UNIVERSIDADE FEDERAL DE PELOTAS
INSTITUTO DE FÍSICA E MATEMÁTICA
DEPARTAMENTO DE INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MINERAÇÃO DE DADOS APOIADA PELA DESCOBERTA DE SUBGRUPOS ATRAVÉS DO PÓS-PROCESSAMENTO DE REGRAS DE ASSOCIAÇÃO

JOEL PINHO LUCAS

PELOTAS, 2006

JOEL PINHO LUCAS

MINERAÇÃO DE DADOS APOIADA PELA DESCOBERTA DE SUBGRUPOS ATRAVÉS DO PÓS-PROCESSAMENTO DE REGRAS DE ASSOCIAÇÃO

Trabalho acadêmico apresentado ao Curso de Bacharelado em Ciência da Computação da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Profª. Msc. Ana Marilza Pernas Fleischmann

Co-orientador: Prof. Dr. Amauri Almeida Machado

Co-orientador: Prof. Dr. Alípio Mário Jorge

PELOTAS, 2006

BANCA EXAMINADORA:

_____

Profª. Ana Marilza Pernas Fleischmann, Msc. (Orientadora)

_____

Profª. Flávia Braga de Azambuja, Msc.

_____

Prof. Lucas Ferrari de Oliveira, Dr.

AGRADECIMENTOS

Primeiramente, agradeço aos meus pais pelo exemplo de vida e por toda dedicação prestada a mim, sem ela não seria possível chegar ao final desta etapa com êxito.

Agradeço também a Deus pela vida.

Agradeço à minha orientadora, professora Ana Marilza, pela dedicação e pelo apoio, os quais foram imprescindíveis para a realização deste trabalho. Ao meu co-orientador, professor Amauri Machado, pela dedicação e incentivo. Ao também co-orientador, professor Alípio Jorge, pela dedicação e pelo voto de confiança.

Agradeço aos meus colegas de grupo de pesquisa Beatriz, Gustavo e Michele, pela amizade e agradável convivência no laboratório. Também agradeço à Beatriz pela paciência e por todo o período de trabalho em conjunto, no qual aprendi muito.

Agradeço também ao ex-colega de curso, Márcio Basgalupp, pela valiosa ajuda na escolha da área do trabalho: a decisão não poderia ter sido mais acertada.

Agradeço aos professores do curso por todos ensinamentos, em especial ao professor Ricardo Silveira pela orientação durante o período de iniciação científica: a orientação e as oportunidades concedidas contribuíram enormemente para minha formação.

Agradeço à professora Rozane Alves pelo auxílio prestado durante todo o curso e também pela orientação durante o período de estágio no sistema acadêmico, onde pude enriquecer significativamente minha formação.

Gostaria de agradecer também aos meus amigos e colegas pelo incentivo e compreensão durante o período da realização deste trabalho.

RESUMO

A utilização da técnica de descoberta de regras de associação para obter conhecimento a partir de bancos de dados, com grandes dimensões, resulta em uma quantidade muito elevada de regras. A exploração dos resultados obtidos torna-se uma tarefa complexa e limita o uso de técnicas para a Descoberta do Conhecimento. Na tentativa de contornar esta limitação, desenvolveu-se uma ferramenta, na linguagem Java, que baseia-se em uma metodologia que propõe a convergência de princípios de regras de associação e da descoberta de subgrupos, onde utiliza-se, na visualização de subgrupos, técnicas utilizadas no pós-processamento de regras de associação. Sendo assim, pode-se fazer uso da grande quantidade de regras obtidas para serem o ponto de partida na busca por subgrupos interessantes em uma população. Através da ferramenta desenvolvida, possibilita-se ao analista participar de um processo interativo de Descoberta de Subgrupos por meio de um ambiente gráfico, no qual pode utilizar seu conhecimento acerca do domínio durante o processo de identificação de tais subgrupos em meio à população. Com o intuito de validar a ferramenta, realizou-se um estudo acerca dos dados do processo seletivo da Universidade Federal de Pelotas, realizado em dezembro de 2005, onde o escore bruto foi a variável tomada como a propriedade de interesse. Através do estudo, pôde-se comprovar a viabilidade da utilização da ferramenta para a descoberta de subgrupos, pois foi possível encontrar subgrupos com características de interesse.

Palavras-chave: Mineração de Dados. Descoberta de Subgrupos. Pós-processamento. Visualização. Regras de Associação.

ABSTRACT

The use of association rules discovery techniques to acquire knowledge from huge databases results in a large number of rules. Exploring the results obtained is a complex task and limits the use of Knowledge Discovery techniques. In order to overcome this shortcoming we develop a tool, which is written in the Java programming language and is based on an approach that suggests to combine association rules' and subgroup discovery's foundations. This approach uses post-processing of association rules' techniques in subgroups visualization. Thus, it is possible to use the large number of rules as a starting point to find interesting subgroups amid a population. Moreover, the tool engages the analyst in an interactive subgroup discovery process by means of a graphical interface, where he can also use his domain knowledge during the identification of such subgroups amid the population. In order to trial the tool, we show, at the end of this work, a study accomplished on the *Universidade Federal de Pelotas* students' admission procedure performed in December 2005. The candidates' punctuation was the variable used as the property of interest in the subgroup discovery process. By means of this study, we verified it was feasible to use the tool for subgroup discovery, since we had found subgroups containing interest features.

Keywords: Data Mining. Subgroup Discovery. Post-processing. Visualization. Association Rules.

LISTA DE FIGURAS

Figura 1 – Objetivos na utilização de KDD. 16

Figura 2 - Etapas do processo de KDD. 19

Figura 3 – Conjunto de regras de associação. 34

Figura 4 – Exemplo de taxonomia. 35

Figura 5 – Exemplo de um grafo direto. 37

Figura 6 – Exemplo de um gráfico representando matrizes 2D. 38

Figura 7 – Exemplo de um gráfico de *grids*. *39*

Figura 8 – Exemplo de um gráfico representando matrizes 3D. 40

Figura 9 - Ambiente Weka 42

Figura 10 - Gráfico obtido pelo PEAR 44

Figura 11 - Exemplo subgrupos em gráficos de setores. 48

Figura 12 - Um diagrama de caixas contendo subgrupos. 49

Figura 13 - Exemplo de visualização da Distribuição de um Atributo Contínuo 50

Figura 14 - Uma Curva ROC utilizada na descoberta de subgrupos. 51

Figura 15 - Um Gráfico de barras utilizado na descoberta de subgrupos 52

Figura 16 - Tela do *ORANGE*. 54

Figura 17 - Espaço em 2D representando os subgrupos de uma população. ... 56

Figura 18 - Representação gráfica de uma DR representando um subgrupo 57

Figura 19 - Diagrama de classe da ferramenta. 60

Figura 20 - Tela de escolha dos parâmetros de entrada para o CAREN. 61

Figura 21 - Diagrama de atividades da ferramenta 62

Figura 22 - Interface gráfica da ferramenta. 64

Figura 23 - Subgrupos de candidatos ao curso de medicina. 69

Figura 24 - Subgrupo S1. 70

Figura 25 - Subgrupo S2. 71

Figura 26 - Subgrupos de candidatos ao curso de medicina. 72

Figura 27 - Subgrupo S3. 72

Figura 28 - Subgrupo S4. 73

Figura 29 - Subgrupos compostos pela faixa etária dos candidatos. 74

Figura 30 - Subgrupo S6. 74

Figura 31 - Subgrupos das ciências agrárias. 75

Figura 32 - Distribuição de um subgrupo de candidatos das ciências agrárias 75

LISTA DE TABELAS

Tabela 1 - Exemplo de um conjunto de transações. 25

Tabela 2 - Possíveis regras a serem geradas a partir da transação 1. 25

Tabela 3 - Conjuntos candidatos de tamanho 1. 30

Tabela 4 - Conjuntos candidatos de tamanho 2. 30

Tabela 5 - Conjuntos candidatos de tamanho 3. 30

Tabela 6 - Regras candidatas 31

Tabela 7 - Exemplo de uma Tabela de Regras 37

Tabela 8 - Amostra hipotética. 53

LISTA DE ABREVIATURAS E SIGLAS

2D	Duas Dimensões
3D	Três Dimensões
CAPPS	*Computer-Assisted Passenger Prescreening System* (Sistema Assistido por Computador para pré-seleção de Passageiros)
CHD	Coronary Heart Disease (Doença Arterial Coronária)
CRISP-DM	Cross-Industry Standard Process for Data Mining (Padrão Inter-industrial para o Processamento de Tarefas de Mineração de Dados)
CSV	Comma Separated Values (Valores Separados por Vírgula)
FPr	False Positive Rate (Taxa de Falso Positivo)
GIS	Geographical Information System (Sistema de Informação Geográfica)
IBGE	Instituto Brasileiro de Geografia e Estatística
JVM	Java Virtual Machine (Máquina Virtual Java)
KDD	Knowledge Discovery in Databases (Descoberta de Conhecimento em Bancos de Dados)
OSJava	Open Sourced Java (Código Fonte Java Aberto)
PEAR	Post-Processing Environment for Association Rules (Ambiente de pós-processamento de regras de associação)
ROC	Receiver Operating Characteristics (Características de Operação do Receptor)
SGBDs	Sistema de Gerenciamento de Bancos de Dados
SQL	Structured Query Language (Linguagem estruturada da consultas)
TPr	True Positive Rate (Taxa de Verdadeiro Positivo)
UFPel	Universidade Federal de Pelotas
Weka	Waikato Environment for Knowledge Analysis (Ambiente Waikato para Análise de Conhecimento)

SUMÁRIO

RESUMO 5

ABSTRACT 6

LISTA DE FIGURAS 7

LISTA DE TABELAS 8

LISTA DE SIGLAS 9

1 INTRODUÇÃO 12
1.1 Motivação 12
1.2 Objetivos 13
1.2.1 Objetivos Específicos 13
1.3 Organização do Trabalho 14

2 DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS 15
2.1 Objetivos da Descoberta de Conhecimento em Bancos de Dados 16
2.2 Áreas de Aplicação 17
2.3 Etapas da Descoberta de Conhecimento em Bancos de Dados 18
2.3.1 Pré-Processamento 19
2.3.2. Seleção dos Dados 20
2.3.3 Transformação dos Dados 20
2.3.4 Mineração de Dados 20
2.3.4.1 Objetivos 21
2.3.4.2 Classes de Problemas de Mineração de Dados 21
2.3.5 Pós-Processamento 23
2.3.6 Representação do Conhecimento 23

3 REGRAS DE ASSOCIAÇÃO 24

3.1 Descoberta de regras de associação 26
3.1.1 Medidas de Interesse 26
3.1.1.1 Suporte 27
3.1.1.2 Confiança 27
3.1.1.3 Lift 28
3.1.1.4 Conviction 28
3.1.2 O algoritmo APRIORI 29
3.2 Pós-Processamento de Regras de Associação 31
3.2.1 Poda e Sumarização 32
3.2.2 Agrupamento 33
3.2.3 Utilização de Taxonomias 35
3.2.4 Auxílio através de Visualização 36
3.2.4.1 Tabela de Regras 36
3.2.4.2 Grafos Diretos 37
3.2.4.3 Matriz de Duas Dimensões 38
3.2.4.4 Matriz de Três Dimensões 39
3.2.5 Utilização de Técnicas de Gestão de Bancos de Dados 40
3.3 Ferramentas para Tratar Problemas da Associação 41

4 DESCOBERTA DE SUBGRUPOS 45
4.1 Conceituação 45
4.2 Objetivos 46
4.3 Visualização de Subgrupos 47
4.3.1 Gráfico em Setores 47
4.3.2 Diagrama de Caixas 48
4.3.3 Visualização da Distribuição de um Atributo Contínuo 49
4.3.4 Utilização da curva ROC 50
4.3.5 Gráfico de Barras 51
4.4 Regras de Distribuição 52
4.5 Ferramentas para a Descoberta de Subgrupos 53

5 A FERRAMENTA PROPOSTA 55
5.1 A Metodologia Utilizada 55
5.2 Aspectos gerais 57
5.3 Aspectos da Implementação 58
5.3 Utilização da Ferramenta 61

6 VALIDAÇÃO 66
6.1 O Conjunto de Dados Escolhido 66
6.2 Variáveis em Análise 67
6.3 Preparação para a Descoberta de Subgrupos 67
6.4 Resultados Obtidos 69
6.5 Considerações sobre a Utilização da Ferramenta 75

7 CONCLUSÕES 77

REFERÊNCIAS 79

1 INTRODUÇÃO

De acordo com Romão, Freitas e Pacheco (2000), a grande quantidade de informação existente nos bancos de dados informatizados de organizações pode esconder conhecimentos valiosos e úteis para a tomada de decisão, planejamento e gestão. Para os referidos autores o aumento no volume dos dados, associado à crescente demanda por conhecimento novo voltado para decisões estratégicas, tem provocado o interesse crescente em descobrir conhecimento em banco de dados. Dentro deste contexto surgiu a área de KDD (*Knowledge Discovery in Databases* – Descoberta de Conhecimento em Bancos de Dados), a qual permite a extração de informações que dificilmente seriam identificadas somente realizando consultas em um banco de dados.

Dentro do processo de KDD, tem-se a etapa de Mineração de Dados, a qual representa a essência de todo o processo. Segundo Fayyad, Piatetsky-shapiro e Smyth (1996) é nesta etapa que ocorre a aplicação de algoritmos específicos, que tenham uma limitação aceitável de eficiência computacional e que sejam capazes de produzir uma enumeração particular de padrões. Estes representam informação implícita sendo ela previamente conhecida ou não.

Uma das técnicas da Mineração de Dados é a descoberta de regras de associação, uma regra deste tipo possui um termo antecedente e um conseqüente, representando assim um padrão extraído. A estrutura de uma regra deste tipo será descrita com mais detalhes no capítulo 3.

1.1 Motivação

A utilização da técnica de descoberta de regras de associação para obter conhecimento a partir de bancos de dados, com grandes dimensões, resulta em uma quantidade muito elevada de regras. Tal fato torna bastante complexa a análise e compreensão das mesmas. Dentro deste contexto este trabalho baseia-se em uma

metodologia proposta por Pereira (2006) e Jorge (2006b), onde se propõe a descoberta de subgrupos por meio do pós-processamento de regras de associação. A metodologia vale-se de conceitos e técnicas tanto da descoberta de subgrupos, como também do pós-processamento de regras de associação por meio de técnicas de visualização.

Para se fazer valer de tal metodologia e usufruir as definições propostas na mesma, o desenvolvimento de uma ferramenta para implementação de tais definições seria de grande proveito. Além disso, a utilização de uma linguagem de programação, como Java, para a implementação permite que a ferramenta possa ser utilizada na Web e também garante sua portabilidade entre vários sistemas operacionais. Além disso, a linguagem possui um vasto aparato de bibliotecas de classes que viabilizam a implementação de uma ambiente gráfico interativo, no qual o usuário seja capaz de visualizar e navegar no espaço de subgrupos descobertos, para que ao final do processo, ele tenha disponível um conjunto de subgrupos que, até então, possuíam propriedades não conhecidas previamente.

1.2 Objetivos

O foco deste trabalho encontra-se em dois objetivos principais: o primeiro é o estudo de metodologias para descoberta de subgrupos e para realização do pós-processamento de regras de associação; o segundo é o desenvolvimento de uma ferramenta, na linguagem de programação Java, que implemente a metodologia proposta por Pereira (2006) e Jorge (2006b).

