASNAOMAT</u></p>
4	<p><i>If</i> CIDADE is <u>ESTADORIO</u></p> <p>and OPCA01 is <u>EXATAS</u></p> <p>and NUMPONTOS is <u>ENTRE7E8MIL</u></p> <p><i>Then</i></p> <p>RESPOSTA is <u>CLASNAOMAT</u></p>
5	<p><i>If</i> OPCA01 is <u>HUMANAS</u></p> <p>and NUMPONTOS is <u>MAISDE8000</u></p> <p>and INSCRVESTI is <u>OUTOUT1OPC</u></p> <p><i>Then</i></p> <p>RESPOSTA is <u>CLASNAOMAT</u></p>
6	<p><i>If</i> NATURAL is <u>RIO</u></p> <p>and BAIRRO is <u>OUTRO</u></p> <p>and OPCA01 is <u>EXATAS</u></p> <p>and NUMPONTOS is <u>ENTRE7E8MIL</u></p> <p><i>Then</i></p> <p>RESPOSTA is <u>CLASNAOMAT</u></p>
7	<p><i>If</i> SEXO is <u>FEMININO</u></p> <p>and NATURAL is <u>RIO</u></p> <p>and IDADE is <u>ENTRE17E18</u></p> <p>and NUMPONTOS is <u>MAISDE8000</u></p> <p>and INSCRVESTI is <u>OUTOUT1OPC</u></p> <p><i>Then</i></p> <p>RESPOSTA is <u>CLASNAOMAT</u></p>
8	<p><i>If</i> NATURAL is <u>RIO</u></p> <p>and IDADE is <u>ENTRE17E18</u></p> <p>and OPCA01 is <u>HUMANAS</u></p> <p>and NUMPONTOS is <u>MAISDE8000</u></p> <p>and RELIGIAO is <u>CATOLPRATI</u></p> <p><i>Then</i> RESPOSTA is <u>CLASNAOMAT</u></p>

Tabela 6.6 - Regras do WizRule. Base: Vestibular da PUC

REGRA	CLASSE	EFICÁCIA (%)	ABRANGÊNCIA (%)	REGISTROS NA CLASSE
1	NÃO MATRICULADO	75	4,1	879
2	NÃO MATRICULADO	78	4,4	879
3	NÃO MATRICULADO	75	3,4	879
4	NÃO MATRICULADO	76,2	3,6	879
5	NÃO MATRICULADO	75	3,8	879
6	NÃO MATRICULADO	75	3,8	879
7	NÃO MATRICULADO	75,6	3,8	879
8	NÃO MATRICULADO	75	3,4	879

Tabela 6.7 - Coeficientes das regras do WizRule. Base: Vestibular da PUC

6.2.3 Resultados do Bramining

A Tabela 6.8 mostra o texto das principais regras geradas pelo Bramining para a base da PUC e a Tabela 6.9 exhibe os coeficientes obtidos por estas regras.

Nº REGRA	TEXTO DA REGRA
1	SE OPCAO1 = IIUMANAS

	E RELIGIAO = JUDEU E INSCRVESTI = OUTOUTIOPC ENTÃO RESPOSTA = CLASNAOMAT
2	SE BAIRRO = ZONANORTE E OPCA01 = HUMANAS E INSCRVESTI = OUTOUTIOPC ENTÃO RESPOSTA = CLASNAOMAT
3	SE NATURAL = OUTRA E OPCA01 = HUMANAS E NUMPONTOS = ENTRE7E8MIL ENTÃO RESPOSTA = CLASMATRIC
4	SE NUMPONTOS = ATE7000 ENTÃO RESPOSTA = CLASMATRIC
5	SE BAIRRO = ZONAOESTE ENTÃO RESPOSTA = CLASMATRIC
6	SE OPCA01 = ADMINETPD ENTÃO RESPOSTA = CLASMATRIC

Tabela 6.8 - Regras do Bramining. Base: Vestibular da PUC

REGRA	CLASSE	EFICÁCIA (%)	ABRANGÊNCIA (%)	REGISTROS NA CLASSE
1	NÃO MATRICULADO	79,4	1,1	879
2	NÃO MATRICULADO	64,5	3	879
3	MATRICULADO	84,3	3,3	1641
4	MATRICULADO	80,7	44	1641
5	MATRICULADO	78,3	18,4	1641
6	MATRICULADO	67,7	24,3	1641

Tabela 6.9 - Coeficientes das regras do Bramining. Base: Vestibular da PUC

6.3 DADOS DE ANIMAIS

A base de dados escolhida para esta avaliação foi retirada do site de acesso público em: “ftp://ftp.ics.uci.edu/pub/machine-learning-databases”. O banco de dados chama-se “zoo”, e contém 18 atributos de 101 diferentes animais, classificados em sete categorias, conforme a descrição da Tabela 6.10 e a amostra na Tabela 6.11.

CAMPO	DESCRIÇÃO/VALORES
name	Nome do animal
hair	Pelos: 1 - Sim, 0 - Não
feathers	Penas: 1 - Sim, 0 - Não
eggs	Ovos: 1 - Sim, 0 - Não
milk	Leite: 1 - Sim, 0 - Não
airborne	Voa: 1 - Sim, 0 - Não
Aquatic	Aquático: 1 - Sim, 0 - Não
predator	Predador: 1 - Sim, 0 - Não
toothed	Dentes: 1 - Sim, 0 - Não
backbone	Espinha dorsal: 1 - Sim, 0 - Não
breathes	Respira ar: 1 - Sim, 0 - Não
venomous	Venenoso: 1 - Sim, 0 - Não
fins	Barbatanas: 1 - Sim, 0 - Não
legs	Nº de pernas
tail	Cauda: 1 - Sim, 0 - Não
domestic	Doméstico: 1 - Sim, 0 - Não
catsize	Maior que um gato : 1 - Sim, 0 - Não
type	Classe (não é a divisão tradicional científica): 1 a 7

Tabela 6.10 - Descrição da base de dados de animais

name	hair	feath ere	eggs	milk	airb orne	aqua tic	pred ator	Toot hed	back bone	brea thes	veno mous	fins	legs	tail	dom estic	catsi ze	type
gorilla	1	0	0	1	0	0	0	1	1	1	0	0	2	0	0	1	1
crow	0	1	1	0	1	0	1	0	1	1	0	0	2	1	0	0	2
seasn ake	0	0	0	0	0	1	1	1	1	0	1	0	0	1	0	0	3
catfish	0	0	1	0	0	1	1	1	1	0	0	1	0	1	0	0	4
frog	0	0	1	0	0	1	1	1	1	1	0	0	4	0	0	0	5
honey bee	1	0	1	0	1	0	0	0	0	1	1	0	6	0	1	0	6
crab	0	0	1	0	0	1	1	0	0	0	0	0	4	0	0	0	7

Tabela 6.11 - Amostra da base de dados de animais

6.3.1 Resultados do WizRule

A Tabela 6.12 mostra o texto das principais regras geradas pelo WizRule para a base "Zoo" e a Tabela 6.13 exhibe os coeficientes obtidos por estas regras. A regra 1, por exemplo, pode ser interpretada da seguinte forma: "Se o animal dá leite, então sua classe é 1". Regras como a Nº 4, do tipo "SE E SOMENTE SE" são geradas eventualmente pelo WizRule. Elas valem nos dois sentidos, ou seja, são equivalentes a duas regras "SE ENTÃO". A regra 4, por exemplo, equivale às duas regras a seguir: "Se o animal põe ovos e tem penas, então é da classe 4"; "Se o animal é da classe 4, então põe ovos e tem penas".

Nº REGRA	TEXTO DA REGRA
1	<i>If milk is <u>1</u> Then type is <u>1</u></i>
2	<i>If feathers is <u>1</u> Then type is <u>2</u></i>
3	<i>If toothed is <u>1</u> and fins is <u>0</u> and legs is <u>0</u> Then type is <u>3</u></i>
4	<i>eggs is <u>1</u> and fins is <u>1</u> if and only if type is <u>4</u></i>

5	<i>If hair is 0 and toothed is 1 and legs is 4 Then type is 5</i>
6	<i>breathes is 1 and legs is 6 if and only if type is 6</i>
7	<i>If airborne is 0 and backbone is 0 Then type is 7</i>

Tabela 6.12 - Regras do WizRule. Base: Animais

REGRA	CLASSE	EFICÁCIA (%)	ABRANGÊNCIA (%)	REGISTROS NA CLASSE
1	1	100	100	41
2	2	100	100	20
3	3	100	60	5
4	4	100	100	13
5	5	80	100	4
6	6	100	100	8
7	7	83,3	100	10

Tabela 6.13 - Coeficientes das regras do WizRule. Base: Animais

6.3.2 Resultados do Bramining

Após a execução do Bramining para a base "Zoo", foi examinada a tabela de regras gerada e avaliados seus coeficientes. A Tabela 6.14 mostra o texto das regras indicadas e a Tabela 6.15 resume os coeficientes obtidos por estas regras. Uma regra como a N° 2 pode ser interpretada da seguinte forma: "Se o animal não dá leite e tem duas pernas, então é da classe 2".

N° REGRA	TEXTO DA REGRA
1	SE milk = 1 ENTÃO type = 1

2	SE milk = 0 E legs = 2 ENTÃO type = 2
3	SE eggs = 1 E aquatic = 0 E toothed = 1 ENTÃO type = 3
4	SE eggs = 1 E aquatic = 0 E backbone = 1 ENTÃO type = 3
5	SE breathes = 0 E fins = 1 ENTÃO type = 4
6	SE aquatic = 1 E toothed = 1 E legs = 4 ENTÃO type = 5
7	SE aquatic = 0 E toothed = 0 E legs = 6 ENTÃO type = 6
8	SE breathes = 0 E venomous = 0 E tail = 0 ENTÃO type = 7
9	SE hair = 0 E toothed = 0 E backbone = 0 ENTÃO type = 7

Tabela 6.14 - Regras do Bramining. Base: Animais

REGRA	CLASSE	EFICÁCIA (%)	ABRANGÊNCIA (%)	REGISTROS NA CLASSE
1	1	100	100	21
2	2	100	100	10
3	3	100	60	3
4	3	22,2	80	3
5	4	100	100	7
6	5	80	100	2
7	6	100	100	4
8	7	100	60	5
9	7	71,4	100%	5

Tabela 6.15 - Coeficientes das regras do Bramining. Base: Animais

6.4 USUÁRIOS DE COMPANHIA TELEFÔNICA

Neste estudo serão utilizados registros com dados de tráfego telefônico, conforme o formato da Tabela 6.16. O objetivo deste estudo é identificar o perfil residencial ou comercial no uso do telefone do cliente.

CAMPO	DESCRIÇÃO/VALORES
ASSINANTE	Assinante de origem, aquele do qual partiu a ligação
TIPO_LIGAC	1 – DDD, 2 – Local, 3 – A cobrar, 4 – 0800, 5 – Celular
FAIXA_TARI	Faixa de tarifação: 1 – Normal, 2 – Reduzido, 3 – Super Reduzido, 4 – Misto, 5 – Diferenciado
DISTANCIA	Distância física entre a central telefônica do assinante de origem e de destino
FIM_SEMANA	S Sábado e Domingo, N Segunda a Sexta-feira
MINUTOS	Duração da ligação, em minutos
RECEITA	Valor gasto com a ligação
LIGAÇÕES	Quantidade de ligações feitas no mês
TIPO_CLIEN	Tipo de cliente: R – Residencial, C – Comercial

Tabela 6.16 - Descrição da base de dados de tráfego telefônico

A base de dados, com cerca de 325.000 registros, foram submetidos ao BRAMINING e ao WizRule. A Tabela 6.17 exibe uma amostra dos registros desta base.

ASSINAN TE	TIPO_ LIGAC	FAIXA_ TARI	DISTAN CIA	FIM_ SEMANA	MINUTOS	RECEITA	LIGA COES	TIPO_ CLIEN
119	1	5	3	N	378	107,78	129	C
158	1	5	4	N	367	141,7	142	C
395	1	5	4	N	363	142,13	114	C
477	1	5	3	N	984	329,81	830	C
898	1	5	4	N	612	235,37	292	C
906	1	5	2	N	619	118,58	320	C
1189	3	1	4	N	22	4,12	8	R
1192	1	5	1	N	37	4,36	16	R

1240	1	2	4	N	50	4,5	2	R
1273	3	2	4	S	55	4,93	3	R
1285	1	5	4	N	11	4,25	6	R

Tabela 6.17 - Amostra da base de dados de tráfego telefônico

6.4.1 Resultados do WizRule

Encontram-se a seguir, as regras geradas pelo WizRule consideradas relevantes para o estudo de caso proposto. A Tabela 6.18 mostra o texto das principais regras geradas e a Tabela 6.19 exhibe os coeficientes obtidos por estas regras. Uma regra como a de Nº 1 é interpretada da seguinte forma: "Se o tipo de ligação é 0800 e o número de ligações é entre 4 e 12, então o cliente é do tipo Comercial".