Para alcançar os objetivos principais do trabalho foram definidos os seguintes objetivos específicos:

1.2.1 Objetivos Específicos

- Fazer um estudo sobre conceitos e etapas do processo de KDD;
- Realizar uma abordagem acerca dos tipos de tarefas e técnicas a serem utilizadas na etapa de mineração de dados;
- Realizar uma abordagem acerca de conceitos e definições de regras de associação;
- Fazer um estudo sobre pós-processamento de regras de associação;

- Fazer um estudo sobre descoberta de subgrupos;
- Escolher um algoritmo que realize o pós-processamento de regras de associação para a descoberta de subgrupos e explicitar seu funcionamento;
- Especificar a ferramenta através do uso de diagramas definidos na área de Engenharia de Software;
- Implementar a ferramenta;
- Escolher um conjunto de dados para demonstrar a utilização e também validar a ferramenta.

1.3 Organização do Trabalho

No capítulo 2 faz-se uma abordagem acerca do processo de Descoberta de Conhecimento em Bancos de Dados, com ênfase especial nas etapas que compreendem tal processo. Dentre tais etapas, dá-se um destaque para a etapa de Mineração de Dados.

Posteriormente, no capítulo 3, são descritos conceitos relativos às Regras de Associação, as quais são utilizadas como técnica de solução da Associação, uma classe de problema de mineração de dados. Ainda no capítulo 3 são abordadas metodologias para a realização do pós-processamento das referidas regras.

No capítulo 4 é abordada a Descoberta de Subgrupos, uma metodologia especial para a realização da etapa de mineração de dados. Em tal abordagem, é dada uma maior ênfase às técnicas de visualização de subgrupos. Ao final do capítulo são apresentadas as Regras de Distribuição, as quais são o ponto de partida para o desenvolvimento da metodologia que este trabalho tomou como base.

No capítulo 5 é apresentada a ferramenta desenvolvida neste trabalho, onde, inicialmente, descreve-se a metodologia que foi utilizada para o desenvolvimento da mesma e, posteriormente, são descritos aspectos relativos à implementação.

No capítulo 6 é apresentado o conjunto de dados que foi escolhido para realizar a validação da ferramenta proposta. Em seguida, descreve-se o processo de análise que foi realizado utilizando tal ferramenta, assim como os resultados obtidos.

Por fim, são apresentadas as conclusões obtidas com o desenvolvimento deste trabalho, bem como sua contribuição e possíveis trabalhos futuros.

2 DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS

O termo Descoberta de Conhecimento em bancos de dados, ou simplesmente KDD, foi criado em 1989 referindo-se ao processo de descobrir conhecimento potencialmente útil existente em bancos de dados (PITONI, 2002).

Fayyad, Piatetsky-shapiro e Smyth (1996), definem KDD como sendo o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis, embutidos nos dados.

De acordo com Frawley, Piatetsky-shapiro e Matheus (1992), nos dias de hoje a quantidade de informação existente no mundo dobra a cada vinte meses. Aliado a este fato, tem-se também toda tecnológia existente que possibilita o armazenamento de grandes volumes de dados e com um custo cada vez menor. Prova disso está no caso mencionado em (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), no qual cita-se a rede americana Wal-Mart, a qual conta com uma base de dados em que são executadas 20 milhões de transações ao dia.

Outro exemplo de grande volume de armazenamento é o gerenciador de endereços eletrônicos *Gmail*, pertencente ao *Google*, no qual prega-se o princípio “não jogar nada fora”, onde cada usuário pode ter um espaço total de 2 gigabytes disponível.

Além da possibilidade de armazenamento, a tecnologia dos dias de hoje também possibilita a rápida difusão e publicação de informações.

Entretanto, toda essa facilidade proporcionada pela tecnologia traz como conseqüência ruim uma “explosão” de informação, onde tem-se um amontoado de dados sem utilidade nas quais torna-se bastante difícil encontrar informações úteis.

Atualmente, o sucesso e prestígio são medidos através da detenção de conhecimento, aquelas organizações que possuem a habilidade na manipulação de suas informações são os detentores da matéria prima essencial para o seu desenvolvimento (BRAUNER, 2003).

Sendo assim, torna-se eminente a necessidade de um meio de lidar com toda esta informação e prover uma automatização maior em sua análise, visto que sem a referida automatização, informações valiosas provavelmente passariam desapercebidas em meio a um enorme volume de dados. Além disso, em meio a forte pressão competitiva existente nos dias de hoje, torna-se necessário utilizar o conhecimento descoberto para prever futuras tendências e comportamentos para, desta forma, proporcionar maior apoio em processos de tomada de decisão.

2.1 Objetivos da Descoberta de Conhecimento em Bancos de Dados

Fayyad, Piatetsky-shapiro e Smyth (1996), argumentam que o processo de KDD pode ser utilizado para atingir dois tipos objetivos básicos: verificação ou descoberta.

Quando o objetivo é o da verificação, o processo de KDD tem como meta final verificar a veracidade de hipóteses definidas previamente, enquanto que na descoberta o processo busca encontrar padrões de forma automática ou semi-automática. Este último objetivo pode ainda ser subdividido em descrição e previsão. Segundo Fayyad, Piatetsky-shapiro e Smyth (1996), a descrição procura encontrar padrões, interpretáveis pelo homem, que descrevam os dados e a previsão parte de diversas variáveis para prever outras variáveis ou valores desconhecidos. A hierarquia dos tipos de objetivos existentes no processo de KDD é ilustrada na Fig. 1

Figura 1 – Objetivos na utilização de KDD.

A realização de KDD com o objetivo de verificação é trivial quando comparado com objetivo de descoberta, pois não é necessário aplicar todas as etapas do processo de KDD, dentre elas a de mineração de dados, etapa em que se

destina a extração de padrões. Sendo assim, o estudo deste objetivo está fora do escopo deste trabalho, pois o tema do trabalho está inserido no objetivo de descoberta.

2.2 Áreas de Aplicação

De acordo com Seifert (2005), áreas da indústria como seguradoras, bancos e vendas de varejo, comumente utilizam KDD para reduzir custos, aprimorar pesquisa e aumentar as vendas. No setor público, aplicações de KDD inicialmente eram utilizadas como meio para detectar fraudes e desperdícios, porém evoluíram para também serem utilizadas como meio de avaliação e melhoria de desempenho em processos (SEIFERT, 2005).

O processo de KDD pode ser utilizado em diversas áreas do conhecimento e aplicado para diferentes contextos, tais como: análise de marketing; finanças; indústria, educação; medicina; detecção de fraude; controle de produção; gestão de negócios; bioinformática; geoprocessamento; etc. A seguir serão descritas três situações específicas onde o KDD é aplicado.

A primeira situação foi retirada de (SANTOS, 2000) e consiste na utilização de KDD pelo *Secutity Pacific/Bank of América.* O banco utiliza KDD para dar suporte à decisão para a concessão empréstimos bancários. Basicamente, realiza-se uma análise para identificar um perfil de clientes que têm propensão a proporcionar risco nas operações financeiras.

A segunda situação é descrita em (SEIFERT, 2005), onde menciona-se a criação do programa CAPPS II (*Computer-Assisted Passenger Prescreening System II* – Sistema Assistido por Computador para pré-seleção de Passageiros) pelo Departamento de Defesa dos Estados Unidos (EUA). O programa foi criado em resposta direta aos ataques terroristas do 11 de setembro. O objetivo do programa é prover um sistema capaz de confirmar a identidade de passageiros e identificar terroristas, ou pessoas ligadas ao terrorismo, antes de embarcarem em aviões nos EUA. Para tanto, o KDD está sendo utilizado, acerca de dados pessoais dos passageiros, para disponibilizar ao sistema três perfis de passageiros: “verde”, “amarelo” e “vermelho”. No caso deste último, o passageiro seria impedido de embarcar. Estava previsto que tal sistema fosse testado em meados de 2004 em

alguns aeroportos, porém o programa encontrou alguns obstáculos para implementar e testar o sistema, sendo assim, o referido teste teve de ser adiado.

A última situação também foi retirada de (SEIFERT, 2005), onde menciona-se que o departamento de justiça dos EUA utiliza KDD para descobrir padrões de crimes e com isso obter auxílio para a tomada de decisão acerca da alocação de recursos para prevenir tais crimes.

2.3 Etapas da Descoberta de Conhecimento em Bancos de Dados

O processo de KDD é caracterizado como sendo um processo interativo e iterativo, composto por várias etapas interligadas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Entretanto existem diversas propostas para a divisão etapas do KDD, dentre as mais referenciadas na literatura encontram-se as abordagens propostas em (REZENDE et al., 2003), (CHAPMAN et al., 2000) e (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Na abordagem proposta em (REZENDE et al., 2003), o KDD é tratado como um ciclo composto por apenas três grandes etapas: Pré-Processamento, Extração de Padrões e Pós-Processamento. A etapa de pré-processamento é precedida por uma fase de identificação do problema, já a etapa de pós-processamento é sucedida por uma fase de utilização do conhecimento.

Chapman et al. (2000) definiram em sua abordagem o modelo CRISP-DM 1.0 (*Cross-Industry Standard Process for Data Mining).* Tal modelo foi desenvolvido para ser uma espécie de metodologia padrão para projetos de mineração de dados em bancos de dados empresariais, sendo voltada principalmente para profissionais de negócios. O modelo é composto por seis etapas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, validação de modelos e publicação. Cabe salientar que nesta abordagem o termo “Mineração de Dados” refere-se ao processo completo de KDD.

Por fim, a abordagem, exposta em (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), define o processo de KDD sendo composto, segundo adaptação feita em (HAN; KAMBER, 2001), pelas seguintes etapas: pré-processamento, seleção dos dados, transformação dos dados, mineração de dados, pós-processamento e representação do conhecimento.

A Fig. 2 ilustra a seqüência de aplicação das etapas citadas anteriormente.

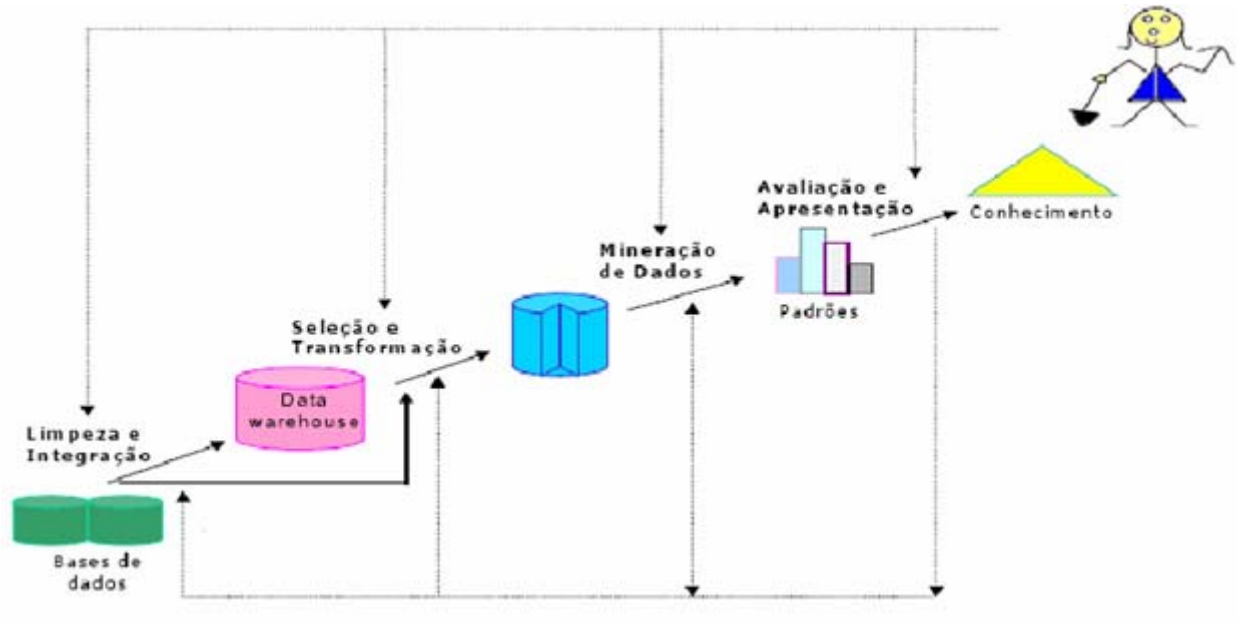

Figura 2 - Etapas do processo de KDD.
Fonte: Adaptação feita a partir de (BORGORI, 2003)

Visto que este trabalho toma como referência esta última metodologia, nas subseções a seguir serão descritas com maior detalhe todas etapas da referida metodologia.

2.3.1 Pré-Processamento

Nos dias de hoje, tem-se a tendência de definir uma única etapa para realizar a limpeza e integração dos dados (HAN; KAMBER, 2001).

A limpeza dos dados tem como objetivo eliminar dados inconsistentes e ruidosos, os quais ocorrem devido a erros originados na entrada dos dados. A realização da limpeza de dados de forma incorreta tende a comprometer as etapas posteriores do processo de KDD.

No caso de se ter várias fontes de dados, deve-se realizar a Integração dos Dados, a qual consiste na integração de várias fontes de dados em apenas um modelo, o qual, geralmente, constitui um *Data Warehouse*.

2.3.2. Seleção dos Dados

A etapa de Seleção dos Dados diz respeito a recuperação dos dados mais relevantes para a realização da análise. Tal seleção é feita de acordo com objetivos de análise traçados anteriormente. Para tal, faz-se necessário que o analista tenha conhecimento prévio acerca do domínio compreendido pelos dados, pois ele deve ser capaz de identificar os dados que podem determinar padrões de comportamento.

2.3.3 Transformação dos Dados

A Transformação dos Dados é a etapa que antecede a Mineração, portanto, é nesse ponto em que os dados devem ser colocados em um formato que possa servir de entrada para um algoritmo de mineração. Portanto, faz-se necessário definir, antes desta etapa, a técnica de mineração a ser aplicada.

Han e Kamber (2001) ressaltam que, em alguns casos, a etapa de transformação dos dados pode ser realizada antes mesmo da etapa de seleção dos dados, particularmente quando os dados são integrados em um *Data Warehouse*.

2.3.4 Mineração de Dados

A mineração de dados é a essência do KDD, pois todo o processo é realizado em função desta etapa. Dentro do KDD, inicialmente preparam-se os dados para aplicar uma técnica para a extração dos padrões e posteriormente os padrões descobertos são filtrados e preparados em um modo de apresentação mais legível.

O termo “Mineração de Dados” é tratado tanto como um sinônimo para todo o processo de KDD, como também uma etapa que compõe o mesmo. Neste trabalho, o referido termo é tratado como uma etapa dentro do processo de KDD, acordando com o que foi definido em Fayyad, Piatetsky-Shapiro e Smyth (1996). Além disso, esta é uma definição que se aproxima mais dos propósitos deste trabalho, pois necessita-se distinguir o processo de extração de padrões (mineração de dados) e o da seleção de padrões novos e úteis (etapa de pós-processamento).

Se a informação (os padrões extraídos) durante a etapa de mineração já era previamente conhecida, então no pós-processamento, estes padrões devem ser eliminados da análise, ficando apenas os padrões novos ou úteis. Tais padrões,

aliados a uma forma de visualização que os torne legíveis ao usuário final, constituem o referido “conhecimento” do termo KDD.

2.3.4.1 Objetivos

Antes de mencionar os objetivos existentes na etapa de mineração de dados, é importante versar sobre o conceito de “Padrão”. Um padrão, sob o ponto de vista do processo de KDD, pode ser definido como um evento, ou combinação de eventos que ocorrem freqüentemente em um banco de dados, onde cada evento é representado por um conjunto de dados.

O tipo de padrão que se deseja encontrar nesta etapa está diretamente relacionado com o tipo de conhecimento a ser descoberto no processo de KDD e também com o objetivo de sua posterior aplicação. Conforme o tipo de padrão desejado, têm-se objetivos diferentes na realização da etapa de mineração, onde, de acordo com Fayyad, Piatetsky-shapiro e Smyth (1996), os de mais alto nível são a predição e a descrição.

O objetivo da predição tem como meta estabelecer valores desconhecidos ou futuros de um ou mais atributos variáveis, tomando como base valores já conhecidos. Já o da descrição busca encontrar padrões através de dados já conhecidos e posteriormente descrevê-los de uma forma legível.

A importância desses dois objetivos de mineração varia bastante conforme as características e objetivos existentes na descoberta do conhecimento, porém, de acordo com Fayyad, Piatetsky-shapiro e Smyth (1996), no contexto de KDD, os padrões descritivos freqüentemente são mais importantes que os preditivos. Tal fato se dá porque estes são eficazes na busca de uma resposta para um problema específico e bem definido, o que não é o caso do contexto de KDD, onde possui-se apenas um conjunto de dados como ponto de partida do qual se quer extrair informações (PITONI, 2002).

2.3.4.2 Classes de Problemas de Mineração de Dados

De acordo com o objetivo de mais alto nível da etapa de mineração, têm-se diferentes classes de problemas de descoberta de conhecimento a serem resolvidos, os quais também podem ser definidos como tipos de tarefas a serem

desempenhadas, onde mais comumente tem-se a classificação, regressão, associação, agrupamento, seqüência, sumarização, dentre outras.

De acordo com a tarefa a ser desempenhada, e também com os dados disponíveis para análise, pode-se escolher uma ou mais técnicas para minerar dados. Dentre elas, pode-se citar: redes neurais, árvores de decisão, árvores de regressão, raciocínio baseado em casos, algoritmos genéticos, descoberta de regras de associação, dentre outras.

Cabe ressaltar que as técnicas utilizadas para extração de padrões disponíveis atualmente, embora sejam capazes de produzir resultados satisfatórios, ainda necessitam muito da interação humana no processo, pois ainda é inviável conceber de forma totalmente automática padrões considerados “valiosos”.

De acordo com a avaliação de Han e Kamber (2001), o agrupamento, a classificação e a associação estão entre as mais importantes classes de problemas de mineração de dados.

O agrupamento parte de um conjunto de dados heterogêneo e constrói grupos de dados que possuam características semelhantes. Tais grupos, mais comumente chamados de classes, não são conhecidos a priori e não possuem um critério bem definido para serem formados. Sendo assim, o agrupamento é considerado uma forma de aprendizagem não supervisionada. Por conseguinte, o agrupamento se enquadra no objetivo da descrição em relação ao processo de KDD, pois, nesse caso, o analista deseja obter classes que caracterizem o conjunto de dados em análise. O terceiro exemplo exposto na seção 2.2, no qual o departamento de justiça dos EUA deseja encontrar padrões de crimes, consiste de um exemplo de agrupamento.

O oposto ocorre na classificação, a qual é considerada uma forma de aprendizagem supervisionada, onde as referidas classes são definidas a priori. Sendo assim, faz-se necessário fornecer um conjunto de treino ao algoritmo utilizado, pois, dessa forma, ele consegue distinguir os dados conforme as classes definidas. O conjunto de treino nada mais é do que um conjunto de dados que define claramente as características da classe que se deseja utilizar. A resolução de problemas pertencentes à classificação se enquadra no objetivo da predição em relação ao processo de KDD, como é o caso do segundo exemplo exposto na seção 2.2, no qual o programa CAPPS II classifica passageiros em alguma das três classes mencionadas no exemplo.

Tanto no agrupamento, como na classificação, pode-se utilizar técnicas oriundas de diferentes áreas do conhecimento, porém tem-se predominância de técnicas da Inteligência Artificial, como redes neurais, algoritmos genéticos, raciocínio baseado em casos, dentre outras.

Por fim, tem-se a classe de problemas da associação, na qual busca-se encontrar itens que ocorram juntos em transações de um conjunto de dados. Dessa forma, compõem-se regras que indicam dependência entre itens do conjunto de dados. Sendo assim, a associação também se enquadra no objetivo da descrição. No capítulo 3 trata-se com mais detalhes esta classe de problemas.

2.3.5 Pós-Processamento

O Pós-processamento, também chamado de Avaliação, é a etapa na qual se realiza a validação dos resultados obtidos na etapa de mineração. São selecionados os padrões que atendem aos objetivos de análise traçados pelo analista. Para tanto, o analista pode valer-se de medidas de interesse, ou avaliação, para selecionar os padrões a serem apresentados como resultado do processo de KDD.

Nesta etapa é possível identificar se os resultados obtidos atingiram ou não os objetivos da análise, caso não tenham, pode-se retornar a qualquer uma das etapas do processo.

2.3.6 Representação do Conhecimento

A etapa de Representação do Conhecimento, ou etapa de Visualização, refere-se ao encerramento do processo de KDD, onde o conhecimento obtido durante o processo é apresentado de uma forma legível ao usuário. Para tanto, podem ser utilizadas técnicas de visualização, de representação do conhecimento, planilhas, relatórios, dentre outros.

3 REGRAS DE ASSOCIAÇÃO

Visto que a associação, uma das classes de problemas de mineração, tem como objetivo encontrar conjuntos de itens (ou eventos) que ocorram juntos em um determinado conjunto de dados, o caminho mais natural para representar tal associação entre atributos seria através de uma Regra de Associação. Uma regra de associação caracteriza o quanto a presença de um conjunto de itens nos registros de um conjunto de dados implica na presença de algum outro conjunto distinto de itens nos mesmos registros (AGRAWAL; IMIELINSKI; SWAMI, 1993). Conceitos inerentes a regras de associação serão apresentados a seguir, para que assim, posteriormente, seja feita uma abordagem à técnica de "descoberta de regras de associação", onde os referidos conceitos serão cruciais ao desenvolvimento desta técnica.

As regras de associação foram inicialmente introduzidas por Agrawal, Imielinski e Swami (1993), onde formalizou-se que: dado um conjunto de itens (também comumente chamados de atributos) I, onde I = $\{i_1, i_2, i_3, ..., i_{n-1}, i_n\}$. Cada elemento i pertencente a I pode assumir valores binários 0 ou 1 (falso ou verdadeiro), conforme esteja presente ou não. Tem-se também um conjunto de transações T acerca dos elementos de I, onde T = $\{t_1, t_2, t_3, ..., t_{n-1}, t_n\}$. Cada elemento t pertencente a T, corresponde a um conjunto de itens presentes em I, tal que $t \subseteq I$.

Ainda segundo a formalização apresentada em Agrawal, Imielinski e Swami (1993), considera-se que se todos os elementos pertencentes a um conjunto de itens A têm valor 1 em uma transação t, então o conjunto A está contido na transação t, ou seja, $A \subseteq t$. Uma regra de associação pode ser representada por uma expressão do tipo: $A \rightarrow B$, ou seja, a presença dos itens pertencentes a A na transação, implicam também na presença de itens pertencentes a B, onde $A \subseteq I$, $B \subseteq I$. Cabe salientar que os itens de A são diferentes dos itens de B, ou seja, $A \cap B =$

∅.

Agrawal, Imielinski e Swami (1993), tratam ainda a existência de variáveis quantitativas dentro de conjunto de itens, para tanto, criam-se intervalos de valores de tais variáveis e posteriormente utilizam-nos como valores binários.

Um exemplo prático do que foi exposto acima seria, por exemplo, a compra de produtos em um supermercado. Cada produto seria um item e sua compra seria uma transação. A tab. 1 ilustra um conjunto hipotético de transações (compras), onde cada uma é composta por conjuntos de itens .

Tabela 1 - Exemplo de um conjunto de transações.

Transação	Conjunto de itens
1	{queijo, refrigerante, manteiga}
2	{queijo, banana, refrigerante, manteiga}
3	{queijo, presunto, manteiga}
4	{banana, refrigerante}
5	{queijo, refrigerante, presunto}
6	{banana, refrigerante}
7	{queijo, banana, refrigerante, manteiga}
8	{manteiga, banana}
9	{queijo, banana, refrigerante}
10	{queijo, refrigerante, banana}

Através da transação 1, poderiam ser geradas as seguintes regras, as quais são expressas na tab. 2:

Tabela 2 - Possíveis regras a serem geradas a partir da transação 1.

Número	Regra
1	queijo → refrigerante, manteiga
2	queijo → refrigerante
3	queijo→ manteiga
4	Refrigerante → queijo, manteiga
5	Refrigerante → queijo
6	Refrigerante → manteiga
7	Manteiga → queijo, refrigerante
8	Manteiga → queijo
9	Manteiga → refrigerante
10	queijo, refrigerante → manteiga
11	Queijo, manteiga → refrigerante
12	Refrigerante, manteiga → queijo

É importante salientar que chamamos de antecedente de uma regra os itens presentes no lado esquerdo da implicação e de conseqüente os itens presentes no lado direito da implicação. Tomando-se a regra número 1, a mesma poderia ser interpretada da seguinte forma: se compra “queijo” então compra “refrigerante” e “manteiga”.

3.1 Descoberta de regras de associação

A motivação do estudo iniciado por Agrawal, Imielinski e Swami (1993), foi a necessidade de obtenção de conhecimento por parte de organizações da área de varejo (supermercados), onde esta procura específica de conhecimento foi designada por “*market basket analysis”* ou “análise da cesta de compras” (NEVES, 2003). O objetivo era fornecer suporte a decisão para o planejamento e disposição de produtos nas prateleiras de um supermercado, de forma que produtos geralmente adquiridos na mesma compra fossem vistos próximos entre si (HARRISON, 1998; GONÇALVES, 1999). Em um estudo posterior Agrawal e Srikant (1994), formularam um algoritmo, chamado de *APRIORI*, que destina-se a descoberta de regras de associação. Na subseção 3.1.2. será feita uma abordagem acerca do funcionamento de tal algoritmo.

A descoberta de regras de associação, segundo Domingues (2004), tem como objetivo encontrar tendências que possam ser utilizadas para entender e explorar padrões de comportamento dos dados. Porém, nem toda regra de associação configura um padrão nos dados. Uma regra representará um padrão se o algoritmo de descoberta de regras de associação gerar a regra, pois neste caso a regra respeitou alguns critérios definidos que condicionaram sua existência e desta forma possui um grau maior de confiabilidade. Tais critérios, ou medidas de interesse, são apresentados na subseção abaixo

3.1.1 Medidas de Interesse

Ao analisar-se a tab. 2., é fácil concluir que o número de regras geradas pode tornar-se excessivamente elevado, principalmente se o número de itens presentes nas transações também for elevado, fato este que acabaria inviabilizando uma análise eficaz e comprometeria todo processo de KDD. Sendo assim, também faz-se necessário estabelecer um meio de reduzir o número de regras geradas. Uma

solução para amenizar esta dificuldade é a definição de medidas de interesse, as quais definem restrições na geração do conjunto de regras.

Os algoritmos de descoberta de regras de associação se utilizam dessas medidas de interesse, como parâmetros de entrada, com o intuito de diminuir o número de regras geradas na saída do algoritmo. As medidas de interesse universalmente mais utilizadas nos algoritmos de descoberta de regras de associação são o suporte e a confiança (NEVES, 2003). Tais medidas são descritas nas subseções seguintes, assim como outras medidas que, apesar de não serem as mais importantes, também são utilizadas.

3.1.1.1 Suporte

O suporte é uma medida que avalia a freqüência com que os termos de uma regra se encontram nos dados, ou seja, o número de transações em que os itens presentes na regra aparecem ao mesmo tempo nos dados.

Tomando o exemplo exposto no início deste capítulo e observando a regra de número 3 (queijo → manteiga), pode-se concluir que o suporte desta regra é de 40% (ou 0,4), pois os itens “queijo” e “manteiga” aparecem quatro vezes juntos nas dez transações existentes.

3.1.1.2 Confiança

A medida de confiança se refere a um valor de correspondência entre os itens que compõem uma regra, ou seja, expressa o percentual de transações em que ocorrendo o antecedente, o conseqüente também ocorre. A confiança pode ser obtida da seguinte forma: suporte(A, B) / suporte(A).

Tomando o exemplo exposto na tab. 1 e se observamos a regra de número 3 (queijo → manteiga), pode-se concluir que a confiança desta regra seria 4/7 = 0,57 = 57%.

3.1.1.3 Lift

O *lift*, ou sustentação em português, é uma medida utilizada para avaliar o grau de dependência dos termos de uma regra. Supondo uma regra de associação A → B, o *lift* representa o quão freqüente tende ser "B" quando "A" ocorrer, ou vice-versa. A medida pode ser obtida através da seguinte forma:

$$\text{lift}(A \rightarrow B) = \text{suporte}(A, B) / (\text{suporte}(A).\text{suporte}(B))$$

A avaliação de uma regra pode ser realizada através da seguinte configuração:

- Se lift(A → B) = 1, então a ocorrência dos itens pertencentes a "B" independe da ocorrência dos itens de "A", e vice-versa.
- Se lift(A → B) > 1, então a ocorrência dos itens pertencentes a "B" influi na probabilidade da ocorrência dos itens de "A".
- Se lift(A → B) < 1, então a ocorrência dos itens pertencentes a "B" influi na probabilidade da não ocorrência dos itens de "A".

Analisando a regra de número 3 da tab. 1, tem-se que *lift* (refrigerante → presunto) = 1/(8.2), o que resulta em 0,062, ou seja, existe pouca possibilidade de refrigerante e presunto serem comprados juntos. Geralmente, define-se que regras com *lift* menor do que um sejam descartadas, como é o caso do exemplo citado.

3.1.1.4 Conviction

A medida *conviction*, ou convicção em português, avalia o quanto o antecedente influencia na ocorrência do conseqüente de uma regra de associação. Ao contrário da medida de *lift*, o *conviction* é uma medida unidirecional, ou seja, o resultado de *conviction*($A \rightarrow B$) será diferente de *conviction*($B \rightarrow A$). O *conviction* pode ser obtido da seguinte forma:

$$\textit{conviction}(A \rightarrow B) = \text{suporte}(A, \neg B) / (\text{suporte}(A).\ \text{suporte}(\neg B))$$

ou

$$\textit{conviction}(A \rightarrow B) = \text{suporte}(A).(N - \text{suporte}(B)) / (\text{suporte}(A) - \text{suporte}(A,B))$$

, sendo "N" o número total de transações.

O valor do *conviction* pode variar entre 0 e +∞, quanto mais alto for o valor, mais o termo conseqüente tenderá a ocorrer quando o antecedente da regra ocorrer. Se o valor for igual a 1, indica independência dos termos da regra.

Cabe salientar que um valor de *conviction* muito alto é um indício de que a regra em questão é pouco interessante. Tomando-se como exemplo as transações da tab. 1 e analisando-se uma regra de associação R = {presunto → queijo}, tem-se: suporte(presunto, queijo) = 2, suporte(presunto) = 2, suporte(queijo) = 7 e conseqüentemente conviction(R) = 2.(10-7)/(2-2) = 6/0 = +∞. A informação de que queijo provavelmente será comprado se presunto também o for, não é relevante para o analista, pois é um fato óbvio.