Nº REGRA	TEXTO DA REGRA
1	<i>If TIPO_LIGAC is 04 and NUM_LIGACO is $8,00 \pm 4,00$ Then TIPO_CLIEN is <u>C</u></i>
2	<i>If TIPO_LIGAC is 04 and FAIXA_TARI is 5 Then TIPO_CLIEN is <u>C</u></i>
3	<i>If TIPO_LIGAC is 04 Then TIPO_CLIEN is <u>C</u></i>
4	<i>If TIPO_LIGAC is 04 and FAIXA_TARI is 1 and MINUTOS is $15,00 \pm 11,00$ Then TIPO_CLIEN is <u>C</u></i>
5	<i>If TIPO_LIGAC is 01 and FAIXA_TARI is 2 Then TIPO_CLIEN is not <u>C</u></i>
6	<i>If FAIXA_TARI is 2 and FIM_SEMANA is N and MINUTOS is $15,00 \pm 11,00$ and NUM_LIGACO is $2,00 \pm 1,00$ Then TIPO_CLIEN is not <u>C</u></i>
7	<i>If FAIXA_TARI is 2 Then TIPO_CLIEN is not <u>C</u></i>

8	<i>If TIPO_LIGAC is 01 and FIM_SEMANA is S Then TIPO_CLIEN is not C</i>
9	<i>If TIPO_LIGAC is 03 and FIM_SEMANA is S Then TIPO_CLIEN is not C</i>

Tabela 6.18 - Regras do WizRule. Base: Tráfego telefônico

REGRA	CLASSE	EFICÁCIA (%)	ABRANGÊNCIA (%)	REGISTROS NA CLASSE
1	C	86,3	10,6	100.000
2	C	82,5	18,5	100.000
3	C	80,2	40,1	100.000
4	C	82,5	8,3	100.000
5	R	80,2	17,6	225.000
6	R	77,4	3,7	225.000
7	R	73,6	25,3	225.000
8	R	73,1	12,8	225.000
9	R	70,9	5,8	225.000

Tabela 6.19 - Coeficientes das regras do WizRule. Base: Tráfego telefônico

6.4.2 Resultados do Bramining

A Tabela 6.20 mostra o texto das principais regras geradas pelo Bramining para a base de tráfego telefônico e a Tabela 6.21 exhibe os coeficientes obtidos por estas regras. Uma regra como a N° 1 é assim interpretada: "Se a faixa de tarifação é a Reduzida e o tempo de uso é maior que 42 minutos, então o cliente é do tipo Residencial".

N° REGRA	TEXTO DA REGRA
1	SE FAIXA_TARIFAÇÃO = 2 and MINUTOS > 42 then

	TIPO_CLIEN = R
2	SE FAIXA_TARIFAÇÃO = 2 and LIGAÇÕES <= 6 then TIPO_CLIEN = R
3	SE TIPO_LIGAÇÃO = 1 and FIM_SEMANA = S and MINUTOS > 1 and RECEITA <= 0.3 then TIPO_CLIEN = R
4	SE TIPO_LIGAÇÃO = 1 and MINUTOS > 4 and LIGAÇÕES <= 2 then TIPO_CLIEN = R
5	SE TIPO_LIGAÇÃO = 1 and FAIXA_TARIFAÇÃO = 3 then TIPO_CLIEN = R
6	SE TIPO_LIGAÇÃO = 3 and FAIXA_TARIFAÇÃO = 3 then TIPO_CLIEN = R
7	SE FAIXA_TARIFAÇÃO = 1 and LIGAÇÕES > 11 then TIPO_CLIEN = C
8	SE TIPO_LIGAÇÃO = 4 then TIPO_CLIEN = C
9	SE TIPO_LIGAÇÃO = 1 and FAIXA_TARIFAÇÃO = 1 and MINUTOS <= 4 then TIPO_CLIEN = C

Tabela 6.20 - Regras do Bramining. Base: Tráfego telefônico

REGRA	CLASSE	EFICÁCIA (%)	ABRANGÊNCIA (%)	REGISTROS NA CLASSE
1	R	91,4	32,3	225.000
2	R	74,4	40,1	225.000
3	R	74,1	14,8	225.000
4	R	70,8	23,6	225.000
5	R	61,5	21,7	225.000
6	R	61,2	16,4	225.000
7	C	75,4	28,2	100.000
8	C	71,9	37,9	100.000
9	C	56,6	23,7	100.000

Tabela 6.21 - Coeficientes das regras do Bramining. Base: Tráfego telefônico

6.5 CONSIDERAÇÕES SOBRE OS RESULTADOS

O conjunto de regras de conhecimento gerado por uma determinada ferramenta para descrever uma base de dados não pode ser facilmente comparado àquele gerado por uma outra, a não ser que se obtenha regras de eficácia e abrangência 100%, caso não muito comum de ocorrer. Para exemplificar o problema da comparação, considerem-se as regras obtidas pelo Bramining e WizRule para descrever a classe 7 da base de animais (item 6.3), trazidas para a Tabela 6.22:

IDENTIFI CADOR	SOFT WARE	REGRA	EFICÁ CIA	ABRAN GÊNCIA
I	WizRule	<i>If airborne is 0 and backbone is 0 Then type is 7</i>	83,3%	100%
II	Bramining	SE breathes = 0 E venomous = 0 E tail = 0 ENTÃO type = 7	100%	60%
III	Bramining	SE hair = 0 E toothed = 0 E backbone = 0 ENTÃO type = 7	71,4%	100%

Tabela 6.22 - Exemplo de comparação de qualidade de regras

A regra I tem bons índices de eficácia e abrangência. Pode-se facilmente dizer que ela tem mais qualidade que a regra III, pela comparação direta dos coeficientes.

E quanto à regra II? Ela seria considerada melhor ou pior que a regra I? Em termos de eficácia, a regra II é superior, mas perde em abrangência por uma diferença maior. O que tem maior peso: eficácia ou abrangência? E o resultado composto por II e III, seria superior ao

resultado isolado de I? Não há uma resposta direta para estas perguntas; cada caso de mineração deve ser avaliado individualmente, de acordo com a necessidade do solicitante.

Desta forma, não há maneira simples e direta de comparar os resultados obtidos pelo WizRule e pelo Bramining, mas pode-se observar na Tabela 6.23, onde foram destacados os coeficientes das "melhores" regras geradas para cada classe, que os índices apresentados pelo Bramining podem ser considerados, de forma geral, equivalentes aos do WizRule, ambos obtendo regras com boa eficácia e abrangência, cobrindo grande parte dos elementos das classes. A exceção ocorreu na base "Vestibular PUC", para a qual nenhum dos dois aplicativos conseguiu gerar regras com boa abrangência. O Bramining, entretanto, gerou regras para a todas as classes de dados desta base, o que não aconteceu com o WizRule.

BASE	CLASSE	WIZRULE		BRAMINING	
		EFICÁCIA (%)	ABRAN GÊNCIA (%)	EFICÁCIA (%)	ABRAN GÊNCIA (%)
Vestibular PUC	Não matriculado	78	4,4	64,5	3
Vestibular PUC	Matriculado	-	-	80,7	44
Animais	1	100	100	100	100
Animais	2	100	100	100	100
Animais	3	100	60	100	60
Animais	4	100	100	100	100
Animais	5	80	100	80	100
Animais	6	100	100	100	100

Animais	7	83,3	100	71,4	100%
Tráfego telefônico	C	80,2	40,1	91,4	32,3
Tráfego telefônico	R	73,6	25,3	71,9	37,9

Tabela 6.23 - Tabela comparativa de resultados

Um aspecto, entretanto, deve ser ressaltado: o WizRule, assim como os demais softwares avaliados no Capítulo 4, limita-se à fase de Mineração de Dados. O Bramining, por sua vez, consiste de um ambiente que abrange outras fases do processo de KDD, tendo permitido que as bases de dados utilizadas nos estudos fossem adequadamente preparadas para a tarefa de mineração, para o uso de ambos os softwares. Desta forma, os resultados obtidos pelo Bramining neste capítulo devem ser avaliados de forma mais abrangente.

No próximo capítulo serão apresentadas as conclusões acerca do trabalho, bem como sugestões para sua continuidade.

7 CONCLUSÕES E TRABALHOS FUTUROS

7.1 CONCLUSÕES

O principal objetivo do trabalho apresentado foi investigar o desempenho da técnica de Rough Sets em Mineração de Dados. Para isso foi feito um estudo sobre o processo de KDD, um estudo sobre as técnicas de Rough Sets aplicadas à mineração de dados, uma análise de ferramentas de mineração de dados do mercado, a implementação de novas características no projeto Bramining e, finalmente, a realização de alguns estudos de caso para avaliar o aplicativo.

Os índices apresentados pelo Bramining nos estudos de caso podem ser considerados, de forma geral, equivalentes aos do WizRule, tendo ambos obtido regras com boa eficácia e abrangência na maioria dos casos. Um aspecto, entretanto, deve ser ressaltado: o WizRule, assim como os demais softwares avaliados no Capítulo 4, limita-se à fase de Mineração de Dados. O Bramining, por sua vez, consiste de um ambiente que abrange outras fases do processo de KDD, tendo permitido que as bases de dados utilizadas nos estudos fossem adequadamente preparadas para a tarefa de mineração, para o uso de ambos os softwares. Desta forma, os resultados obtidos pelo Bramining devem ser avaliados de forma mais abrangente.

Os resultados obtidos comprovaram, através da aplicação desenvolvida, a adequação dos conceitos de Rough Sets à tarefa de classificação de dados. Alguns pontos frágeis da técnica foram identificados, como a necessidade de um mecanismo de apoio para a redução de atributos e a dificuldade em trabalhar com atributos de domínio contínuo. Porém, ao se inserir

a técnica em um ambiente mais completo de KDD, como o Bramining, estas deficiências foram sanadas. As opções de preparação da base que o Bramining disponibiliza ao usuário para executar, em particular, a redução e a codificação de atributos permitem deixar os dados em estado adequado à aplicação de Rough Sets.

A mineração de dados é uma questão bastante relevante nos dias atuais, e muitos métodos têm sido propostos para as diversas tarefas que dizem respeito a esta questão. A teoria de Rough Sets não mostrou significativas vantagens ou desvantagens em relação a outras técnicas já consagradas, mas foi de grande valia comprovar que há caminhos alternativos no processo de descoberta de conhecimento. Podem ser destacadas algumas das principais vantagens do seu uso neste contexto [23] :

- ◆ Oferece resultados facilmente interpretáveis
- ◆ Avalia a significância de dados
- ◆ Seus conceitos são de fácil compreensão

7.2 TRABALHOS FUTUROS

Questões surgidas no decorrer deste trabalho sugerem algumas direções para sua continuidade, como:

1. Criação de novas opções de algoritmos no ambiente Bramining;
2. Aprimoramento do módulo C4.5, para exibir a evolução da execução;
3. Refinamento do módulo de Rough Sets, para aprimoramento dos campos que balizam a variedade de combinações de atributos a serem exploradas;
4. Aprimoramento do pós-processamento no Bramining, possibilitando a customização de relatórios e incluindo opções de gráficos.

APÊNDICE A - WIZRULE PASSO A PASSO

A tela inicial do WizRule é apresentada abaixo e, nela é permitido escolher a fonte em que a sua base de dados se encontra. No caso de uma tabela em formato Access, Excel, DBase (.mdb, .dbf, .xls), é preciso escolher a opção Data File. E no caso de uma tabela Oracle, por exemplo, basta escolher SQL Source.

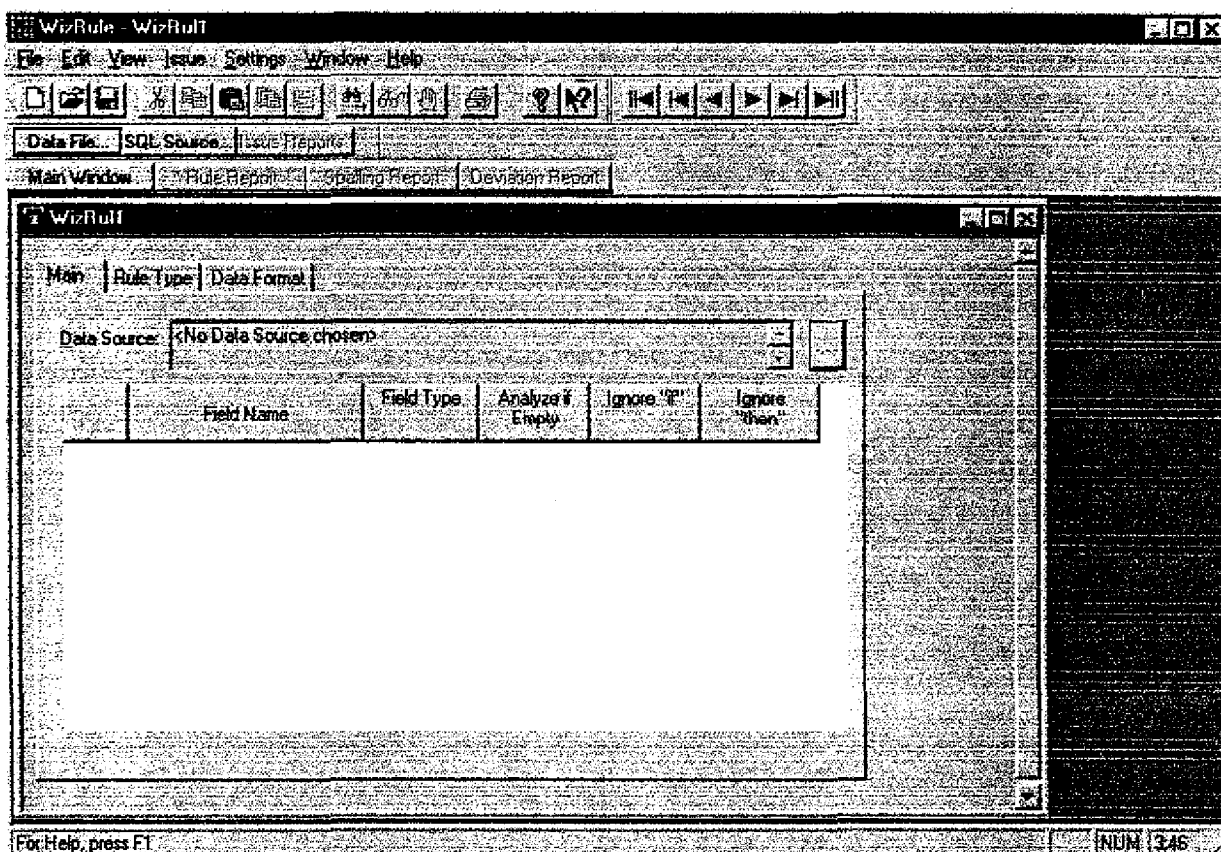


FIGURA A1: Tela inicial do WizRule.