3.1.2 O algoritmo APRIORI

Encontram-se na literatura vários algoritmos de descoberta de regras de associação, tais como: Basic (MANNILA; TOIVONEN; VERKAMO, 1994), DHP (PARK; CHEN; YU, 1997), ECLAT (ZAKI; PARTHASARATHY; LI, 1999), FPGrowth (HAN; YIN, 2000) e DIC (BRIN; MOTWANI; ULLMAN, 1997). Porém, de acordo com Neves (2003), o algoritmo padrão atualmente mais utilizado é sem dúvida o APRIORI. Os conceitos empregados neste algoritmo estão presentes em quase todos os algoritmos utilizados atualmente, sendo que a maioria de tais algoritmos são especializações do APRIORI, tais como o AprioriTid (AGRAWAL; SRIKANT, 1994) e o AprioriHybrid (AGRAWAL; SRIKANT, 1994).

A primeira tarefa a ser realizada pelo algoritmo APRIORI é obtenção dos chamados conjuntos de itens freqüentes, ou seja, obter aqueles conjuntos em que todos seus itens respeitam a uma medida de suporte mínimo (SupMin). Devido à ampla utilização do algoritmo APRIORI, desde quando formalizou-se o problema da descoberta de regras de associação, assume-se que a tarefa de encontrar conjuntos de itens freqüente é uma tarefa padrão em algoritmos de descoberta de regras de associação.

Agrawal e Srikant (1994) apresentam uma importante propriedade quando propõem o algoritmo APRIORI: todo subconjunto de um conjunto de itens freqüentes também é um conjunto de itens freqüentes. Sendo assim, a execução do algoritmo começa com a obtenção de conjunto de itens freqüentes de tamanho 1 e posteriormente, os de tamanho 2 e assim por diante. Tomando o exemplo exposto

na tab.1, tem-se os seguintes conjuntos de itens de tamanho 1 (também chamados de conjuntos de itens candidatos a freqüentes) expressos na tab. 3.

Tabela 3 - Conjuntos candidatos de tamanho 1.

Conjunto de itens	Num. De transações	Suporte
{banana}	7	0,7
{manteiga}	5	0,5
{presunto}	2	0,2
{queijo}	7	0,7
{refrigerante}	8	0,8

Assumindo que SupMin seja 0,6, os conjuntos de itens freqüentes obtidos seriam os seguintes: {banana}, {queijo} e {refrigerante}. Partindo-se desses conjuntos, teria-se os conjuntos candidatos de tamanho 2 que estão expressos na tab.4, os quais constituem-se da combinação dos conjuntos de itens freqüentes de tamanho 1.

Tabela 4 – Conjuntos candidatos de tamanho 2.

Conjunto de itens	Num. de transações	Suporte
{banana, queijo}	4	0,4
{banana, refrigerante}	6	0,6
{queijo, refrigerante}	6	0,6

Os conjuntos de itens freqüentes obtidos seriam os seguintes: {banana, refrigerante} e {queijo, refrigerante}. Partindo-se desses conjuntos, ter-se-ia o conjunto candidato de tamanho 3, o qual é expresso na tab.5.

Tabela 5 – Conjuntos candidatos de tamanho 3.

Conjunto de itens	Num. de transações	Suporte
{banana, queijo, refrigerante}	4	0,4

Como agora não poderiam ser mais obtidos conjuntos de itens freqüentes de tamanho 4, a tarefa de obtenção de conjuntos de itens freqüentes se encerraria por aqui. As regras de associação que poderiam ser geradas (ou regras candidatas) a partir dos conjuntos de itens freqüentes obtidos estão expressas na tab. 6.

Tabela 6 - Regras candidatas

Conjunto de itens freqüentes	Regras
{banana, refrigerante}	banana => refrigerante refrigerante => banana
{queijo, refrigerante}	queijo => refrigerante refrigerante = >queijo

O próximo passo do algoritmo é o de verificar se as regras candidatas satisfazem a alguma ou a algumas medidas de interesse que foram citadas na subseção 3.1.1. As que passassem por esta espécie de “filtro” seriam o output do algoritmo. Cabe salientar que as versões inicias do APRIORI utilizam como filtro a medida de confiança, porém, em extensões do mesmo são utilizadas outras medidas também.

De acordo com Neves (2003), recomenda-se que o analista defina, como entrada do algoritmo, um suporte baixo e uma confiança alta, pois primeiro gera-se um número grande (mas não excessivo) de regras e depois verifica-se a coesão de seus itens através da confiança, descartando as que estiverem abaixo da medida estipulada. Uma regra de associação com confiança baixa não refletiria um padrão de comportamento e um suporte demasiadamente elevado tenderia a proporcionar a perda de possíveis padrões.

3.2 Pós-Processamento de Regras de Associação

Pode-se pensar que o processo de geração de regras de associação assim como a sua compreensão são tarefas simples. Entretanto, se este raciocínio é verdadeiro para pequenos conjuntos de regras, tal já não acontece quando se procura analisar um grande conjunto de regras de associação (NEVES, 2003).

Sendo assim, somente a utilização de medidas de interesse, em grandes conjuntos de transações, não é suficiente para obter um resultado que seja interpretável pelo analista. Para se ter um resultado interpretável é necessário ainda realizar um pós-processamento no conjunto de regras gerado, pois desta forma é possível tornar o espaço de regras reduzido e/ou passível de melhor interpretação. Além disso, mesmo que uma regra satisfaça a todas medidas de interesse, não significa que ela seja relevante para o analista, pois o padrão expressado por ela pode já ser conhecimento pelo analista.

Um outro ponto favorável à realização do pós-processamento de regras de associação é o impasse que o analista se encontra no momento de definir as medidas de interesse no input do algoritmo de descoberta das regras, pois caso ele estipule valores muito altos, pode ter a perda de regras relevantes, por outro lado, se estipular valores muito baixos gera-se um número excessivo de regras. Tal fato torna a definição das medidas de interesse um processo praticamente empírico. Sendo assim, propõe-se que sejam estipulados valores baixos para tais medidas, pois posteriormente as regras são filtradas e/ou apresentadas de uma forma mais organizada no pós-processamento das regras de associação.

O pós-processamento das Regras de Associação é tratado tanto como sendo parte da etapa de mineração de dados dentro do processo de KDD, e também como sendo parte da etapa de pós-processamento. Jorge (2004) e Domingues (2004) não consideram que a tarefa de mineração termina logo após as regras terem sido descobertas. Porém, outros autores, como é o caso de Melana (2004), consideram que as regras de associação são pós-processadas em uma etapa posterior a mineração. Neste trabalho será considerado que o pós-processamento das regras de associação ainda pertence à etapa de mineração de dados, visto que a realização do pós-processamento das regras é incorporada muitas vezes dentro do algoritmo de descoberta de regras.

Nas subseções a seguir são apresentadas algumas das principais metodologias para a realização de pós-processamento em regras de associação, onde cada uma dessas metodologias pode ser usada em separado ou em conjunto com outras metodologias em uma forma híbrida.

3.2.1 Poda e Sumarização

A poda, ou “pruning”, constitui-se da definição de um critério para diminuir o conjunto de regras que for gerado. O estudo desta metodologia foi iniciado por Toivonen et al. (1995), o qual utilizou-se de um princípio, chamado de *cobertura de uma regra*, para realizar uma espécie de filtro nas regras geradas, objetivando eliminar (ou podar) regras que não tragam informação adicional.

O princípio da cobertura de regras de associação consiste na identificação de regras que expressem a mesma informação, onde a regra (ou as regras) que for mais específica (ou seja, com maior número de itens) deve ser podada caso possua

um valor de confiança igual ou menor que a regra menos específica. Um exemplo da utilização deste princípio poderia ser aplicado nas seguintes regras hipotéticas:

R1) queijo, presunto, manteiga → refrigerante; sup=2, conf=6

R2) queijo, presunto → refrigerante; sup=3, conf=8

Segundo este princípio, R1 poderia ser eliminada, pois ela é mais específica (ou menos geral) que R2 e possui confiança menor. Portanto, diz-se que R1 é redundante e é "coberta" por R2. Toivonen et al. (1995), propôs ainda a ordenação das regras, conforme seu grau de interesse, e em seguida a realização do agrupamento das mesmas, metodologia a qual será abordada na próxima subseção.

Um outro estudo relevante versando a técnica de poda foi o realizado por Liu, Hsu e Ma (1999). Neste estudo os autores assumem o conceito de sumarização (neste caso também chamado de resumo), onde diz-se que uma regra mais geral sumariza uma ou mais regras menos gerais, como é o caso exemplo exposto acima. Desta forma, busca-se encontrar um conjunto especial de regras (as mais gerais) que sumarizam, ou resumem, todas as demais.

A maioria dos autores, como é o caso de Toivonen et al. (1995) e Liu, Hsu e Ma (1999), assumem as regras mais gerais como as mais relevantes, entretanto, ainda que poucos, existem autores que consideram as regras menos gerais como as mais relevantes, como é o caso de Pitoni (2002).

Atualmente existem inúmeras ferramentas que incorporam esta metodologia, várias delas inclusive incorporam a técnica de poda no processo de geração das regras de associação.

3.2.2 Agrupamento

Realizar agrupamento em regras de associação foi primeiramente proposto por Toivonen (1995). O autor utilizou-se desta metodologia como um aprimoramento da metodologia citada na subseção anterior, porém o estudo foi apenas inicial e não foi aprofundado.

Entretanto, Jorge (2004) propôs uma metodologia na qual realiza-se agrupamento temático das regras de associação geradas, ou seja, objetiva-se encontrar grupos de regras que correspondam a diferentes temas dentro de um determinado domínio. Para tanto, as regras são agrupadas conforme o conteúdo dos itens presentes em seus termos.

Diferentemente das transações apresentadas na tab.1, geralmente transações a serem mineradas possuem diferentes temas representados por seus itens. No exemplo retirado de (JORGE, 2004) e exposto na Fig. 3, apresenta-se um pequeno conjunto de regras geradas a partir de um conjunto de transações.

```
r1 :  customs & imports -> taxes
r2 :  customs -> taxes
r3 :  imports -> taxes
r4 :  customs & imports & tax payer -> taxes
r5 :  customs & tax payer -> taxes
r6 :  taxes & imports -> customs
r7 :  taxes -> customs
r8 :  crisis -> taxes
r9 :  crisis & government -> taxes
r10:  taxes -> crisis
r11:  management -> books
r12:  books -> management
r13:  management & success -> books
r14:  Iraq & USA -> Europe
r15:  Iraq -> USA
r16:  USA -> Iraq
```

Figura 3 – Conjunto de regras de associação.
Fonte: JORGE, 2004.

Nas regras exibidas na figura acima percebe-se claramente que podem ser encontrados três grupos de regras: o primeiro grupo sendo composto pelas dez primeiras regras; o segundo pelas regras r11, r12 e r13; o terceiro composto pelas regras r14, r15 e 16.

A identificação dos temas presentes nas regras, entretanto, nem sempre é uma tarefa trivial como a do exemplo exposto anteriormente. Tal identificação varia conforme um determinado contexto, onde o analista define critérios no número de grupos de regras a se obter.

Jorge (2004) propõe ainda a utilização da técnica de sumarização para encontrar uma regra que sumarize todas as regras de um grupo e assim, conseqüentemente, facilitar a interpretação do analista.

A metodologia exposta nesta subseção ainda encontra poucas implementações desenvolvidas, dentre elas pode-se citar a ferramenta *Caren*

(AZEVEDO, 2003), que em uma versão mais atual possui a funcionalidade de gerar regras de associação de acordo com a metodologia exposta por Jorge (2004).

3.2.3 Utilização de Taxonomias

Uma taxonomia permite realizar uma classificação hierárquica de itens por meio de uma caracterização coletiva ou individual (DOMINGUES, 2004). A utilização de taxonomias em regras de associação foi proposta por SirKank e Agrawal (1997), onde se propõe utilizar taxonomias para generalizar regras, ou seja, transformar regras específicas em conceitos mais gerais. A aplicação de taxonomias pode ser utilizada para reduzir o volume de regras extraídas e, por conseqüência, facilitar a análise e compreensão do conhecimento (DOMINGUES, 2004).

Um exemplo de taxonomia utilizada neste contexto pode ser observado através da árvore presente na Fig.4, onde as folhas representam itens de transações e os outros nodos representam generalizações formadas a partir dos itens.

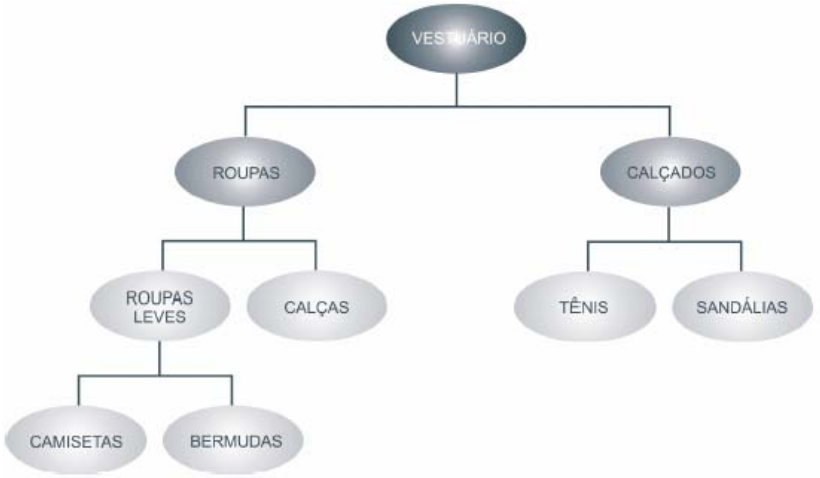

Figura 4 – Exemplo de taxonomia.
Fonte: DOMINGUES, 2004.

Através do exemplo, observa-se que os itens “camisetas” e “bermudas” podem compor uma classe chamada “roupas leves”. Ainda em relação ao exemplo, supondo as seguintes regras de associação: camisetas → tênis, camisetas → sandálias, bermudas → tênis e bermudas → sandálias. Tais regras podem ser generalização de uma única regra do tipo: roupas leves → calçados.

Entretanto, especificar uma taxonomia não é uma tarefa trivial, pois para um único domínio podem existir várias taxonomias, cada uma refletindo um ponto de vista do domínio. Além disso, em alguns casos não é possível especificar a classe de um item na hierarquia, constituindo assim um item dito sem classificação.

De acordo com Srikant (2001), dentre os principais argumentos para utilizar esta metodologia, está a obtenção de regras de simples interpretação, generalização de regras e maior facilidade de encontrar regras interessantes, podendo assim as não interessantes serem eliminadas através de um critério especificado em uma ou mais medidas.

Já é possível encontrar na literatura vários autores que implementam esta metodologia, tais como: *Cumulate* (SIRKANK; AGRAWAL, 1997), *Stratify* (SIRKANK; AGRAWAL, 1997), *Gart* (DOMINGUES, 2004) e *Genex* (WEBER, 1998).

3.2.4 Auxílio através de Visualização

Visualização é o processo de transformação de dados, informações e conhecimento em uma forma visual, de forma que se faça uso da capacidade humana de visualização (GERSHON; EICK, 1998).

A aplicação de técnicas de visualização em regras de associação é uma metodologia utilizada em processos de KDD que também se encontra na literatura, onde propõe-se, basicamente, aa exibição das regras em um formato que o analista possa visualizar todo ou parte do espaço das regras geradas.

Nas subseções a seguir são explanadas os principais métodos encontrados na literatura para a visualização de regras de associação.

3.2.4.1 Tabela de Regras

A chamada “Tabela de Regras” foi um dos primeiros esforços realizados na tentativa de exibir ao analista o espaço de regras de uma forma interpretável visualmente. Segundo Zhang (2000), é o método mais simples e direto para visualizar regras de associação.

Na tab. 7 encontra-se um exemplo de uma tabela de regras obtida de um pequeno conjunto de regras hipotéticas. Em cada linha da tabela se encontra uma regra distinta e nas colunas denotam-se suas propriedades.

Tabela 7 - Exemplo de uma Tabela de Regras

Item1	Item2	Item3	Item4	NullM	RuleM	Confiança	Suporte
Queijo	Banana			2	1	90%	10%
Refrigerante	Queijo	Banana		3	2	85%	7%
Banana	Presunto	Refrigerante	Manteiga	4	2	80%	5%
Manteiga	Queijo	Banana		3	1	60%	3%

As quatro primeiras colunas referem-se aos termos de cada regra, a coluna “NullM” indica quantas dessas colunas não são nulas e a coluna “RuleM” indica quantos itens estão presentes no termo antecedente (lado esquerdo) de cada regra.

Tomando como base a terceira linha da tab. 7, tem-se a seguinte regra:

banana & presunto → refrigerante & manteiga

3.2.4.2 Grafos Diretos

Segundo Wong, Whitney e Thomas (1999), este método é, juntamente com o método da Matriz de 2 Dimensões, o mais utilizado para visualizar regras de associação. O método consiste de exibir as regras em grafos diretos, onde os nodos representam os itens das regras e as arestas representam as associações.

A Fig.5 exibe um grafo direto que denota a seguinte regra: A → B & C.

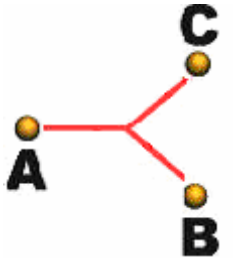

Figura 5 – Exemplo de um grafo direto.
Fonte: WONG; WHITNEY; THOMAS, 1999

As medidas de suporte e confiança podem ser identificadas através da utilização de diferentes cores e/ou larguras nas arestas. Porém, se temos um conjunto muito grande de regras, torna-se difícil analisar as regras, pois haverá um número excessivo de arestas entre os nodos.

Com a finalidade de contornar tal problema, Hetzler et al. (1998), sugere que algumas arestas do grafo fossem animadas, pois assim seria possível visualizar mais claramente as associações de certos itens. Entretanto, esta técnica requer constante interação humana, pois é necessário selecionar cada um dos itens em que se deseja visualizar suas associações.

3.2.4.3 Matriz de Duas Dimensões

Este método possibilita a visualização de todo ou parte do espaço de regras em um gráfico, onde cada uma das regras de associação é exibida através de uma matriz de duas dimensões.

Neste método, uma regra de associação é representada através de uma barra presente em uma coordenada do gráfico, onde um dos eixos representa o termo antecedente e o outro representa o termo conseqüente. As medidas de suporte e confiança podem ser representas através da altura e cores das barras presentes no gráfico. A Fig.6 ilustra tal sistema de representação, na qual tem-se duas regras de associação, A → C e B → C, sendo que a última possui maior suporte.

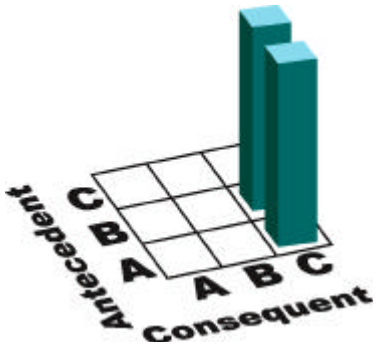

Figura 6 – Exemplo de um gráfico representando matrizes 2D.
Fonte: WONG; WHITNEY; THOMAS, 1999.

Uma outra possibilidade de exibição da matriz de duas dimensões é através de um gráfico de *grids*. Porém, nesta representação os eixos cartesianos expressam medidas relativas ao conjunto de regras analisado e em cada ponto do gráfico encontram-se uma ou mais regras de associação. Na Fig.7 tem-se um exemplo desse tipo de gráfico que foi retirado da ferramenta CrystalClear (ONG et al., 2002).

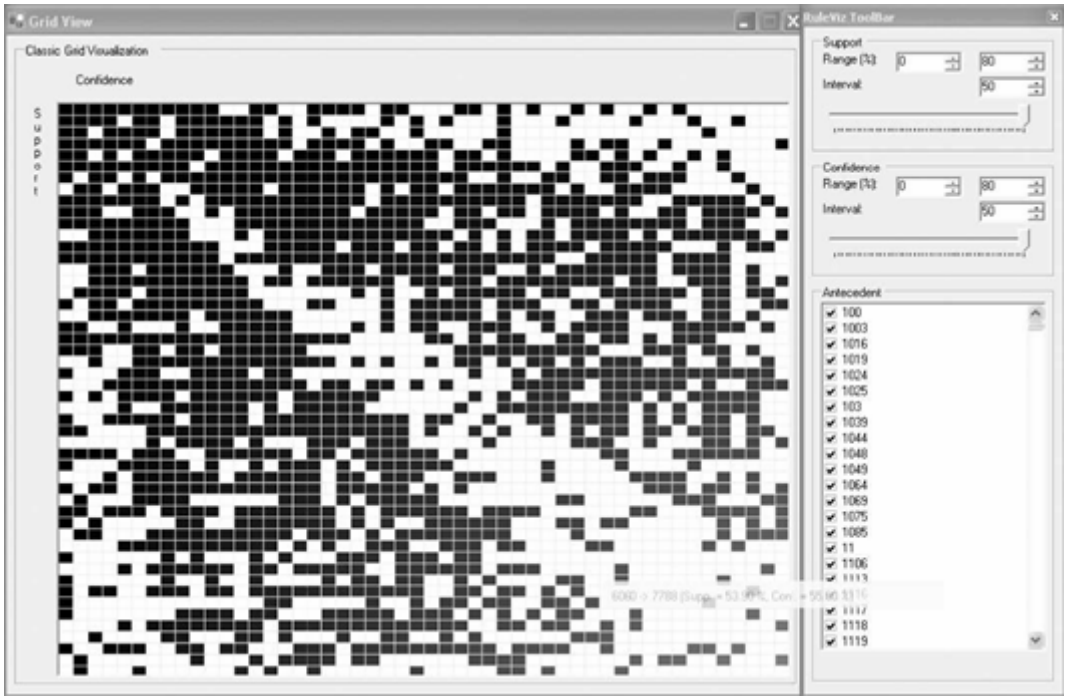

Figura 7 – Exemplo de um gráfico de *grids*.
Fonte: ONG et al., 2002.

Em tal ferramenta, o eixo das abscissas reflete a medida confiança e o das ordenadas reflete a medida de suporte relativa ao conjunto de regras analisado. Em cada ponto do gráfico pode haver várias regras de associação, pois algumas destas podem ter as mesmas medidas de suporte e confiança. Sendo assim, a ferramenta permite que, através de um clique do mouse em um ponto do gráfico, o analista tenha a descrição das regras que ocupam um determinado ponto.

3.2.4.4 Matriz de Três Dimensões

Uma outra forma de representação visual de um conjunto de regras de associação foi proposta por Wong, Whitney e Thomas (1999). Em tal representação, ao invés de se ter um gráfico contendo itens de regras nos eixos, tem-se um gráfico expressando as próprias regras em um eixo e no outro os itens que compõem as regras. Além disso, as medidas de suporte e confiança são representadas no mesmo gráfico através da altura de barras referentes a cada regra. Tal representação é ilustrada na Fig.8.

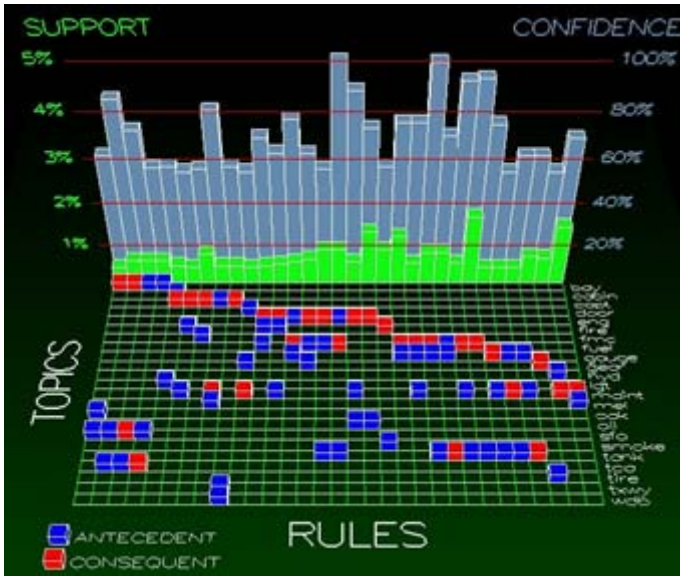

Figura 8 – Exemplo de um gráfico representando matrizes 3D.

Fonte: ONG et al., 2002

Através da figura observa-se que a diferenciação dos termos das regras é realizada através da utilização da cor azul, para identificar os termos antecedentes, e vermelho, para os termos conseqüentes. Além disso, utiliza-se a cor azul claro para identificar as barras referentes às medidas de confiança (barras mais ao fundo) e verde para as referentes as medidas de suporte (barras mais à frente). Tomando-se a regra presente mais à esquerda do gráfico, tem-se "*sfo* & *oak* → boy" com suporte de 0,5% e confiança de 61%.

3.2.5 Utilização de Técnicas de Gestão de Bancos de Dados

A popularidade que os sistemas de gestão de bancos de dados têm vindo a assumir, motivada, em grande parte, pela utilização da linguagem *Structured Query Language* (SQL), levou a que diversos pesquisadores procurassem tirar partido das capacidades destes sistemas na mineração de dados (NEVES, 2003).

A utilização de recursos de banco de dados para a mineração de dados, em especial para a descoberta de regras de associação, compreende uma estrutura de armazenamento de transações, ou mesmo de regras de associação, e também de uma linguagem para manipular as estruturas armazenadas. Tal linguagem quase sempre é baseada na linguagem SQL.

Uma das vantagens apontadas pelos idealizadores dessas técnicas é que se pode fazer uso o conhecimento do analista para direcionar o descobrimento das regras através das consultas.

Um dos primeiros estudos desenvolvido neste sentido foi o de Meo, Psaila e Ceri (1996), no qual foi desenvolvido o MINE RULE, um operador, baseado na linguagem SQL, destinado a descoberta de regras de associação a partir de transações. Neste estudo, partindo-se de transações armazenadas, o MINE RULE possibilita extrair regras de associação através de condições especificadas pelo analista. Tais condições podem ser relativas a medidas de interesse, neste caso, suporte e confiança. Além disso, tem-se a possibilidade de utilizar uma cláusula (semelhante ao “GROUP BY” do SQL) que permite agrupar as regras geradas.

Outros dois estudos relevantes realizados foram o de Han et al. (1996) e o de Morzy e Zakrzewicz (1997), onde foram idealizadas duas linguagens baseadas também em SQL. No primeiro, idealizou-se a DMQL, uma linguagem que também permite extrair regras de associação, porém é bem mais ampla que o MINE RULE, pois permite extrair também outros tipos de regras utilizadas em outras classes de problemas de mineração de dados. O segundo estudo compreendeu o desenvolvimento da linguagem MineSQL, a qual permite a extração de regras de associação dentro de um processo iterativo e interativo, no qual o analista pode especificar parâmetros na consulta para gerar as regras e, posteriormente, analisar o resultado obtido e se necessário modificar um ou mais parâmetros para obter um novo conjunto de regras.

Também com o propósito de otimizar o processo de extração de regras de associação a partir de consultas a bancos de dados, Hipp et. al. (2002) desenvolveram um estudo, o qual descreve uma metodologia, designada por *Rule Cache*, para o armazenamento de regras de associação em um banco de dados relacional.

3.3 Ferramentas para Tratar Problemas da Associação

Vários sistemas para a realização de KDD vêm sendo desenvolvidos nos últimos anos, os quais utilizam metodologias oriundas da área de estatística, banco de dados, aprendizado de máquina e visualização (KLÖSGEN, 1999). Em Kddnuggets (2006), cita-se softwares tanto no âmbito acadêmico, como da indústria.

Existem sistemas tanto no âmbito acadêmico, quanto na indústria. Dentre estes últimos pode-se destacar o *Oracle 9i,* o qual constitui-se de um poderoso SGBD que também fornece suporte a decisão através de KDD. Neste sistema é dada maior ênfase para as etapas anteriores a mineração de dados, porém só é possibilitado ao analista trabalhar com as classes de problemas da associação e classificação.

Quanto aos softwares acadêmicos, pode-se citar o *Weka* (*Waikato Environment for Knowledge Analysis*) (WEKA, 2006) como um dos mais populares. O *Weka* é um software livre para a realização de KDD, onde também é fornecida uma ampla biblioteca de classes em Java relativas a vários algoritmos relativos a diversas classes de problemas de mineração. Na Fig. 9 é exibida uma tela retirada de tal sistema.

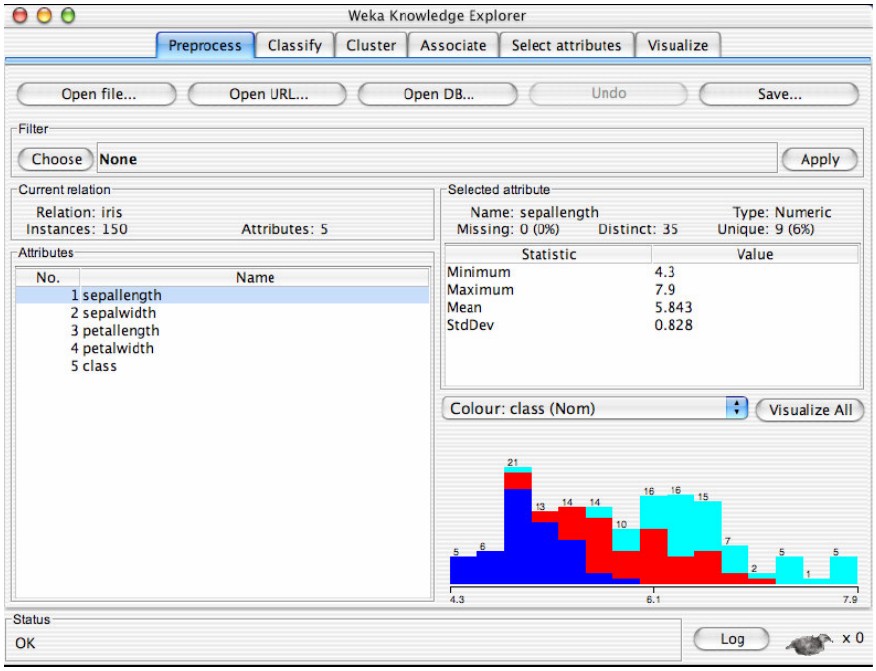

Figura 9 - Ambiente Weka.