Escolhida a base de dados a ser trabalhada, o software necessita que alguns parâmetros sejam ajustados e que os campos relevantes ao processamento, bem como o campo de saída sejam definidos na janela Main. No caso em questão, o campo Tipo_cliente foi estabelecido

como saída para cada regra gerada e alguns campos foram descartados por não serem de grande relevância para o resultado final.

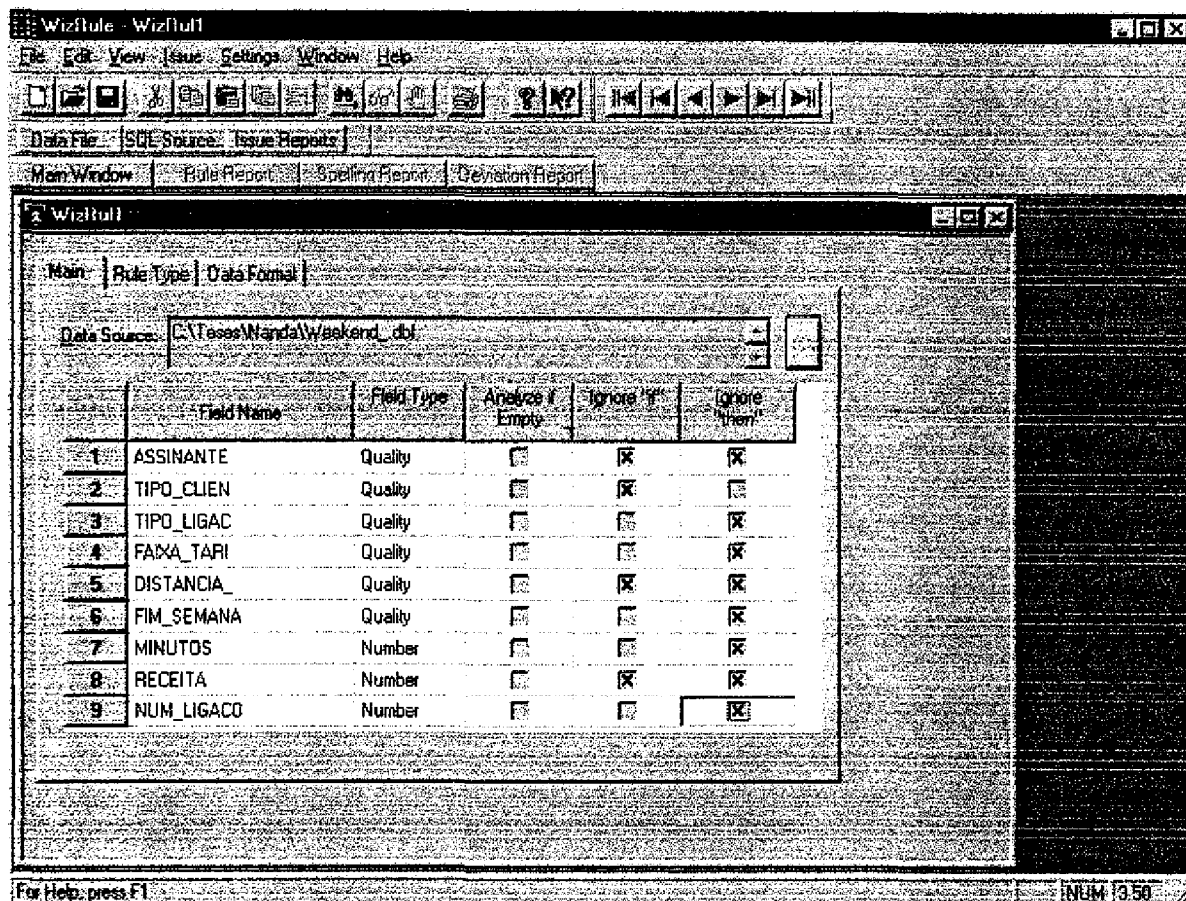


FIGURA A2: Especificação dos atributos relevantes e do atributo de saída.

Na janela Rule Type é possível especificar o número mínimo de casos para que uma regra seja gerada, assim como a quantidade de casos de desvio que deve ser mostrada.

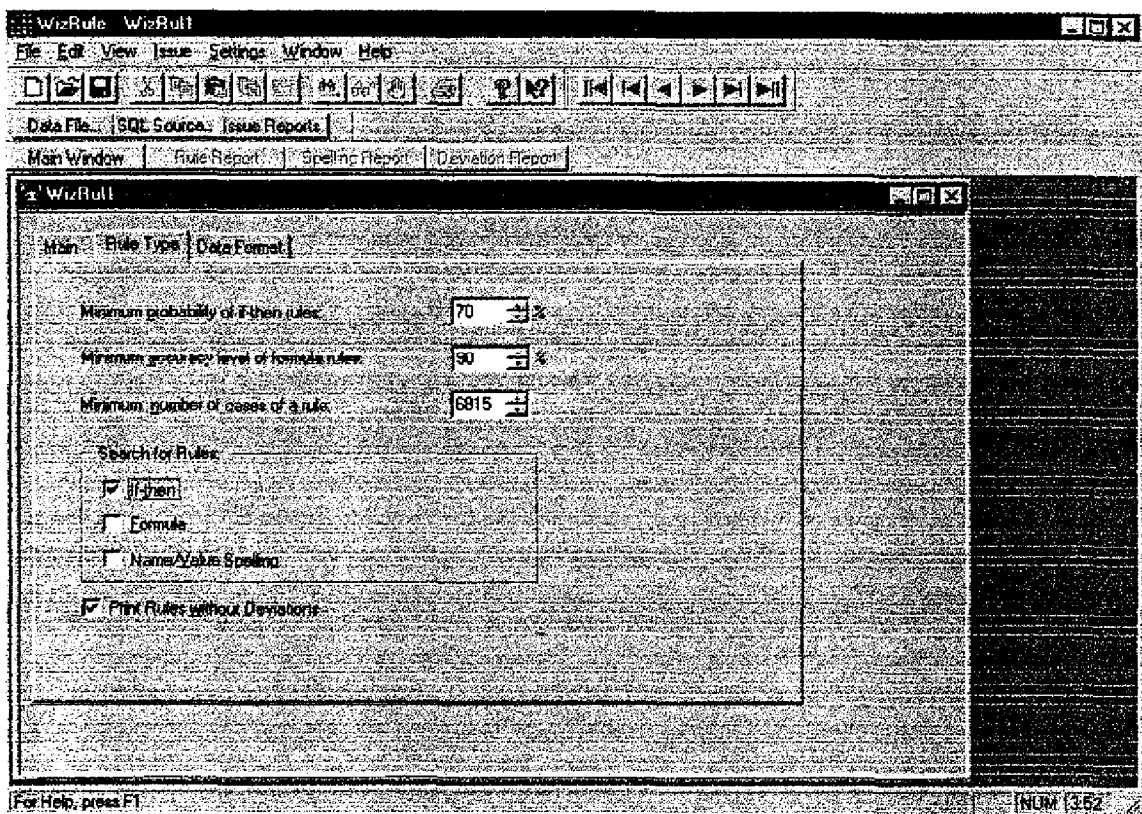


FIGURA A3: Definição de parâmetros para as regras se-então.

Feito isto, basta clicar em Issue Reports para que a geração de regras seja iniciada. O processamento é iniciado e o resultado é mostrado em forma de relatório.

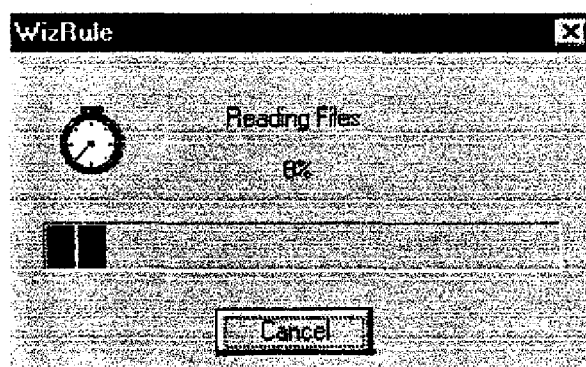


FIGURA A4: Barra indicando o andamento da geração das regras.

The screenshot shows the WizRule application window with the following content:

321996, 321997, 321998, 321999, 322000, 322001

2) **FIM_SEMANA is N**
Rule's probability: 0,798
The rule exists in 260106 records.

IF-THEN RULES:

3) **if FAIXA_TARI is 2**
Then
TIPO_CLIEN is R
Rule's probability: 0,736
The rule exists in 56749 records.
Significance Level: Error probability < 0,1
Deviations (records' serial numbers):
6, 7, 8, 15, 23, 31, 34, 42, 43, 44,
47, 48, 66, 88, 94, 98, 99, 104, 105, 114,
115, 118, 133, 134, 137, 138, 141, 142, 143, 144,
158, 164, 168, 171, 181, 183, 198, 203, 205, 206,
207, 208, 218, 227, 229, 230, 233, 234, 248, 249,
250, 251, 252, 261, 262, 291, 299, 324, 325, 326

Record: 1

Field	Value
ASSINANTE	112013311-
TIPO_CLIEN	C
TIPO_LIGAC	03
FAIXA_TARI	1
DISTANCIA_	01
FIM_SEMANA	N
MINUTOS	31.00000
RECEITA	1.70000
NUM_LIGACO	9.00000

Field	Rule #
FAIXA_TARI	3, 4, 7, 8, 9
FIM_SEMANA	2, 5, 6, 7, 9
MINUTOS	8, 9
NUM_LIGACO	8, 9

For Help, press F1

NUM 14,48

FIGURA A5: Relatório das regras geradas e dos casos de desvio.

APÊNDICE B - WIZWHY PASSO A PASSO

O WizWhy, já apresentado no item 4.2, é uma ferramenta de predição de casos futuros, baseados nas regras geradas pelo WizRule.

Para dar início à etapa de predição, é preciso clicar em Issue Prediction. Alguns valores precisam ser especificados e o resultado será mostrado a seguir.

WizWhy Predictor

Data Source: MEUS DOCUMENTOS/TESE_NAN

Field to Predict: TIPO_CLIENTE

Condition Fields:

	Field Name	Field Value
1	TIPO_LEAD	Unknown
2	FAIXA_TARI	Unknown
3	FIM_SEMANA	Unknown
4	NUM_LIGACO	
5	MINUTOS	

FIGURA B1: Definição de alguns valores relevantes para a predição de novos casos.

WizWhy Predictor

Data Source: MEUS DOCUMENTOS/TESE_NAN

Field to Predict: TIPO_CLIENTE

Condition Fields:

	Field Name	Field Value
1	TIPO_LEAD	01
2	FAIXA_TARI	5
3	FIM_SEMANA	N
4	NUM_LIGACO	5
5	MINUTOS	10

FIGURA B2: Exemplo de valores definidos para a Figura B1.