Existem alguns sistemas de KDD, no âmbito comercial, que permitem a realização de pós-processamento de regras de associação utilizando-se recursos de visualização, dentre estes destacam-se o IBM Intelligent Miner, MineSet e IBM QUEST. O primeiro utiliza grafos diretos, o segundo utiliza matriz de duas dimensões e o terceiro também utiliza matriz de duas dimensões, porém através de gráfico de grids.

Apesar de existir uma grande variedade de sistemas de KDD disponíveis na atualidade, os quais integram, em um único sistema, várias metodologias encontradas na literatura, poucos destes apresentam implementações de metodologias eficientes para a realização do pós-processamento de regras de associação.

Zhang (2000) argumenta que, apesar do empreendimento de esforços na área de regras de associação, a grande maioria dos sistemas comerciais de KDD apresentam diversas limitações em implementações relativas a tal área.

Os que apresentam maior eficiência são oriundos, em sua maioria, do âmbito acadêmico e não costumam ter suporte a todas etapas do KDD, pois seu objetivo maior é a eficiência de uma metodologia especifica. Somente após ter-se a metodologia consolidada, parti-se então para um passo posterior em que se inclui a implementação da metodologia em um sistema mais amplo de KDD.

Quanto a implementações de pós-processamento de regras de associação utilizando-se recursos de visualização, no âmbito acadêmico, podem ser citadas as seguintes implementações: AViz (HAN; CERCONE; HU, 2002) e PEAR (NEVES, 2003).

O AViz é um ambiente interativo de visualização, em três dimensões, que proporciona a descoberta de regras de associação. O sistema fornece acompanhamento ao analista em todas etapas do KDD, porém todo processo é direcionado para a descoberta de regras de associação, as quais são encontradas somente através de variáveis numéricas.

Já o PEAR (Post-Processing Environment for Association Rules) é um ambiente Web direcionado para a navegação de regras de associação. O sistema toma como entrada um arquivo no formato PMML contendo as regras e, posteriormente, realiza o pós-processamento das regras utilizando técnicas de agrupamento de regras. Por fim, o sistema utiliza o método da matriz de duas dimensões para realizar um segundo pós-processamento das regras, no qual, exibi-se ao analista alguns gráficos contendo a descrição das regras obtidas. Na Fig. 10 é exibido um gráfico obtido pelo PEAR, no qual são representadas as medidas de suporte e confiança de um determinado conjunto de regras de associação.

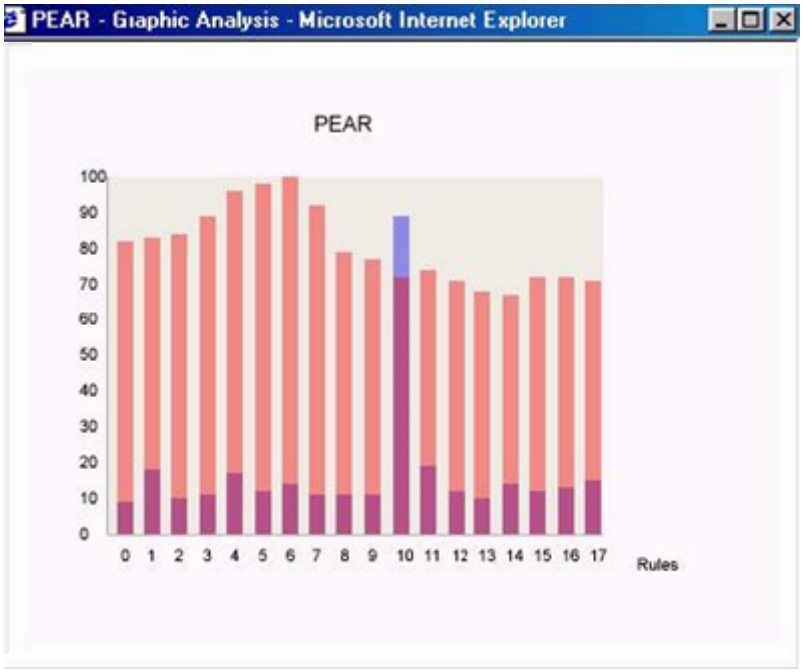

Figura 10 - Gráfico obtido pelo PEAR.
Fonte: NEVES, 2003.

4 DESCOBERTA DE SUBGRUPOS

Ao longo deste capítulo é abordada uma metodologia especial para a realização da etapa de mineração de dados. Tal metodologia, intitulada Descoberta de Subgrupos, foi proposta inicialmente por Klösgen (1994), na qual direciona-se o processo de KDD para encontrar padrões na forma de subgrupos, os quais são conceituados na seção abaixo.

4.1 Conceituação

Em seu estudo, Klösgen (1994) conceitua os chamados subgrupos, os quais correspondem a padrões encontrados nos dados que correspondem a subgrupos de indivíduos de uma população com alguma característica inesperada, discrepante ou atípica.

Esses padrões não expressam um comportamento existente em todo volume de dados, mas sim em subconjuntos específicos do mesmo. Mesmo assim, os subgrupos descobertos podem ter grande relevância para o analista, pois o objetivo da análise pode ser bastante específico ou os dados disponíveis podem não permitir a realização de uma análise confiável com relação a toda população (KLÖSGEN, 1999).

A característica inesperada, discrepante ou atípica, mencionada anteriormente, pode se referir a uma variável específica do conjunto de dados, sendo esta chamada de *Propriedade de Interesse*. Esta propriedade é a base para a descoberta de subgrupos, pois, neste processo, leva-se em conta relações existentes entre a propriedade de interesse, tomada como uma variável dependente, com o resto das variáveis, tomadas como variáveis independentes, as quais correspondem a descrição dos subgrupos. A avaliação acerca de tais relações é realizada através de alguma medida de interesse definida pelo analista ou pela metodologia em questão.

De acordo com Klösgen (1999) para um subgrupo ser útil ou relevante, ele necessita satisfazer, essencialmente, duas condições. A primeira condiciona que o subgrupo descoberto deve ser interpretável pelo analista. A segunda condiciona que ele deve satisfazer algum critério de avaliação a respeito do interesse do mesmo. Esta avaliação pode ser realizada utilizando-se tanto uma medida estatística apropriada, como também pode ser baseada no conhecimento do analista sobre do domínio em questão (JORGE, 2006b).

No atual estado da arte, segundo Kavsek e Lavrac (2004), os algoritmos mais comumente empregados são o *APRIORI-SD* (KAVSEK; LAVRAC; JOVANOSKI, 2003), o algoritmo *CN2-SD* (LAVRAC et al., 2004) e o *SubgroupMiner* (KLÖSGEN; MAY, 2002). Os dois primeiros acrescentam o sufixo SD (oriundo do termo *Subgroup Discovery*), pois ambos são algoritmos adaptados para a descoberta de subgrupos e que foram inicialmente idealizados para o aprendizado de regras de associação e classificação, respectivamente. Já o *SubgroupMiner* é um algoritmo voltado à visualização de subgrupos, no qual permite-se identificar subgrupos de maior tamanho, porém tão precisos.

4.2 Objetivos

Visto que a descoberta de subgrupos é uma metodologia para se realizar mineração de dados, seus objetivos de mais alto nível são os mesmos: predição e descrição. Tais objetivos condicionam o tipo de padrão a ser encontrado que, de acordo com Klösgen (1999), pode pertencer a um dos seguintes tipos: associações entre dois subgrupos, subgrupos seqüenciais e subgrupos discrepantes.

O primeiro tipo, associação entre dois grupos, busca identificar dois subgrupos que possuam alguma associação relevante em relação à propriedade de interesse.

O segundo tipo de padrão, subgrupos seqüenciais, identifica subgrupos que apresentam uma ordenação parcial, a qual satisfaz alguma restrição definida e, geralmente, é relativa ao tempo (KLÖSGEN, 1999).

Por fim, o tipo de padrão grupos discrepantes se refere a uma forma qualquer de desvio no valor de uma variável (propriedade de interesse) em relação ao resto da população.Tal variável pode ser binária, nominal ou numérica, fato que é determinante na escolha do tipo de técnica a ser utilizada para encontrar o referido

desvio. Jorge (2006b) menciona que encontrar subgrupos dessa classe pode ser útil para a identificação de desvios sistemáticos e para a compreensão das causas de tal desvio. E, ainda, cita um exemplo da área médica, no qual são encontrados subgrupos da população que apresentam um nível de colesterol discrepante em relação a toda população. Encontrar as causas destes desvios, através da descrição destes subgrupos, pode ajudar a prever futuras ocorrências dos mesmos.

4.3 Visualização de Subgrupos

Conforme citado na seção anterior, para o analista identificar um subgrupo como relevante, ele deve ser capaz de reconhecer e interpretar as características do subgrupo. Para tanto, faz-se necessário encontrar meios de descrever subgrupos ao analista. Dentre estes, pode-se destacar a *Visualização de Subgrupos* como um dos mais utilizados.

Na Visualização de Subgrupos, a principal finalidade e desafio é permitir que o analista consiga identificar subgrupos com características diferenciadas em relação ao resto da população. Para tanto, proporciona-se ao analista a possibilidade de realizar uma comparação da distribuição de um subgrupo com outro e, principalmente, com a distribuição de toda população. Além disso, os subgrupos são dispostos visualmente conforme o valor de sua propriedade de interesse, a qual, de acordo com Pereira (2006) , necessita ser uma variável categórica ou discretizada.

Segundo Kralj et al. (2005), as principais técnicas de visualização de subgrupos disponíveis atualmente são as seguintes: gráfico em setores, diagrama de caixas, visualização da distribuição de um atributo contínuo, utilização da curva ROC e gráfico de barras. Tais técnicas serão descritas nas próximas subseções.

4.3.1 Gráfico em Setores

Nesta técnica, cada subgrupo é mostrado em um gráfico de setores composto por dois níveis. O mais alto corresponde à distribuição da propriedade de interesse em relação a toda população, enquanto que o mais baixo corresponde a distribuição do subgrupo em questão.

Cada setor do gráfico representa uma classe referente à propriedade de interesse, a qual pode representar tanto um valor categórico como também um valor numérico que divide os subgrupos em duas classes. Um exemplo desta divisão poderia ser ilustrado através de uma dada propriedade de interesse referente ao nível de açúcar no sangue de pacientes, onde certo valor poderia dividir os pacientes em duas classes: a dos pacientes com diabetes e a dos pacientes saudáveis. O raio de cada setor representa o tamanho do subgrupo, ou seja, o número de amostras que o compõe. Na Fig. 11 tem-se uma ilustração de seis gráficos de setores, onde o primeiro representa toda população e os outros cinco representam, respectivamente, os subgrupos A1, A2, B1, B2 e C1.

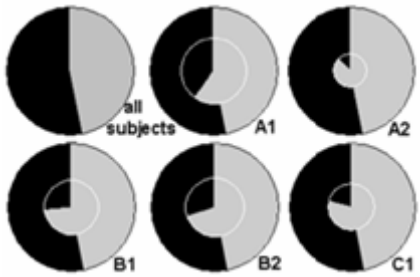

Figura 11 - Exemplo subgrupos em gráficos de setores.
Fonte: KRALJ et al., 2005

Apesar de possuir uma interpretação trivial, este tipo de gráfico possui algumas limitações em relação à visualização de subgrupos. Uma delas está atrelada ao fato de se ter a representação de apenas um subgrupo por gráfico, pois desta forma torna-se difícil realizar comparações entre subgrupos.

4.3.2 Diagrama de Caixas

Uma outra forma de representar graficamente subgrupos é através do diagrama de caixas em que, ao contrário da técnica anterior, é possível visualizar todos os subgrupos em apenas um gráfico. Cada subgrupo é representado por uma caixa, sendo esta dividida em duas partes: uma à esquerda, contendo várias linhas horizontais, e outra em branco mais à direita. A parte à esquerda da caixa representa as amostras que pertencem a uma dada classe, enquanto que a parte a esquerda representa as amostras que não pertencem a classe. Na Fig 12 é ilustrado

um gráfico desse tipo, onde se pode observar também que a caixa no ponto mais alto representa toda população.

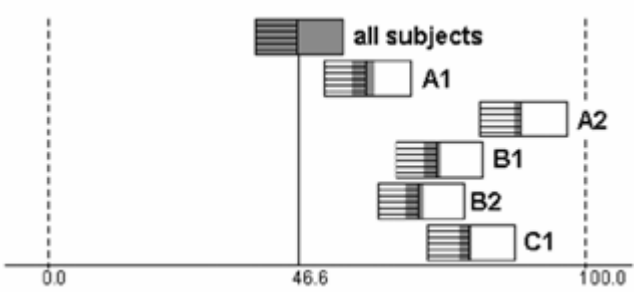

Figura 12 - Um diagrama de caixas contendo subgrupos.
Fonte: KRALJ et al., 2005

As áreas hachuradas com cinza representam o percentual do tamanho do subgrupo em relação a toda população. Pode-se observar ainda que 46.6% das amostras de toda população pertencem a classe referida anteriormente. O restante das caixas são ordenadas no eixo x de acordo com a porcentagem de amostras pertencentes a classe em cada subgrupo.

4.3.3 Visualização da Distribuição de um Atributo Contínuo

Nesta técnica é possível visualizar a distribuição de um atributo contínuo em relação aos subgrupos, o qual é representado no eixo x de um gráfico. Cada subgrupo é representado através de um par de linhas, uma referente às amostras pertencentes a uma dada classe e a outra referente às amostras não pertencentes à classe. O eixo y é divido em duas partes, a primeira se refere às amostras que pertencem a dada classe (parte superior) e a segunda se refere às que não pertencem a classe (parte inferior). Para cada valor do atributo contínuo escolhido é possível visualizar, no eixo y, uma estimativa da freqüência de amostras de cada classe. A Fig. 13 exibe um gráfico relativo a dados de pacientes com problemas cardíacos, os quais são divididos em duas classes, os com a doença coronária (CHD) e os saudáveis.

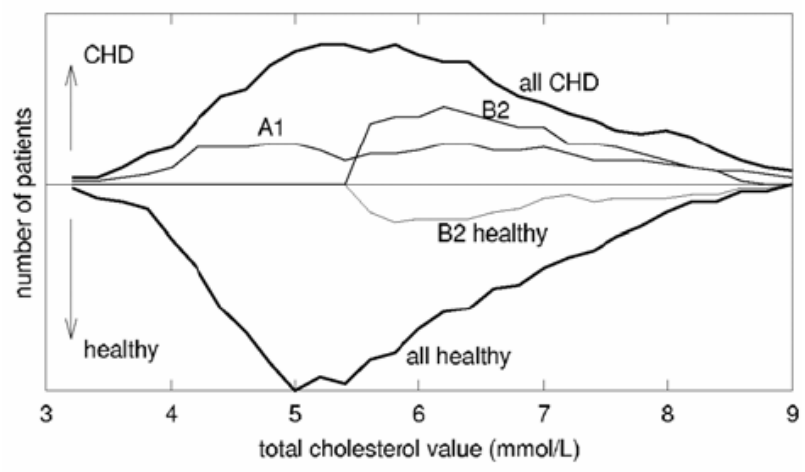

Figura 13 - Exemplo de visualização da Distribuição de um Atributo Contínuo.

Fonte: KRALJ et al., 2005.

As linhas em negrito da figura representam toda população. A variável contínua em análise representa a taxa de colesterol. Tomando-se como base o subgrupo B2, o qual é composto por pacientes com nível de colesterol entre 5.5 e 9.0, observa-se que para este subgrupo o número de pacientes com a doença coronária é predominante em relação aos que não possuem a mesma.