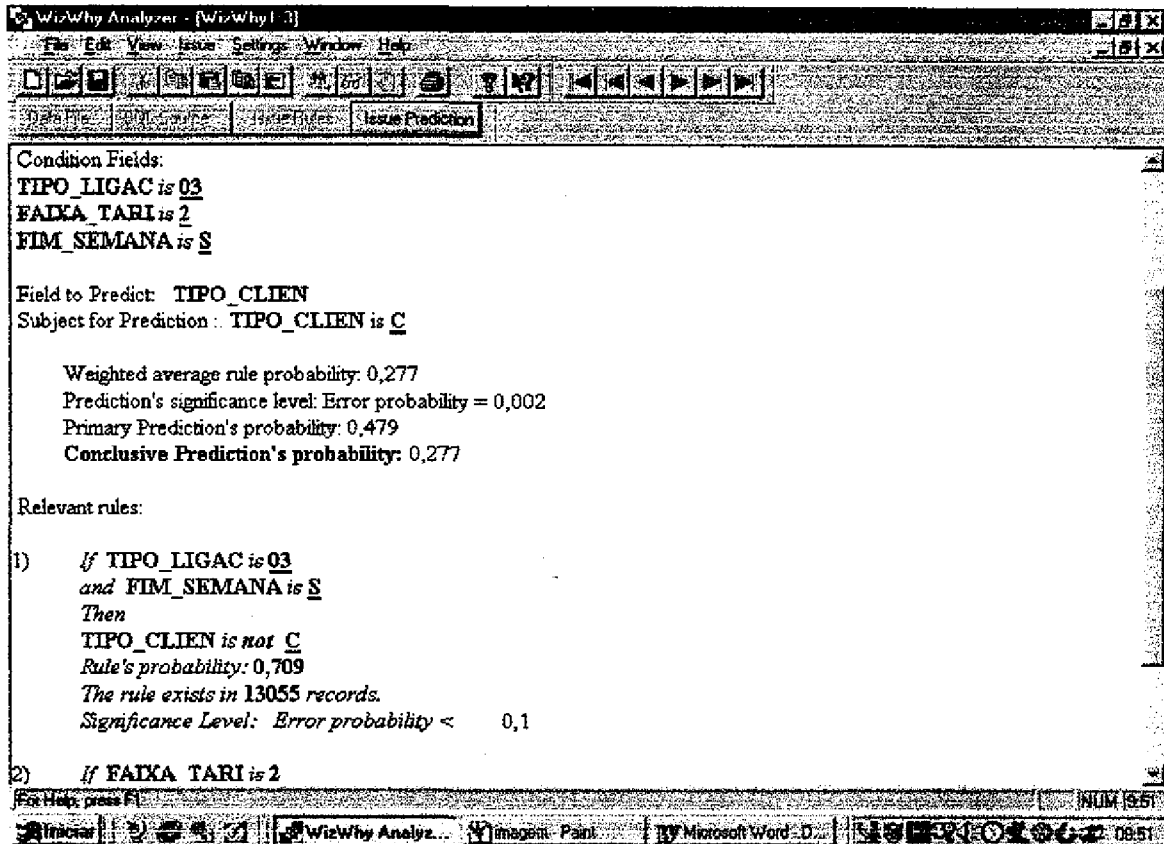


FIGURA B3: Resultado da predição dos valores especificados na Figura B2.

As regras geradas podem ser exportadas para a linguagem SQL e utilizadas em bases de dados de maior porte.

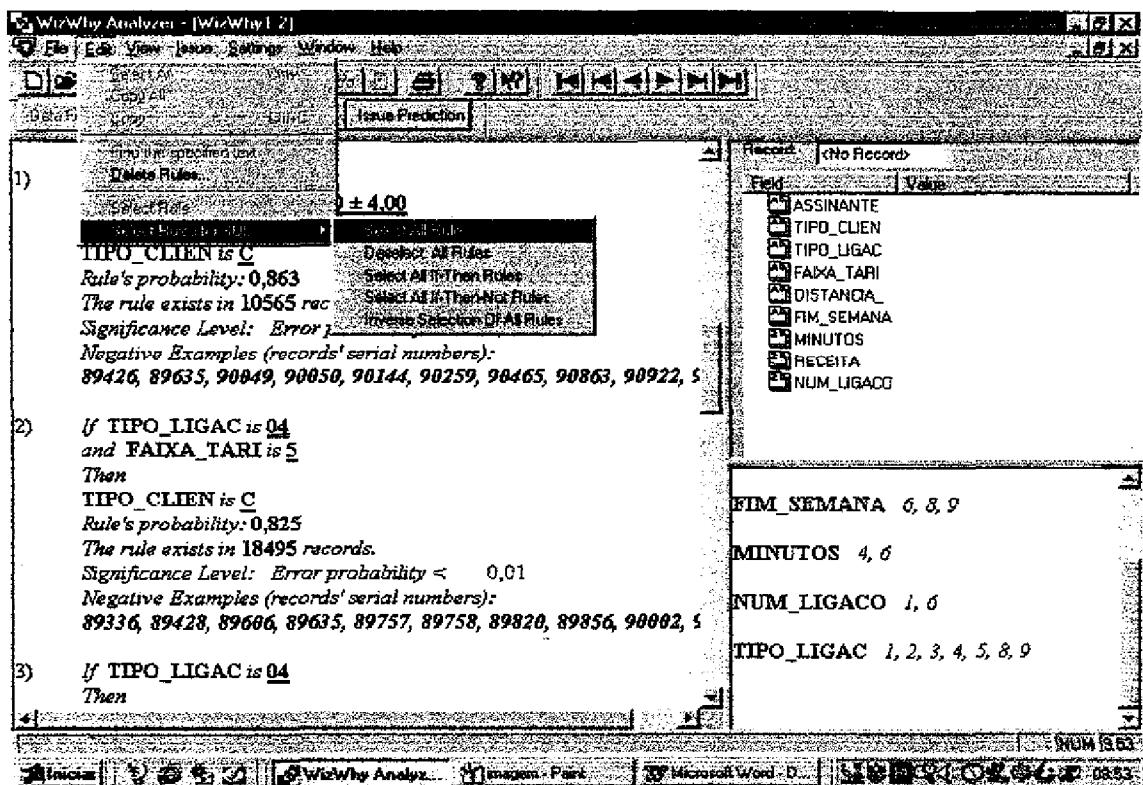


FIGURA B4: Seleção das regras de interesse para exportação.

As regras destacadas são aquelas de interesse para o estudo de caso em questão e, portanto, selecionadas para exportação.

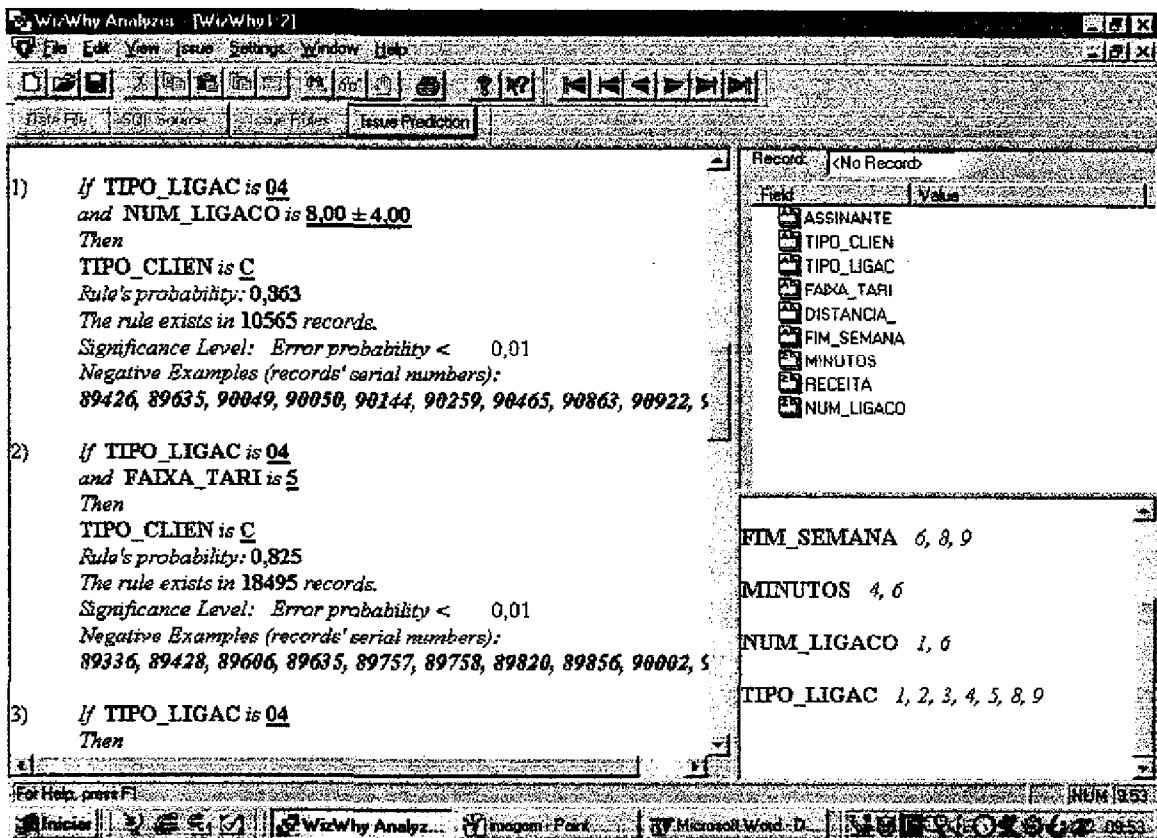


FIGURA B5: Resultado da seleção ocorrida na Figura B4.

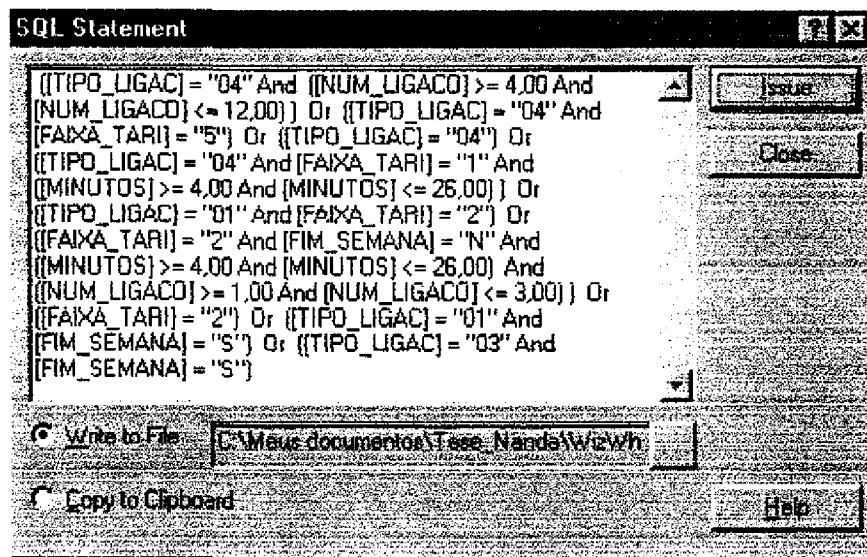


FIGURA B6: Resultado da exportação feita.

APÊNDICE C - XPERTRULE MINER PASSO A PASSO

O script de geração das regras encontra-se na janela a seguir. O objetivo do caso em questão, é gerar uma árvore de mineração capaz de classificar os clientes de acordo com o perfil de suas ligações.

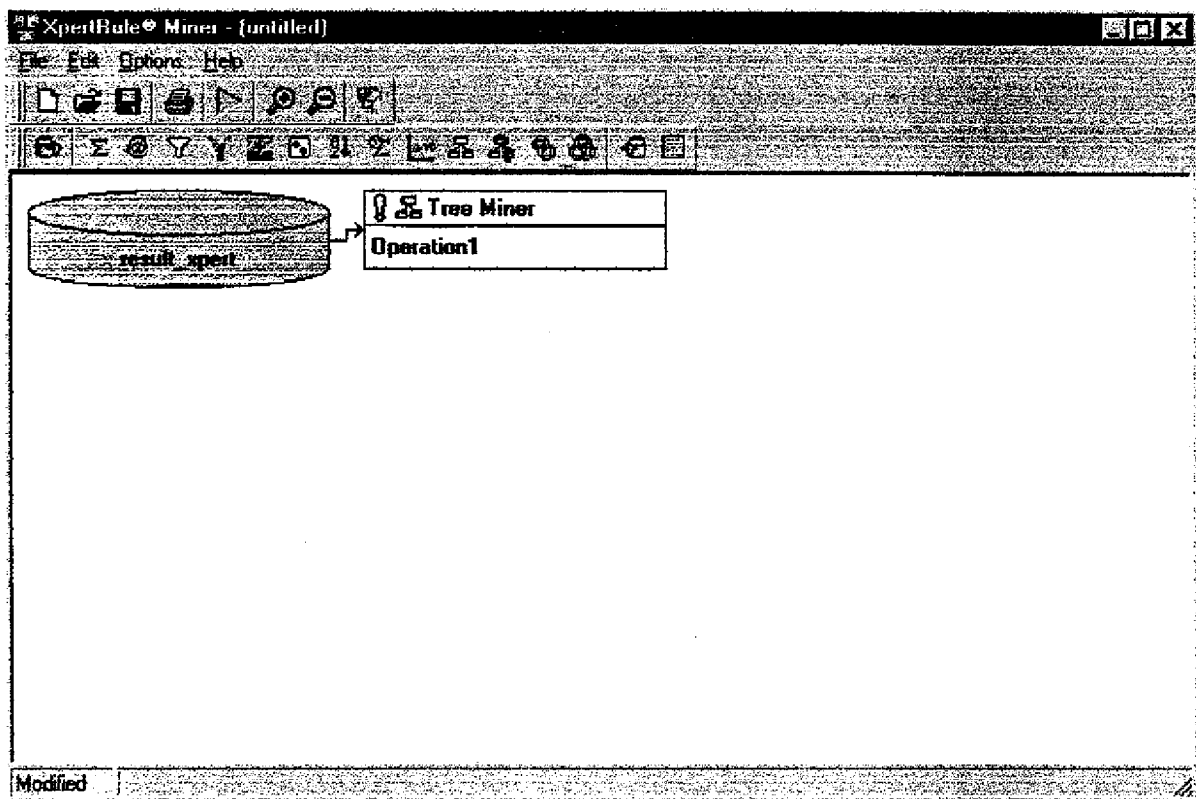


FIGURA C1: Script de mineração da base de dados selecionada.