4.3.4 Utilização da curva ROC

Esta técnica é mais utilizada como um meio de avaliação dos resultados obtidos pelo processo de descoberta de subgrupos, visto que não possui uma visualização intuitiva (KRALJ et al., 2005).

A utilização da curva ROC *(Receiver Operating Characteristics)* na descoberta de subgrupos consiste em um espaço de visualização composto por um gráfico de duas dimensões, no qual ilustra-se o grau de interesse de cada subgrupo descoberto. Para a avaliação do grau de interesse, utiliza-se a medida FPr (*false positive rate*) no eixo x do gráfico e a medida TPr (*true positive rate*) no eixo y do gráfico. A TPr corresponde a taxa de amostras de um subgrupo que pertence a uma dada classe, atribuída através da propriedade de interesse, em relação ao total de amostras pertencentes a classe. A FPr corresponde a taxa de amostras do subgrupo que não pertence a tal classe. Na Fig. 14 é ilustrado uma curva ROC contendo quatro subgrupos.

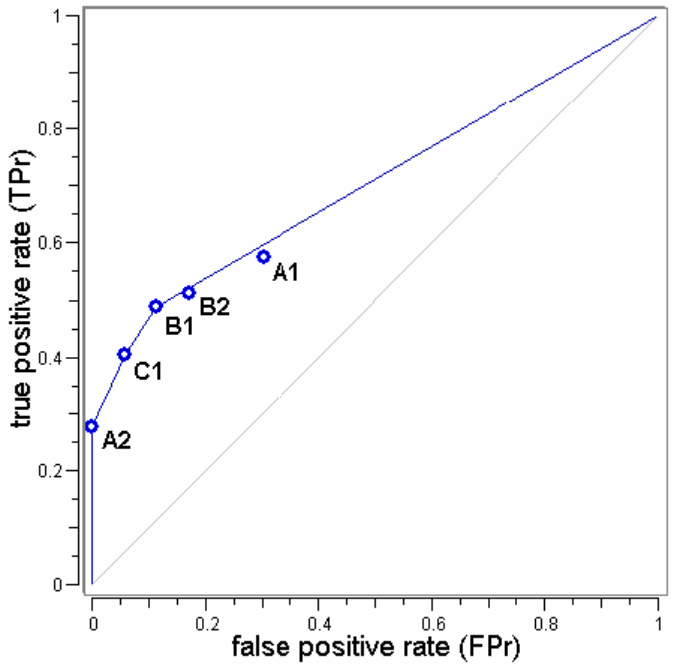

Figura 14 - Uma Curva ROC utilizada na descoberta de subgrupos.

Fonte: KRALJ et al., 2005

Observa-se que o gráfico possui uma linha diagonal que divide o mesmo em duas partes, os subgrupos que estiverem próximos de tal linha devem ser considerados como não interessantes, pois tal linha representa igualdade nos valores de TPr e FPr. Esta igualdade torna o subgrupo desinteressante, pois expressa a existência uma freqüência similar de amostras das duas classes no subgrupo. Tomando-se o subgrupo A2 da figura anterior e supondo o mesmo exemplo da subseção anterior, nota-se que A2 somente terá pacientes com CHD, fato que permitirá uma análise mais confiável em relação à ocorrência da doença em tal subgrupo.

4.3.5 Gráfico de Barras

Uma outra técnica existente para visualização de subgrupos, proposta recentemente por Kralj et al. (2005), permite visualizar cada subgrupo em uma linha de um gráfico de barras, onde a primeira linha representa toda população. O gráfico é dividido em duas partes, uma à direita, representando as amostras que pertencem a uma dada classe, e outra à esquerda, representando as amostras não

pertencentes a classe. A Fig. 15 ilustra a representação de cinco subgrupos através de tal técnica.

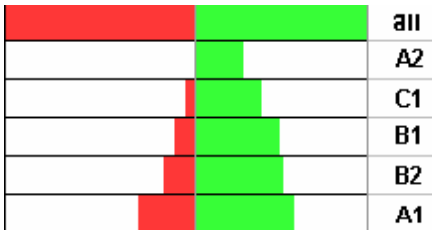

Figura 15 - Um Gráfico de barras utilizado na descoberta de subgrupos.

Fonte: KRALJ et al., 2005.

Os subgrupos são ordenados de forma descendente em relação à proporção de elementos presentes na classe em análise.

4.4 Regras de Distribuição

Uma forma de representar subgrupos é através de uma Regra de Distribuição, DR, (JORGE, 2006a). Visto que este trabalho se baseia na metodologia proposta por Pereira (2006) e Jorge (2006b), na qual propõe-se utilizar regras de distribuição na descoberta de subgrupos, a compreensão desse tipo de regra é de fundamental importância para o desenvolvimento deste trabalho.

Uma DR pode ser vista como uma regra que abrange uma ou mais regras de associação, pois todos termos antecedentes idênticos destas são reunidos no termo antecedente da DR. O termo conseqüente da regra corresponde a distribuição da propriedade de interesse acerca dos termos antecedentes das regras de associação, as quais formam o termo antecedente da DR.

Jorge (2006a) formaliza a seguinte representação relativa a uma DR:

$$A \rightarrow y = Dy|A$$

Onde, *A* corresponde ao conjunto de itens que expressam o termo antecedente da DR, *y* corresponde a propriedade de interesse (categórica ou numérica) e *Dy|A* corresponde a distribuição de *y* onde *A* está presente. D*y|A* é composto por um conjunto de elementos y*i* / freq(y_i), onde y_i é um dos valores de y encontrado quando os itens de A ocorrerem e freq(y_i) corresponde a sua freqüência, ou seja, quantas vezes os itens de A ocorrem juntos quando y_i também ocorrer.

Um exemplo desta formalização pode ser ilustrado através da seguinte amostra hipotética contida na tab. 8, a qual é relativa aos dados de identificação de pacientes diagnosticados com resfriado, gripe ou pneumonia.

Tabela 8 - Amostra hipotética.

Paciente	Cidade	Diagnóstico	Idade
1	Pelotas	Resfriado	35
2	Bagé	Resfriado	18
3	Bagé	Pneumonia	45
4	Canguçu	Gripe	58
5	Pelotas	Pneumonia	60
6	Bagé	Gripe	14
7	Pelotas	Resfriado	32
8	Canguçu	Pneumonia	38
9	Pelotas	Pneumonia	74
10	Pelotas	Pneumonia	60

Supondo que o analista esteja interessado somente em analisar a idade dos pacientes diagnosticados com pneumonia que residam na cidade de Pelotas. Neste caso, o analista teria um possível subgrupo representado pela seguinte DR:

{Pelotas & pneumonia} -> idade = {60/2, 74/1}

4.5 Ferramentas para a Descoberta de Subgrupos

Apesar de existir uma grande quantidade de ferramentas de KKD, atualmente existem poucas ferramentas que possuem implementações de metodologias eficientes para a realização da descoberta de subgrupos.

Dentre as ferramentas que dão suporte à descoberta de subgrupos, pode-se destacar o ORANGE (DEMŠAR; ZUPAN; LEBAN, 2004) e *SubgroupMiner* (KLÖSGEN; MAY, 2002). O ORANGE é um *framework* que possibilita a implementação de algoritmos e aplicações de aprendizado de máquina e mineração de dados, possuindo uma interface visual de programação que possibilita a criação de componentes para implementação de aplicações para a descoberta de subgrupos. Cabe salientar que o método de visualização de subgrupos em gráfico de barras, descrito na seção 4.3, foi desenvolvido através do *ORANGE*, o qual também possui implementações de algoritmos de descoberta de subgrupos,

possibilitando assim o desenvolvimento de novos algoritmos que sejam especializações desses.

Dentre as implementações existentes no ORANGE, pode-se destacar a implementação da curva ROC para a visualização de subgrupos, conforme é ilustrado no lado direito da Fig. 16. Na parte inferior da figura tem-se a representação dos subgrupos através de gráfico de barros e no lado esquerdo tem-se uma tela inicial de opções que permite escolher com qual método se deseja visualizar os subgrupos.

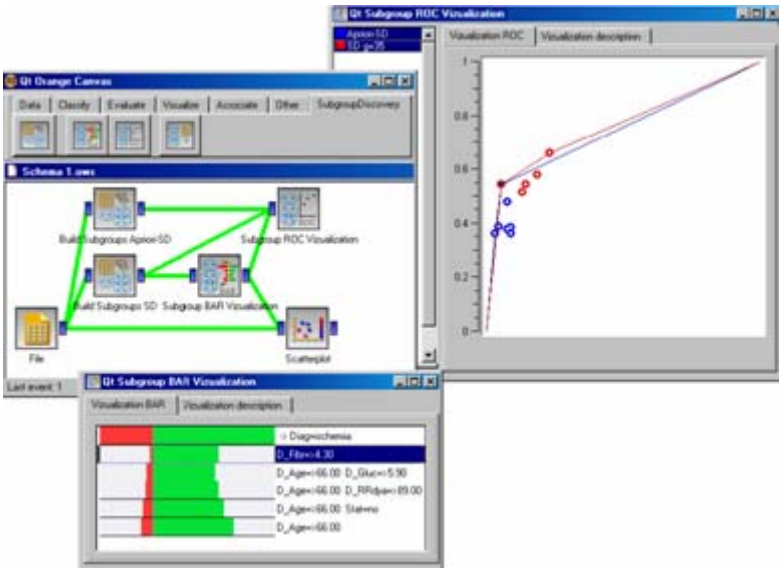

Figura 16 - Tela do *ORANGE*.
Fonte: Adaptado de KRALJ et al., 2005.

Já o *SubgroupMiner* constitui-se de um sistema de KDD direcionado para a descoberta de subgrupos, fornecendo suporte, apesar de limitado, para as etapas anteriores a mineração de dados. Por outro lado, o sistema possui um poderoso componente de visualização integrado com um GIS (Sistema de Informação Geográfica) que permite a análise de subgrupos através de um mapa.

5 A FERRAMENTA PROPOSTA

O objetivo principal da ferramenta desenvolvida neste trabalho é fornecer suporte ao analista em processos de tomada de decisão, proporcionando a ele um ambiente interativo de visualização. Tal ambiente é composto, basicamente, pela descrição gráfica de subgrupos por meio de medidas estatísticas referentes à distribuição dos mesmos. Cada subgrupo é representado por meio de uma regra de distribuição, sendo esta composta por um conjunto de regras de associação.

A ferramenta desenvolvida neste trabalho baseia-se em uma metodologia proposta por Pereira (2006) e Jorge (2006b), para a descoberta de subgrupos por meio do pós-processamento de regras de associação. A referida metodologia é descrita com mais detalhes na seção a seguir.

5.1 A Metodologia Utilizada

Na metodologia desenvolvida por Pereira (2006) e Jorge (2006b), propõe-se o uso de técnicas de visualização para possibilitar uma forma interativa de descobrir subgrupos, não apenas limitando-se a exibição do resultado gerado por um algoritmo de descoberta de subgrupos. Através dessa interação, o analista pode utilizar seu conhecimento acerca do domínio durante a realização da análise.

Para a realização da metodologia é necessário que uma variável seja escolhida para ser a propriedade de interesse para que, em seguida, sejam geradas regras de distribuição contendo tal propriedade no termo conseqüente das mesmas, permitindo assim que a análise seja guiada para a descoberta de subgrupos.

Propõe-se que tais regras de associação sejam obtidas na forma de regras de distribuição, onde cada uma destas representa um subgrupo da população. Faz-se necessário também construir uma DR que contenha a distribuição da propriedade de interesse em toda população, a qual é chamada de distribuição *apriori*. Nessa

metodologia, ao contrário da maioria das técnicas existentes, a propriedade de interesse não necessita ser discretizada e não pode ser categórica, pois se fará uso da distribuição da propriedade de interesse durante a análise.

Visto que o conjunto de DRs obtido pode ser bastante grande, propõe-se a realização do pós-processamento das mesmas utilizando recursos de visualização, pois assim o analista pode interpretar mais facilmente e também explorar o conjunto de DRs obtido. O método para a visualização de regras sugerido para ser utilizado é o da matriz de duas dimensões, sendo representação por gráficos de *grids*.

Com a utilização de regras de distribuição tem-se disponível a distribuição da propriedade de interesse de todos subgrupos da população. Sendo assim, é possível então analisar a distribuição de cada um por meio de medidas estatísticas, onde estas são utilizadas como coordenadas para a construção de gráficos de grids contendo a representação dos subgrupos da população. Na Fig. 17. é ilustrado um gráfico contendo a representação dos subgrupos de uma população, através das medidas de desvio padrão e média. Cada ponto vermelho do gráfico representa um subgrupo e o ponto azul representa toda população (distribuição apriori).

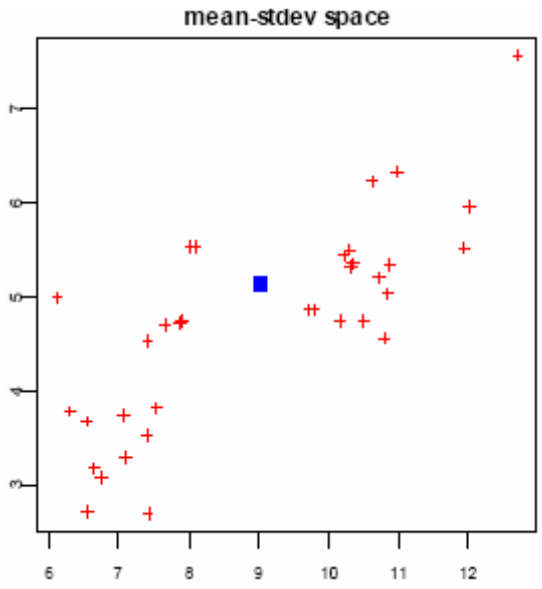

Figura 17 - Espaço em 2D representando
os subgrupos de uma população.

Fonte: PEREIRA, 2006.

Como última etapa da análise, propõe-se que seja possibilitado ao analista escolher e visualizar a distribuição da propriedade de interesse de um subgrupo

qualquer e compará-la com a distribuição da mesma em toda população, conforme é ilustrado na Fig. 18. A linha preta no gráfico representa tal distribuição população.

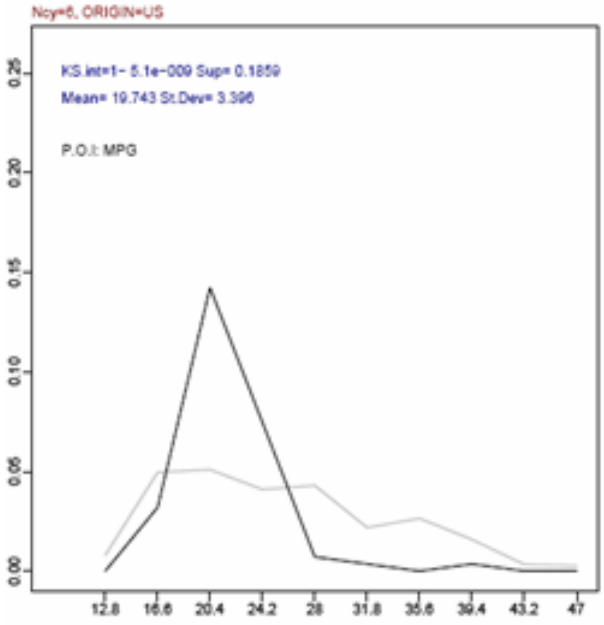

Figura 18 - Representação gráfica de uma DR representando um subgrupo.
Fonte: PEREIRA, 2006.