É preciso, neste passo, especificar quais os campos a serem trabalhados e qual a saída a ser gerada (Outcome – folha da árvore). Veja que o XpertRule Miner gera uma estatística referente ao campo especificado como saída.

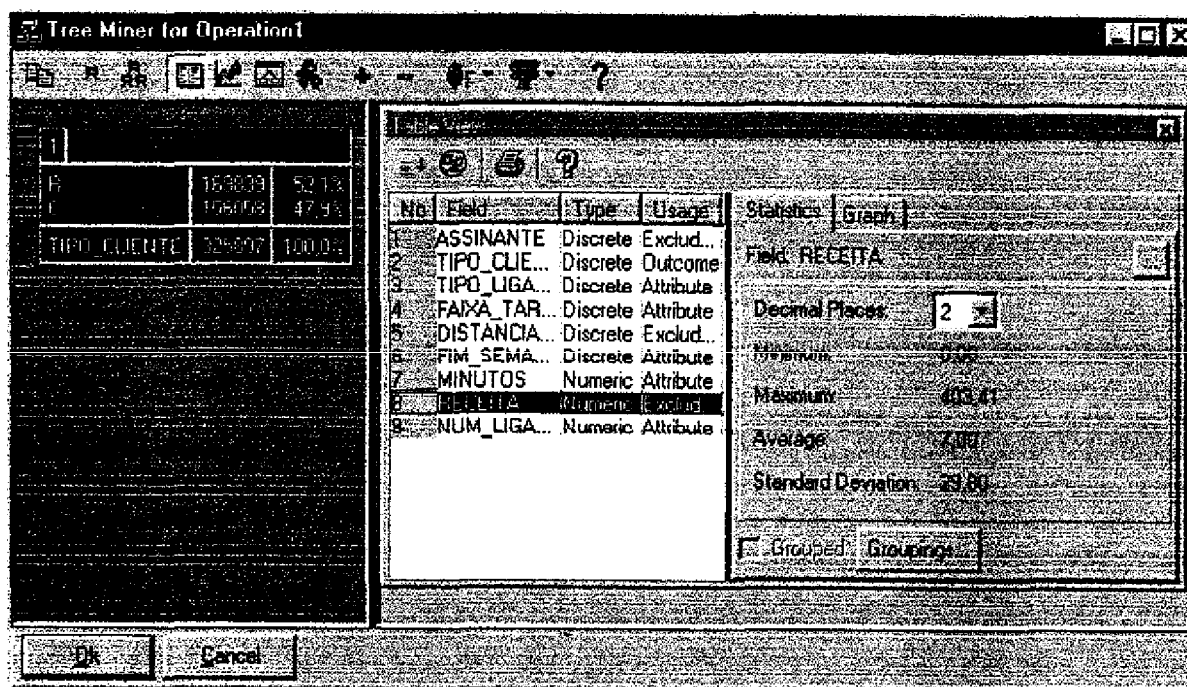


FIGURA C2: Estatística gerada acerca da base de dados em questão.

Alguns parâmetros precisam ser definidos, tais como número mínimo de casos para uma regra, nível máximo de significância para cada nó e o critério de divisão dos ramos da árvore (Entropia ou Chi-Quadrado).

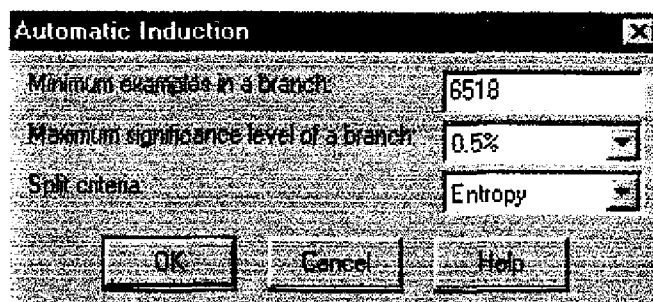


FIGURA C3: Especificação de parâmetros para a geração da árvore de decisão.

Feito isto, a árvore é gerada e o resultado é mostrado na tela a seguir.

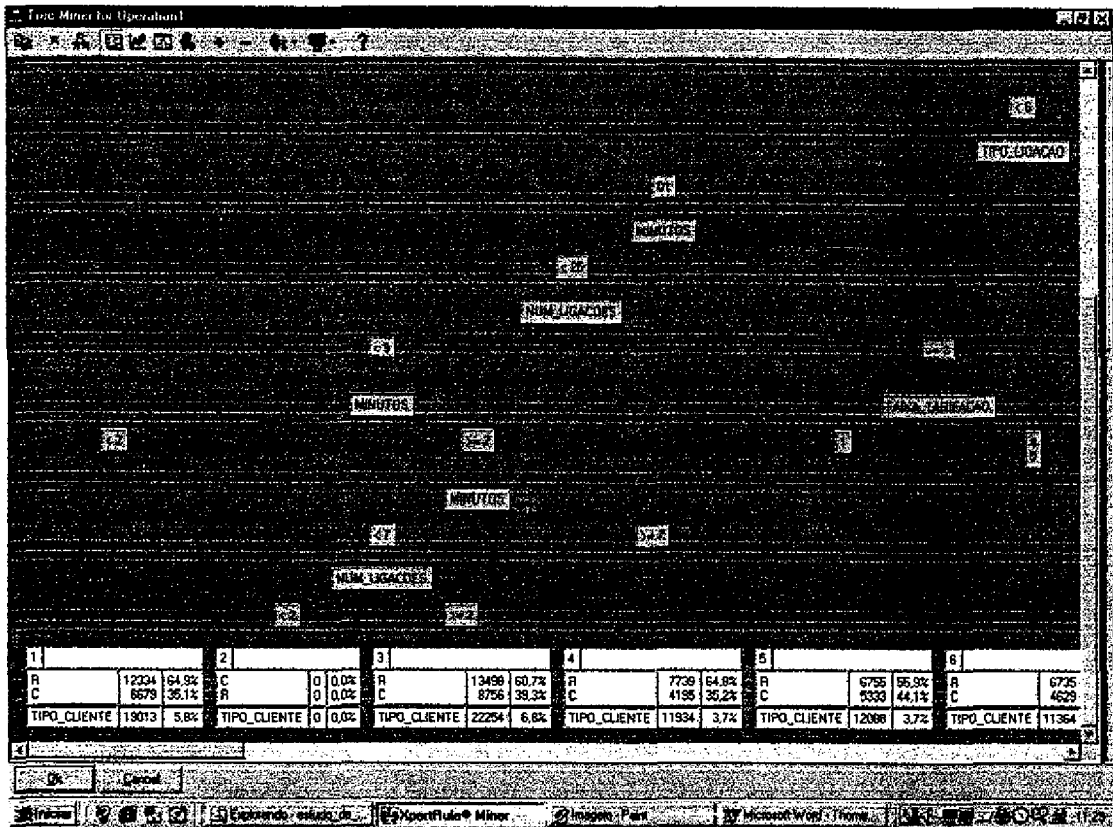


FIGURA C4: Geração da árvore de decisão.

O resultado da árvore pode ser exportado para vários formatos: linguagem SQL, texto, código C, código Pascal, SAS e figura.

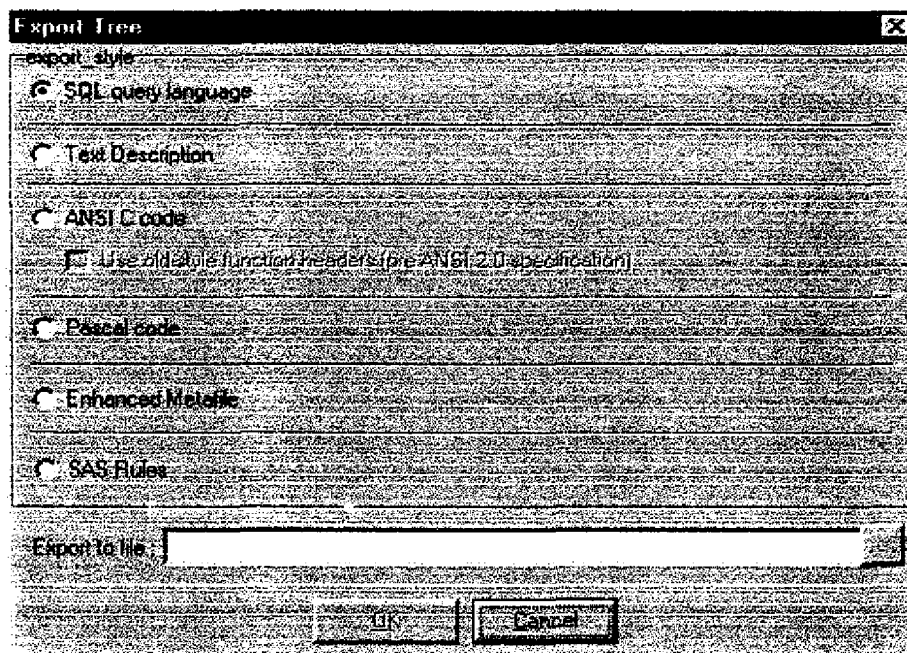


FIGURA C5: Escolha da forma de exportação da árvore de decisão gerada.

APÊNDICE D - BUSINESSMINER PASSO A PASSO

Ao iniciar um novo projeto, a seguinte tela do Business Miner aparecerá para que seja especificada a fonte dos dados a serem trabalhados (banco de dados Access, tabela do Excel, arquivo texto ou do SPSS, etc.). Feito isto, é preciso também, definir o caminho em que se encontram estes dados, bem como aquele em que o resultado deste novo projeto será salvo.



FIGURA D1: Tela inicial do BusinessMiner.



FIGURA D2: Seleção da base de dados a ser trabalhada.

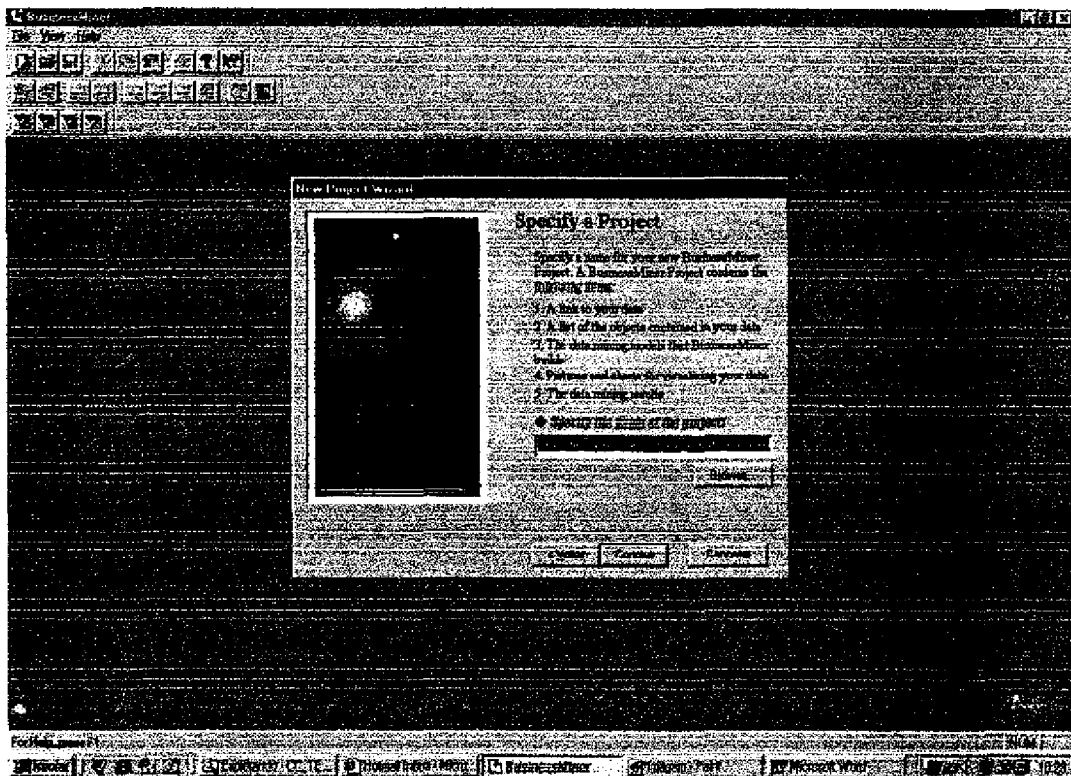


FIGURA D3: Especificação de um novo arquivo, no qual as informações obtidas a partir deste processamento serão armazenadas.

Aqui é iniciada a etapa de mineração de dados propriamente dita. Os campos existentes na base de dados são descritos e, nesta janela, devem ser selecionados os atributos relevantes para a geração das regras de produção e da árvore de decisão. Caso não se tenha uma idéia de quais atributos não contribuem positivamente para o bom desempenho do Business Miner, é recomendado que todos os campos estejam definidos como Yes na opção "Mine".

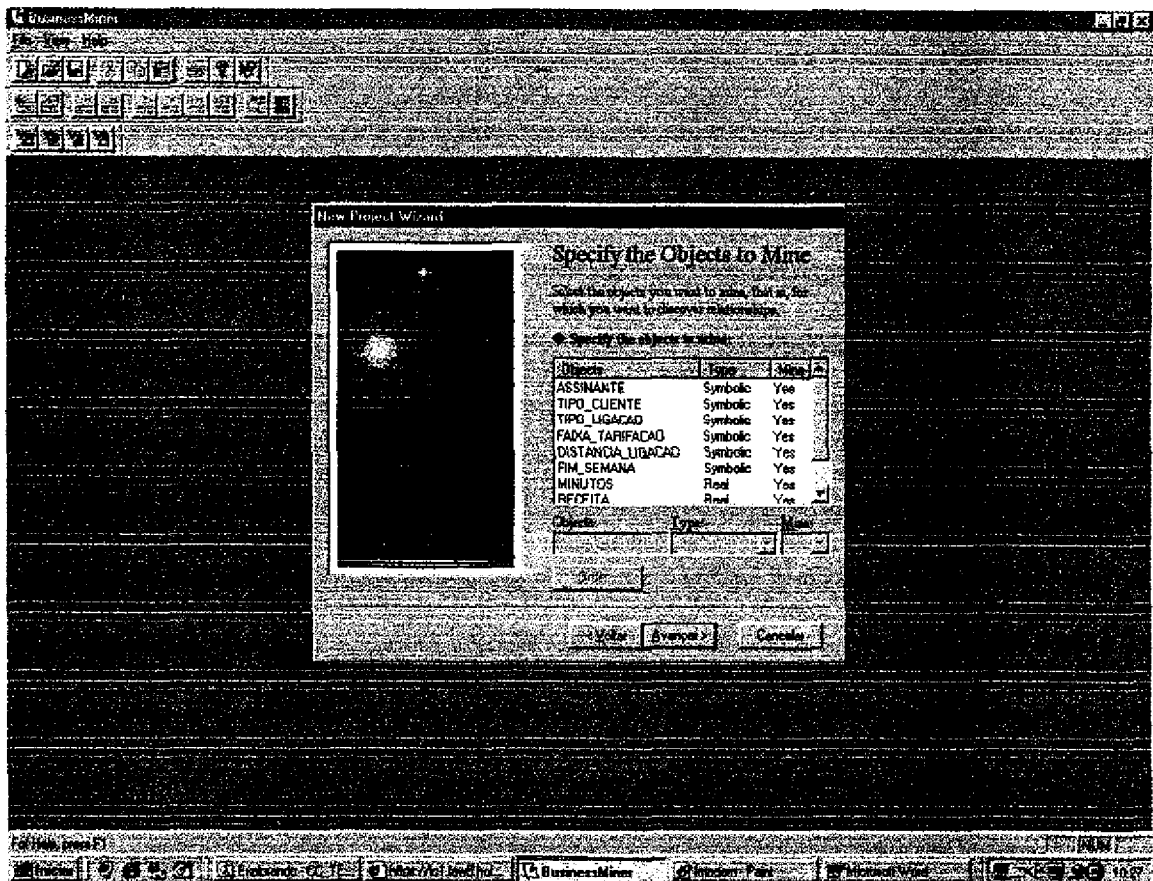


FIGURA D4: Apresentação dos atributos contidos na base de dados.

No caso em questão, alguns atributos serão especificados como relevantes e outros, como irrelevantes para o processo de mineração de dados. Vejamos a seguir:

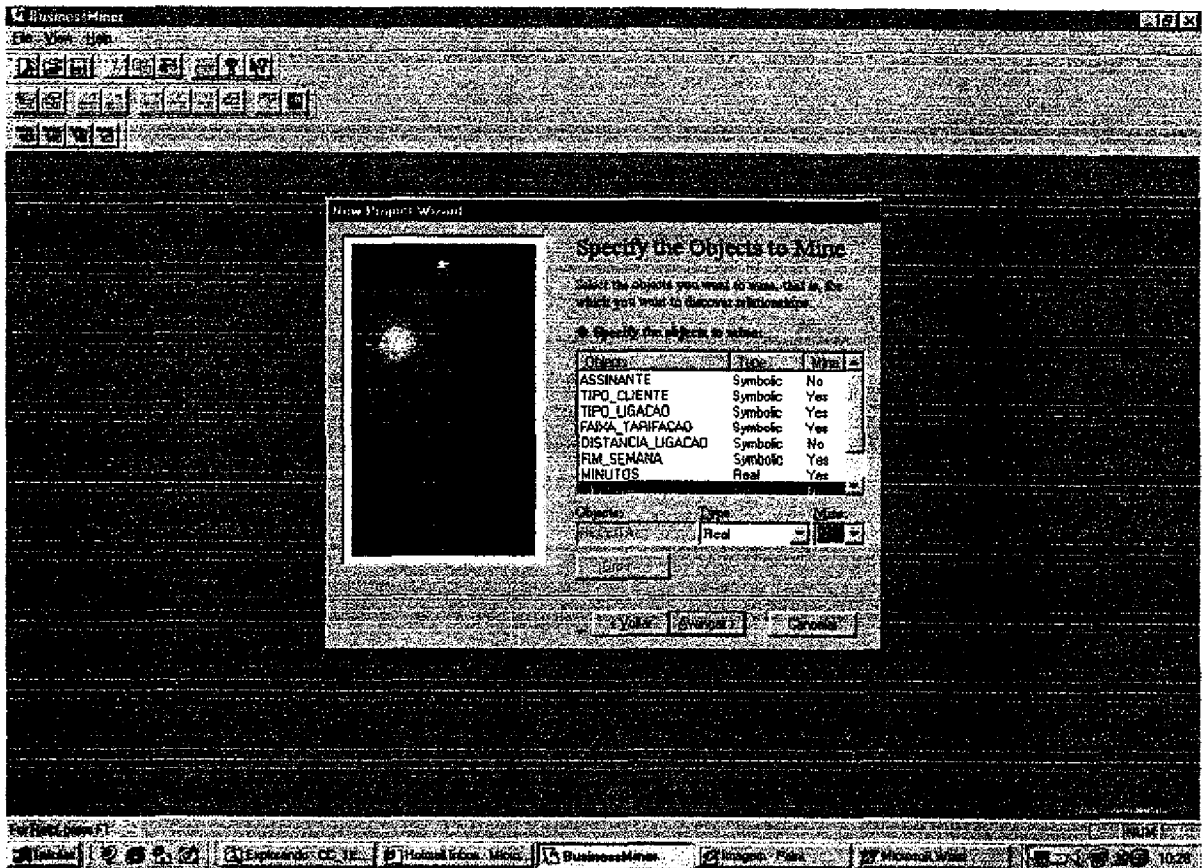


FIGURA D5: Especificação dos atributos relevantes para o caso em questão.

O próximo passo é a definição do atributo de saída, ou seja, aquele que representa o objetivo do estudo de caso: Tipo_cliente (residencial ou comercial).



FIGURA D6: Especificação do atributo de saída.

A progressão do processo de importação dos dados é demonstrada na janela abaixo.

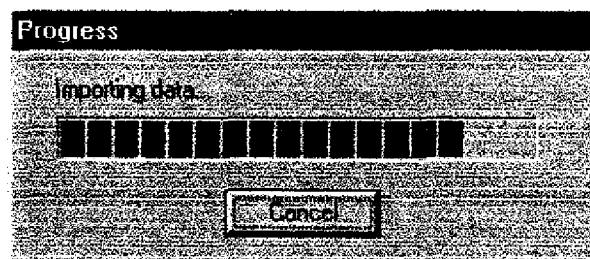


FIGURA D7: Indicando o andamento da criação da árvore de decisão.

Ao final do processamento, uma janela indicará a quantidades de nós criados na árvore de decisão. A seguir, a visualização desta se fará possível.

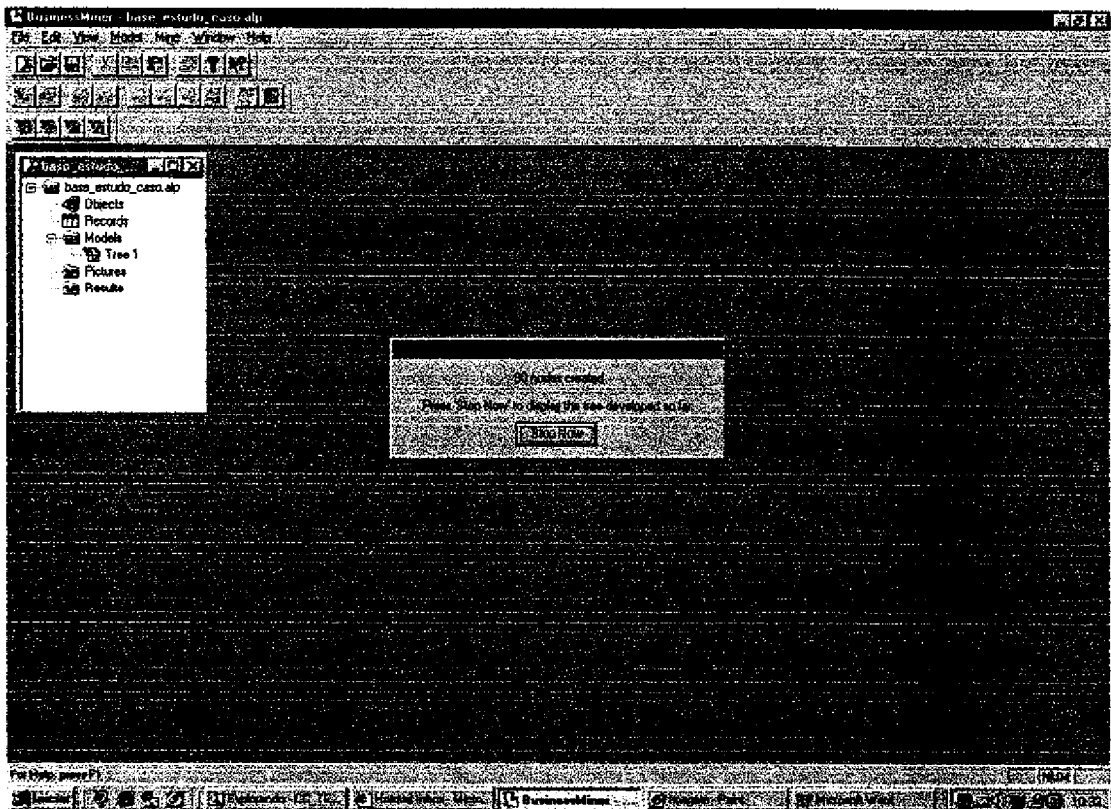


FIGURA D8: Indicação do número de nós criados na árvore.

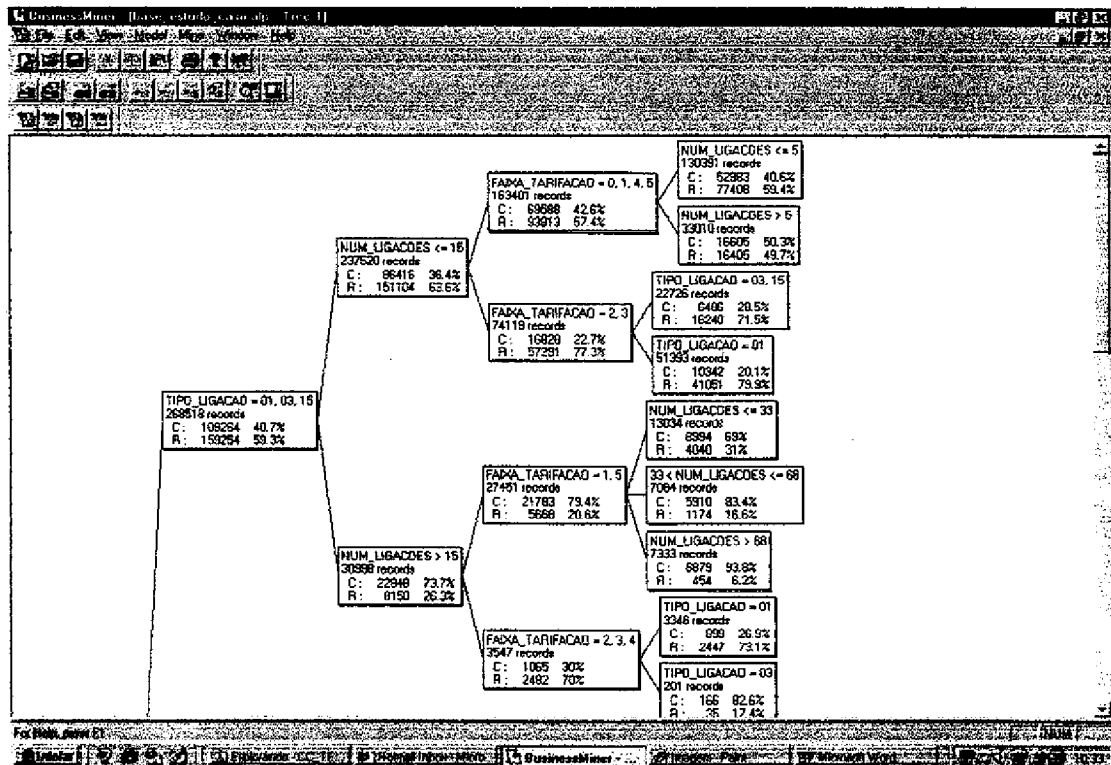


FIGURA D9: Resultado da árvore de decisão.

Uma outra opção do Business Miner, além da árvore de decisão, é o módulo de geração de regras de produção. Nesta janela, é possível definir o parâmetro desejado como saída: Tipo_cliente = R (residencial) ou Tipo_cliente = C (comercial); bem como a confiabilidade mínima para a geração de cada regra: 70%.

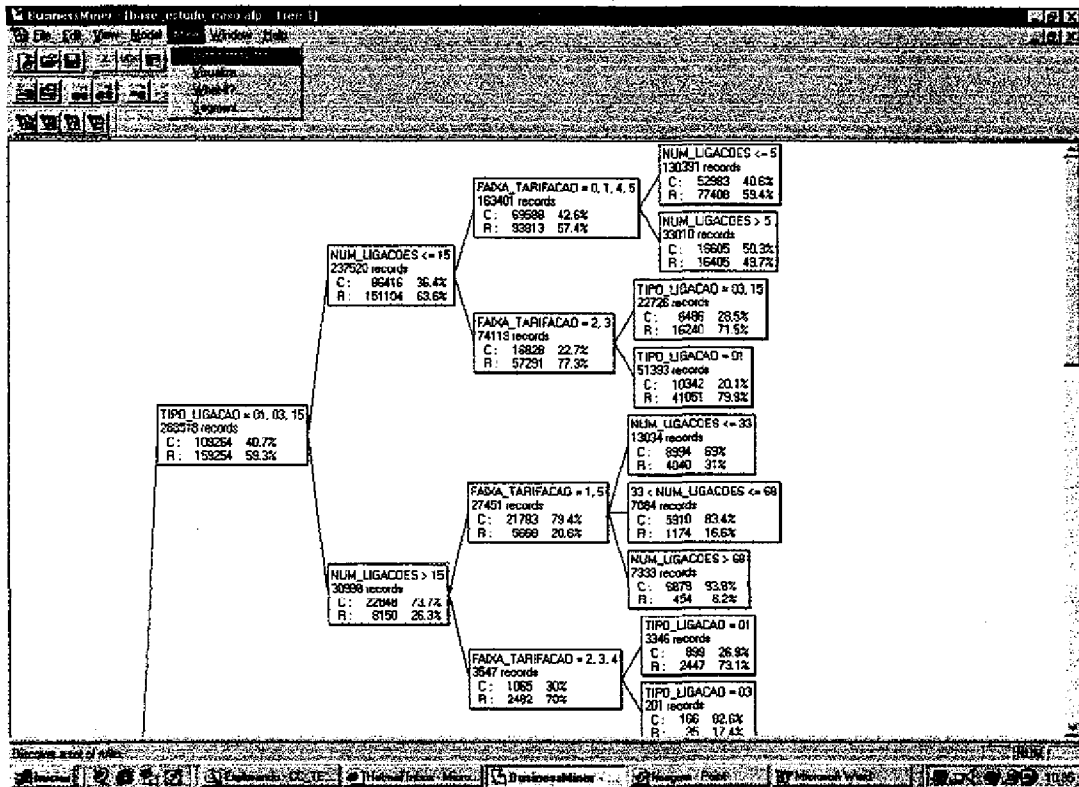


FIGURA D10: Geração de regras a partir da árvore apresentada na Figura D9.

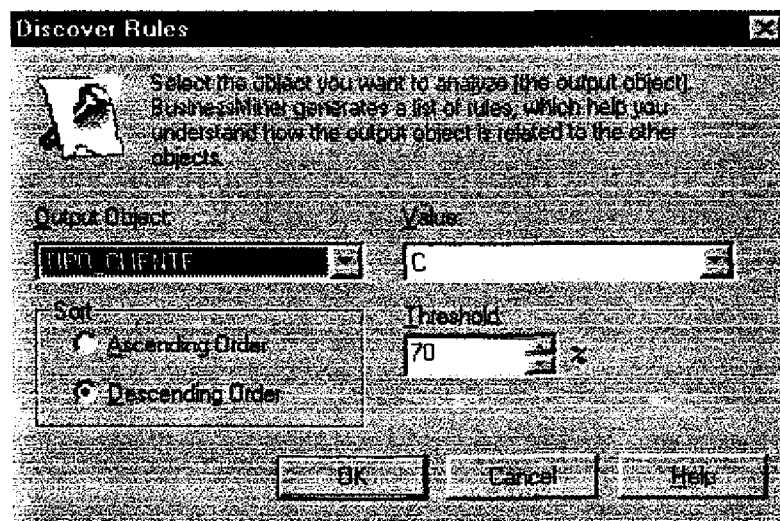


FIGURA D11: Especificação do atributo de saída para as regras geradas.

A janela, Figura D12 abaixo, descreve todas as regras encontradas, a população abrangida por cada regra e o grau de confiança destas.

Percentage of population	in node
71.9	562 TIPO_LIGACAO = 04, 12, 06, 08 NUM_LIGACOES <= 3 FAIXA_TARIFACAO = 2 TIPO_LIGACAO = 06, 08, 12 TIPO_LIGACAO = 04, 12, 06, 08 NUM_LIGACOES > 10 TIPO_LIGACAO = 06
72.8	125 TIPO_LIGACAO = 04, 12, 06, 08 NUM_LIGACOES > 10 TIPO_LIGACAO = 06
73.7	30998 TIPO_LIGACAO = 01, 03, 15 NUM_LIGACOES > 15
73.7	1008 TIPO_LIGACAO = 04, 12, 06, 08 3 < NUM_LIGACOES <= 10 FAIXA_TARIFACAO = 2, 3 NUM_LIGACOES <= 6
76	34866 TIPO_LIGACAO = 04, 12, 06, 08 NUM_LIGACOES <= 3
77	27097 TIPO_LIGACAO = 04, 12, 06, 08 NUM_LIGACOES <= 3 FAIXA_TARIFACAO = 1, 2, 4, 5 TIPO_LIGACAO = 04
77.5	1548 TIPO_LIGACAO = 04, 12, 06, 08 3 < NUM_LIGACOES <= 10 FAIXA_TARIFACAO = 2, 3
78.4	29874 TIPO_LIGACAO = 04, 12, 06, 08 NUM_LIGACOES <= 3 FAIXA_TARIFACAO = 1, 2, 4, 5
79.4	27451 TIPO_LIGACAO = 01, 03, 15 NUM_LIGACOES > 15 FAIXA_TARIFACAO = 1, 3
81.6	57379 TIPO_LIGACAO = 04, 12, 06, 08
82.2	2393 TIPO_LIGACAO = 04, 12, 06, 08 3 < NUM_LIGACOES <= 10 FAIXA_TARIFACAO = 1, 4, 5

FIGURA D12: Apresentação das regras geradas para o atributo de saída especificado.

Outra opção de ajuda no entendimento dos dados analisados é a visualização de gráficos relacionados aos atributos da base de dados.

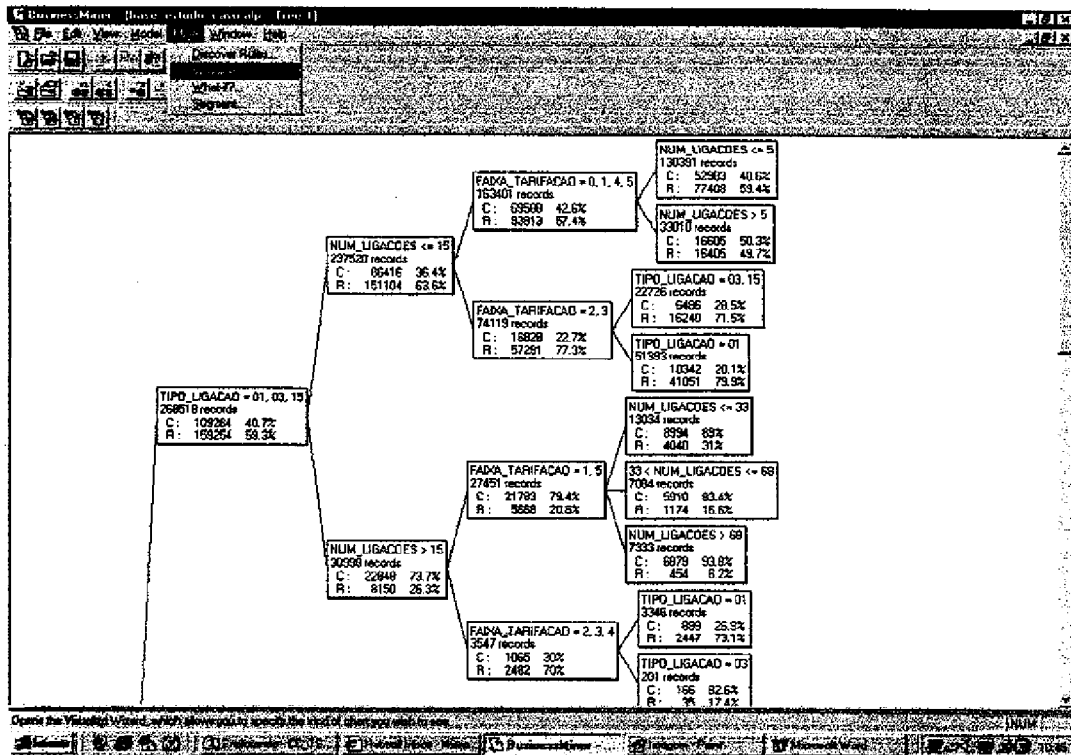


FIGURA D13: Visualização dos dados da base na forma gráfica.

Uma variedade bastante grande de gráficos 2D e 3D está disponível para visualização dos dados.

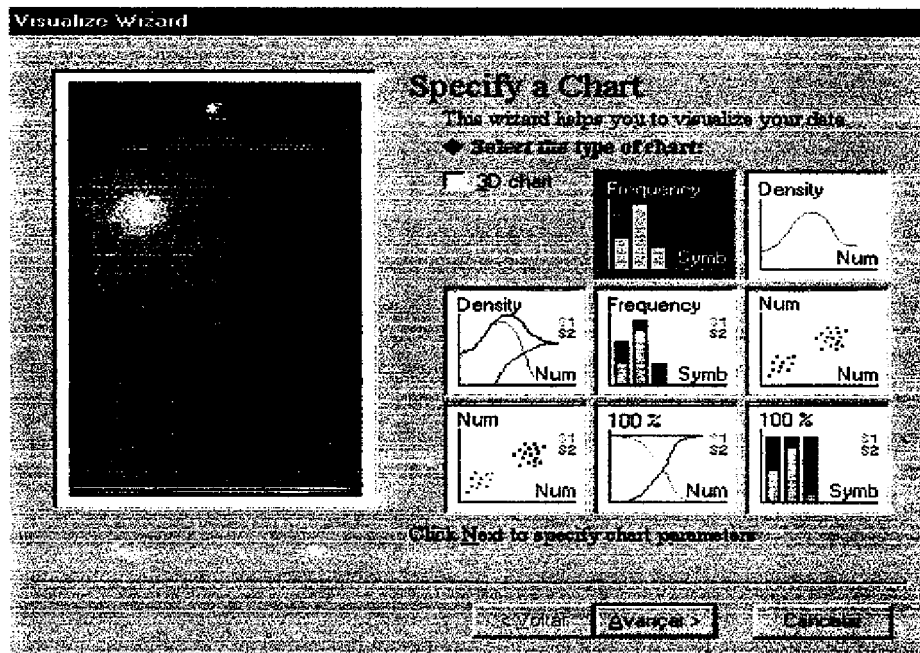


FIGURA D14: Tipos de gráficos suportados pelo BusinessMiner.

A parte de predição é apresentada na Figura, D15, a seguir.

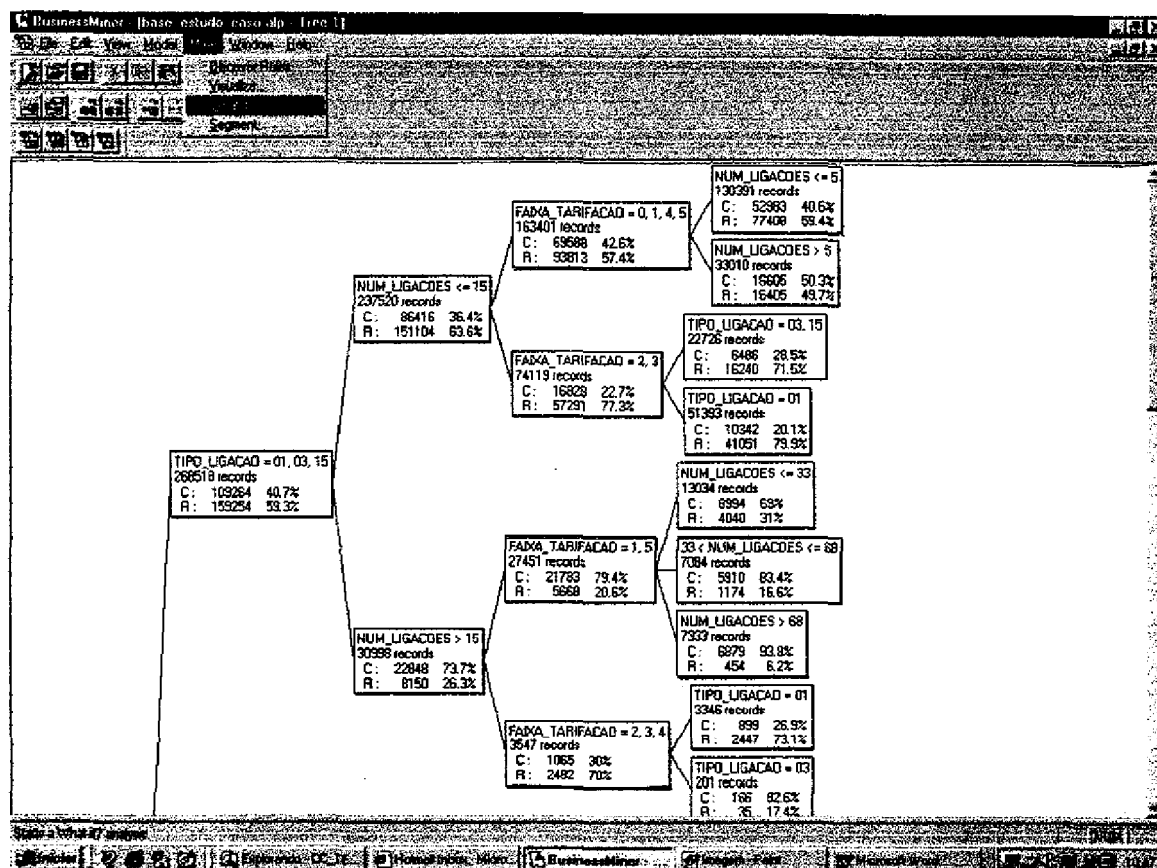


FIGURA D15: Predição de casos futuros.

É preciso especificar supostos valores para os atributos solicitados ou defini-los como desconhecidos ("unknown"). O resultado da predição é representado pela probabilidade da regra se adequar ao perfil residencial ou ao perfil comercial.

No caso a seguir, tipo da ligação é definido como "DDD", a faixa de tarifação é super-reduzida (50% mais barata em relação à normal) e o número de ligações é = 10. Como resultado, temos que esta regra abrange um cliente com perfil residencial.

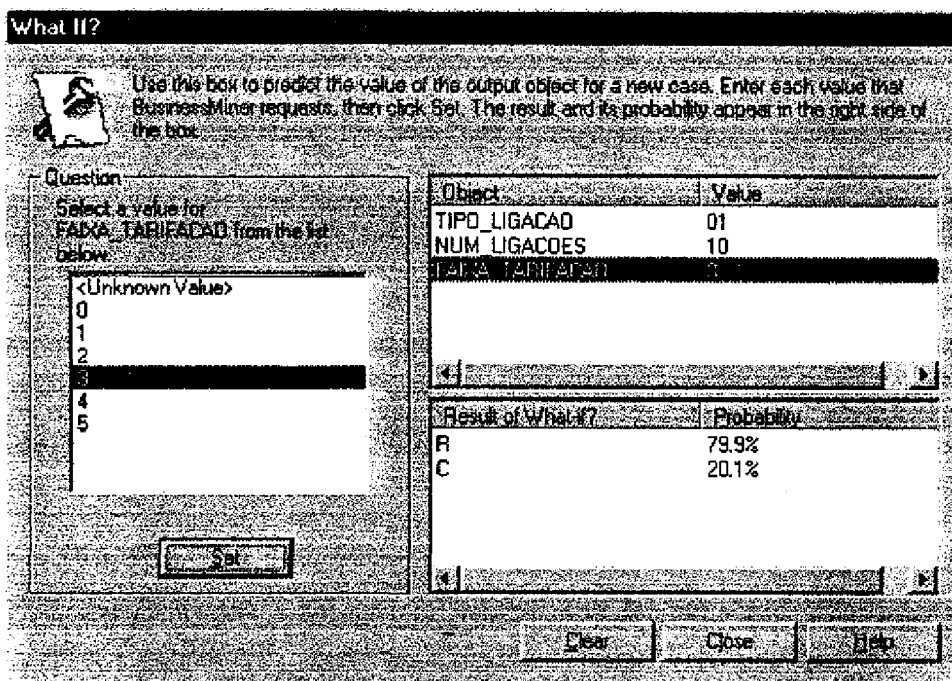


FIGURA D16: Especificação de valores para um novo caso.

Vejamos uma outra predição em que o tipo da ligação é definido como “à cobrar”, o número de ligações é = 55 e a faixa de tarifação é diferenciada (80% mais cara que a normal).

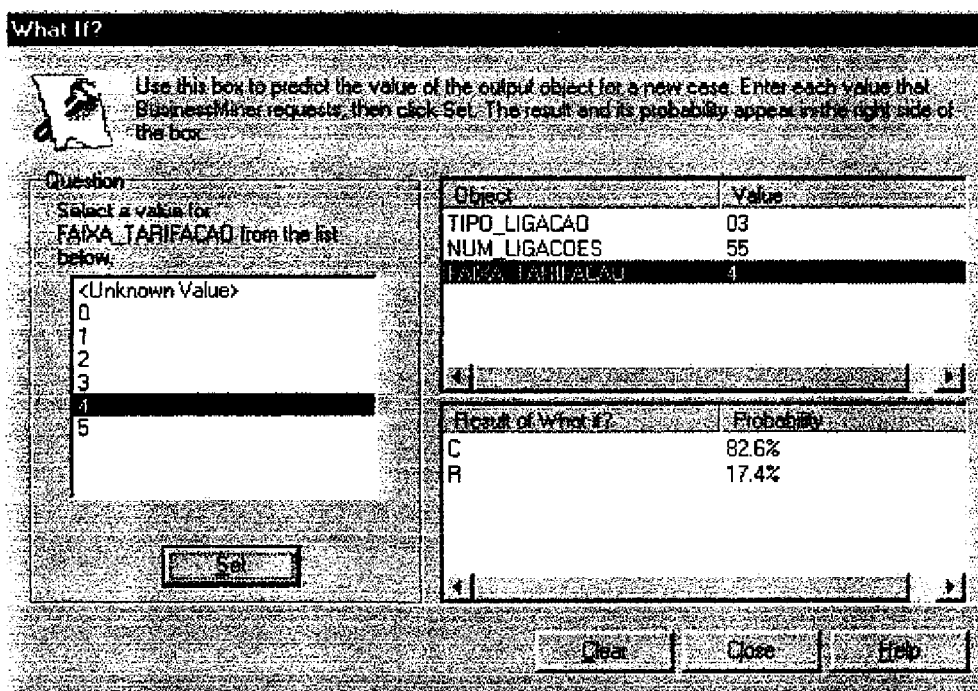


FIGURA D17: Outro exemplo de um novo caso.

REFERÊNCIAS BIBLIOGRÁFICAS

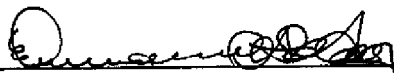
- [1] Basserville, M., Nikiforov, I. V. Detection of Abrupt Changes: Theory and Application, Englewoods Cliffs - NJ, Prentice Hall, 1993.
- [2] Bigus, J.P. Data mining with neural networks:solving business problems - from application development to decision support, McGraw Hill, Inc, 1996.
- [3] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C.J. Classification and Regression Trees, Belmont, California: Wadsworth, 1984.
- [4] Collins, T; Rapp S. The great marketing turnaround. Englewood Cliffs, NJ: Prentice Hall, 1990.
- [5] Cortes, C. Giga Mining. Knowledge Discovery in Databases, 1998.
- [6] Cover, T.M.; Hart, P.E. Nearest Neighbor Pattern Classification, IEEE Transactions on Information Theory 13:21-17, 1967.
- [7] Diaconis, P.; Friedman, D. Asymptotics of Graphical Projection Pursuit, Annals of Statistics, 1984.
- [8] Fayyad, U. M., Shapiro, G. P.; Smyth, P.; Uthurusamy, R. Advances in Knowledge Discovery and Data Mining, AAAI Press, California, 1996.
- [9] Friedman, J.H.; Tukey, J.W. A Projection Pursuit Algorithm for Exploratory Data Analysis, IEEE Transations on Computers 23: 881-889, 1974.
- [10] Frawley, W.J.; Piatesky-Shapiro,G.; Matheus, C.J. Knowledge Discovery in Databases: An Overview, Piatsky-Shapiro, G. and Frawley, W.J. Eds., Knowledge Discovery in Databases, MIT Press, Cambridge, MA, 1991.
- [11] Genaro, S. Sistemas Especialistas - O Conhecimento Artificial, LTC Editora, 1987.
- [12] Hand, D. J. Discrimination and Classification, Chichester, U.K.: John Wiley and Sons, 1981.
- [13] Hastie, T.; Tibshirani, R.; Buja, A. Flexible Discriminant Analysis by Optimal Scoring, Journal of American Statistical Association 89 (428): 1255-1270, 1994.
- [14] Jackson, R; Wang, P. Strategic database marketing. Lincolnwood, IL: NTC Business Book, 1994.
- [15] Komorowski, J.; Polkowski, L.; Skowron, A. - Rough Sets: A Tutorial, Institute of Mathematics, Warsaw University.

- [16] Kotler, P. Administração de Marketing: análise, planejamento, implementação e controle. Editora Atlas, 1998.
- [17] Lowe, D.; Webb, A.R. Optimized Feature Extraction and the Bayes Decision in Feed-Forward Classifier Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 13: 355-364, 1991.
- [18] Luiz, João A. B.- Aplicação de Técnicas de Mineração de Dados, Tese de Mestrado, IME, 1996
- [19] Machado, B. P. Aplicação de Redes Neurais para o tratamento de Grandes Massas de Dados, Tese de Mestrado em Sistemas e Computação, IME, RJ, 1992.
- [20] McLachlan, G. Discriminant Analysis and Statistical Pattern Recognition, New York: Wiley, 1992.
- [21] Naliato, Fernanda C. - Aplicação de Técnicas de Mineração de Dados - Estudo de Caso em Marketing Direto, Tese de Mestrado, IME, 1999
- [22] Passos, Emmanuel P. - Data Mining, Notas de Aula, PUC-Rio.
- [23] Pawlak, Z. - Data Mining – A Rough Set Perspective, Polish Academy of Sciences.
- [24] Polkowski L., Skowron A. - Rough Sets in Knowledge Discovery, Methodology and Applications, Vol. 1-2, Physica-Verlag.
- [25] Quinlan, J. C4.5: Programs for Machine Learning, San Francisco: Morgan Kaufmann, 1993.
- [26] Scott, D.W. Multivariate Density Estimation: Theory, Practice, and Visualization, New York: Wiley, 1992.
- [27] Silverman, B. Density Estimation for Statistics and Data Analysis, New York: Chapman and Hall, 1986.
- [28] Simoudis, E.B., Livezey, Kerber R. Using RECON for data cleaning, Proceedings of First International Conference on Knowledge Discovery and Data Mining, AAAI Press, pp. 282-287, 1995.
- [29] Skowron, A.; Stepaniuk, J. - Generalized Approximations Spaces. Proceedings of the Third International Workshop on Rough Sets and Soft Computing, San Jose, 1994.
- [30] Slowinski, R. - A generalization of the indiscernibility relation for rough set analysis of quantitative information. Rivista di Matematica per le Scienze Economiche e Sociali, 1992.
- [31] Stefanowski J. - On Rough Set Based Approaches to Induction of Decision Rules, Rough Sets in Knowledge Discovery Vol. 1, Physica-Verlag.

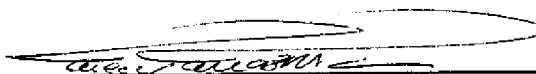
- [32] Weiss, S.I.; Kulikowski, C. *Computer Systems that Learn: Classification e Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems*, San Francisco, California, Morgan Kaufmann, 1991.
- [33] Werbos, P. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*, Ph. D. diss., Harvard, August, 1974.
- [34] Wergend, A.; Gershenfeld, N. *Predicting the Future and Understanding the Past*, Redwood City, California: Addison-Wesley, 1993.
- [35] Ziarko, W. - Variable Precision Rough Set Model., *Journal of Computer and System Sciences*, 1993

MINERAÇÃO DE DADOS COM TÉCNICAS DE ROUGH SETS

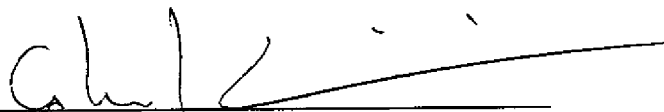
Dissertação de Mestrado apresentada por *Dante José Alexandre Cid* em 18 de dezembro de 2000 ao Departamento de Engenharia Elétrica da PUC-Rio e aprovada pela Comissão Julgadora, formada pelos seguintes membros:



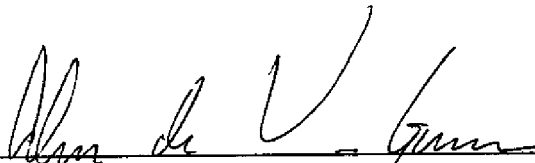
Prof. Emmanuel Piseces Lopes Passos
DEE/PUC-Rio - (Orientador)



Profa. Marley Maria Bernardes Reuzzi Vellasco
DEE/PUC-Rio (Co-Orientadora)



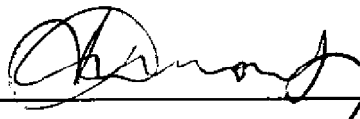
Prof. Josefino Cabral M. Lima
UFRJ



Prof. Alex de V. Garcia
IME

Visto e permitida a impressão

Rio de Janeiro, 20/12/2000



Prof. Ney Augusto Dumont
Coordenador Programas de Pós-Graduação e
Pesquisa do Centro Técnico Científico