Observa-se que na parte superior da figura tem-se o antecedente da DR (Ncy=6 & ORIGIN=US), o qual pode ser visto como uma descrição das características do subgrupo, que no contexto dos dados do gráfico representa carros com seis cilindros e fabricados nos Estados Unidos. O eixo Y do gráfico expressa a freqüência relativa de cada um dos elementos que compõem o subgrupo em questão.

Tanto a Fig. 17, como a 18, foram extraídas de um protótipo implementado junto com a elaboração desta metodologia.

5.2 Aspectos gerais

O objetivo do desenvolvimento da ferramenta proposta neste trabalho é o de proporcionar ao analista um meio viável de descobrir subgrupos visualmente em conjuntos de dados. Para tanto, a ferramenta implementa a metodologia descrita na seção anterior e, dessa forma, vale-se de princípios relativos tanto ao pós-

processamento de regras de associação, como também da descoberta de subgrupos.

A vantagem de se ter uma implementação híbrida, abrangendo tais princípios, é a possibilidade de poder aplicar, na visualização de subgrupos, técnicas de visualização utilizadas no pós-processamento de regra de associação. Atualmente, a maioria das ferramentas que possibilitam a visualização de subgrupos utilizam somente as técnicas expostas na seção 4.2 do capítulo anterior. Em tais técnicas só é possível analisar um número limitado de categorias da propriedade de interesse, ao contrário da ferramenta proposta neste trabalho que lida com variáveis numéricas, pois seu objetivo é possibilitar uma análise focada em características relativas a distribuição da propriedade de interesse dos subgrupos.

Um outro aspecto inovador desta ferramenta é relativo a interação proporcionada durante a análise, pois, atualmente, as ferramentas visuais para descoberta de subgrupos apresentam interação somente na escolha da forma de visualizar os subgrupos, não apresentando interação acerca dos gráficos que representam os subgrupos. A ferramenta desenvolvida neste trabalho possibilita ao analista navegar no espaço de subgrupos, no qual é possível escolher algum destes para ser analisado com maior detalhamento. Além disso, é possível realizar uma espécie de filtragem no espaço de subgrupos, onde são exibidos somente aqueles que respeitarem a uma determinada medida de suporte atribuída pelo analista em tempo de execução.

5.3 Aspectos da Implementação

O primeiro passo realizado para implementação da ferramenta foi a escolha da linguagem de programação. A plataforma Java foi escolhida para implementação devido, em parte, a característica de seu compilador, o qual, segundo Brandão e Moreira (2002), traduz programas fonte em um código intermediário e independente de plataforma, chamado *Java Byte Code*, o qual é interpretado por uma Máquina Virtual Java (JVM). Tal característica torna as aplicações desenvolvidas em Java independentes de plataforma. Com isso, pode-se também disponibilizar uma aplicação em um ambiente Web para ser executada por um navegador qualquer, na forma de um *Java Applet*.

Um outro fator determinante nessa escolha foi a vasta coleção de bibliotecas de classes que a plataforma disponibiliza, possibilitando-se assim implementar uma interface gráfica amigável e interativa.

Além da biblioteca padrão disponível na plataforma Java, utilizou-se na implementação algumas bibliotecas adicionais: a *Jakarta Mathematics* e a *GenJava-CSV*.

A *Jakarta Mathematics* é uma biblioteca de classes desenvolvida no projeto Jakarta Commons (2006), que disponibiliza, sob a licença Apache versão 2.0, uma ampla biblioteca de classes aplicável a resolução de problemas oriundos de diversos domínios, dentre eles o da matemática. Porém, neste trabalho foi utilizada somente o pacote contendo implementações de algoritmos aplicados à estatística descritiva.

A outra biblioteca de classes utilizada, a *GenJava-CSV*, consiste de implementações para o tratamento de leitura e escrita de arquivos CSV. A *GenJava-CSV* é distribuída sob a licença BSD e pertence a um conjunto de bibliotecas de código fonte aberto implementadas por um grupo de desenvolvedores, os quais compõem uma organização chamada OSJava (Open Sourced Java) (OSJAVA, 2006).

Através das bibliotecas de classes da plataforma Java, em especial a Java Swing e a Java AWT, juntamente com as duas citadas anteriormente, desenvolveu-se as seguintes classes que compõem a ferramenta proposta neste trabalho: SDiscoveryTool, CSVProcessor, PreProcessor, Subgroup, SubgroupSpace, PropIntDistribution, Explorer, Language e outras classes acessórias relativas aos idiomas disponíveis para exibição dos textos presentes na interface gráfica. As classes da ferramenta se relacionam conforme exposto no diagrama de classes presente na Fig. 19.

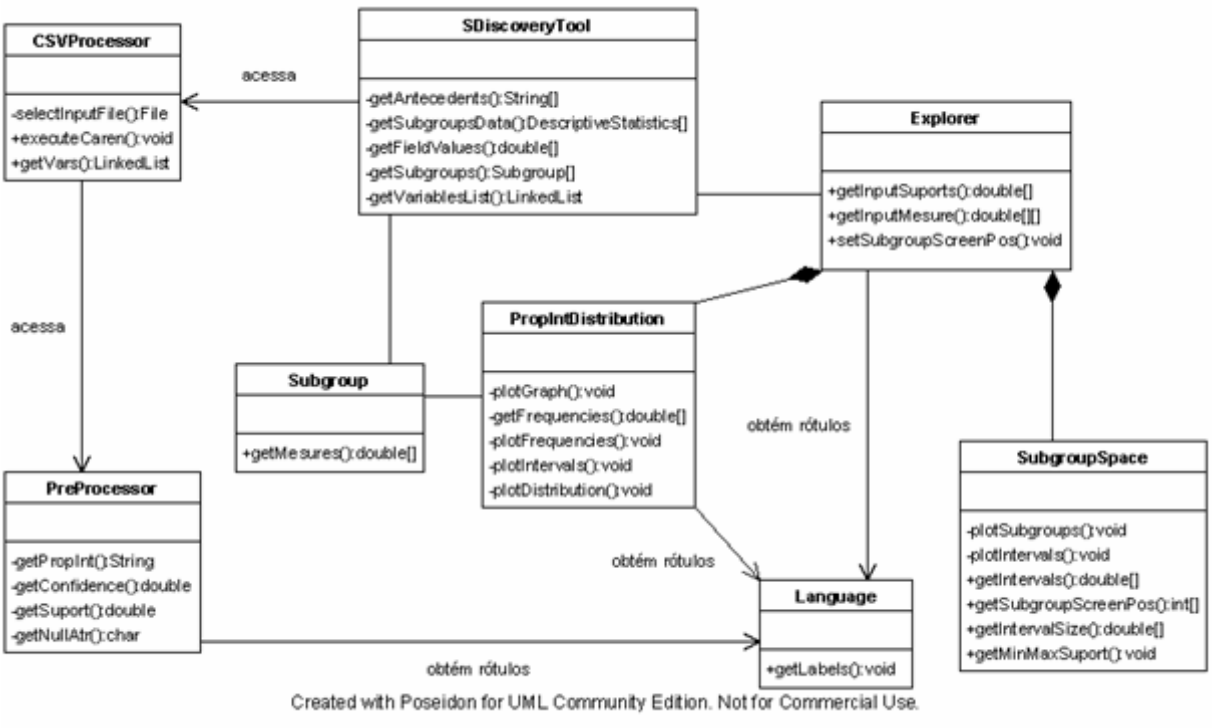

Figura 19 - Diagrama de classe da ferramenta.

A classe *SDiscoveryTool* é a classe principal da ferramenta, ela é quem dispara o início da execução e também cria uma instância da classe *PreProcessor,* sendo esta responsável por exibir uma tela de escolha de parâmetros a serem fornecidos para invocar a ferramenta CAREN-DR (JORGE, 2006a), a qual também foi desenvolvida em Java. Tal ferramenta possui o papel de gerar as regras de distribuição a serem utilizadas na composição dos subgrupos, sendo sua invocação realizada de forma transparente ao analista. A arquitetura do CAREN-DR é baseada na ferramenta CAREN (AZEVEDO, 2003), que implementa, para gerar as regras de associação, uma especialização do algoritmo *Apriori*, o qual foi descrito na subseção 3.1.2. A classe *CSVProcessor* é quem realiza todo tratamento acerca do arquivo CSV que contém os dados de entrada fornecidos pelo analista. A *Explorer* é a classe que contém a tela principal da ferramenta, na qual estão presentes todos elementos da interface gráfica. Sendo assim, ela possui um relacionamento de agregação com as classes *SubgroupSpace* e *PropIntDistribution*, pois estas são responsáveis, respectivamente, por exibir dois gráficos na parte superior e um na parte inferior da tela. A classe *SubgroupSpace* é instanciada duas vezes, pois é ela quem constrói os dois gráficos da parte superior da tela. A classe *Subgroup* fornece, às classes

responsáveis pela exibição dos gráficos, todo conteúdo e medidas relativas aos subgrupos obtidos através dados. Por fim, tem-se a classe *Language*, ela é responsável por gerenciar qual idioma será utilizado nos textos presentes na interface gráfica. Inicialmente, foram implementados dois idiomas (português e inglês), porém pode-se trivialmente adicionar outros.

5.3 Utilização da Ferramenta

Para dar início ao processo de análise, primeiramente é necessário que o analista forneça como entrada para a ferramenta um conjunto de dados representados em um arquivo CSV, no qual já se tenha realizado as etapas iniciais (anteriores a mineração) do processo de KDD. Um arquivo CSV (*Comma Separated Values*) contém a representação de dados em colunas separadas por vírgula (ou ponto e vírgula). Além de ser um formato praticamente universal para migração de dados, sendo utilizado em diversos SGBDs atuais, pode ser lido em diversos editores de planilhas.

Após fornecida a entrada dos dados, é necessário agora dispor de um conjunto de regras de distribuição geradas a partir dos mesmo. Para tanto, antes de invocar a ferramenta CAREN-DR, exibe-se a tela de opções relativa aos parâmetros a serem fornecidos ao CAREN-DR, conforme pode-se observar na Fig. 20.

Figura 20 - Tela de escolha dos parâmetros de entrada para o CAREN.

Os primeiro campo de preenchimento é relativo ao suporte mínimo que as regras geradas deverão apresentar. O segundo especifica qual símbolo irá representar o valor nulo dos atributos, onde, por padrão, tem-se o símbolo “?”. Por último tem-se o campo de seleção relativo a variável que será a propriedade de interesse. A especificação deste último parâmetro só é possível graças a uma funcionalidade da ferramenta CAREN-DR, onde pode-se condicionar a geração de regras que contenham somente uma determinada variável em um de seus termos. Sendo assim, torna-se possível a construção das regras de distribuição, pois dessa forma é possível condicionar uma determinada variável para ser o termo conseqüente de todas as regras a serem geradas e também para tal variável não ser nenhum dos itens que compõem os termos antecedentes das regras.

O próximo passo após a obtenção das regras de distribuição é agrupar as regras que possuam antecedentes iguais, para, logo após, iniciar o processo de pós-processamento e análise. O panorama geral do conjunto dos subgrupos identificados é dado por dois gráficos de navegação que representam cada um dos subgrupos em função de medidas estatísticas da distribuição da propriedade de interesse nesses subgrupos. Nesse momento o analista pode selecionar um subgrupo específico para visualizar, em um terceiro gráfico, sua distribuição e freqüência relativa dos indivíduos que o compõem. Na Fig. 21 é exibido um diagrama de atividades relativo a tal processo.

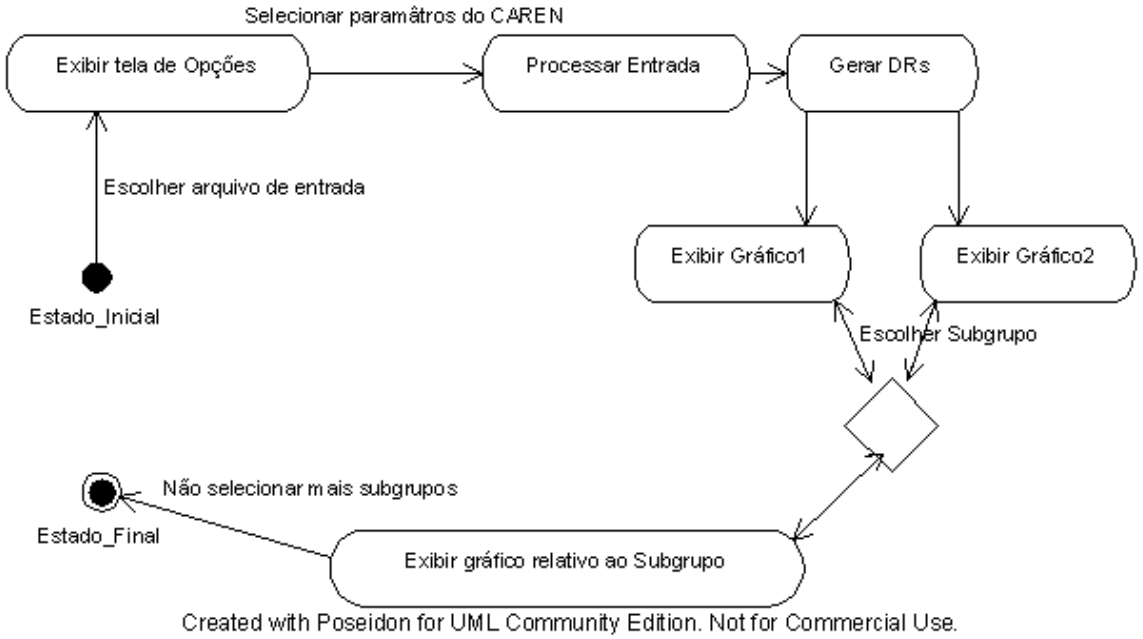

Figura 21 - Diagrama de atividades da ferramenta.

Relativamente aos gráficos de navegação 1 e 2, o analista pode escolher dentre seis medidas estatísticas que irão compor seus eixos cartesianos: mediana, média aritmética, moda, desvio padrão, assimetria e curtose. As três primeiras são medidas de posição, onde a mediana refere-se ao valor que divide o conjunto de dados em duas partes iguais e a moda ao valor que se repete o maior número de vezes no conjunto. A quarta medida, o desvio padrão, é uma medida de dispersão que indica o quão dispersos são os valores da distribuição em relação à média. Por fim, tem-se as medidas relativas ao formato da distribuição: a assimetria e curtose. A primeira indica se a maior parte dos valores da distribuição encontram-se à esquerda (assimetria maior que zero), direita (assimetria menor que zero) ou igualmente distribuídos em torno da média. A segunda indica o grau de achatamento da distribuição em relação à distribuição normal. Um valor de curtose menor que três, indica grau de achatamento maior que a da distribuição normal, um valor maior que três indica menor achatamento e um valor igual a três indica o mesmo grau de achatamento.

A escolha ideal das medidas a serem utilizadas varia de acordo com as características de cada população e também com o objetivo de análise. Por padrão, em um gráfico são exibidas as medidas relativas ao formato da distribuição e, no outro, medidas de posição (mediana e moda).

A Fig. 22 exemplifica interface gráfica da ferramenta durante sua execução, onde, na parte superior, apresentam-se os dois gráficos citados anteriormente.

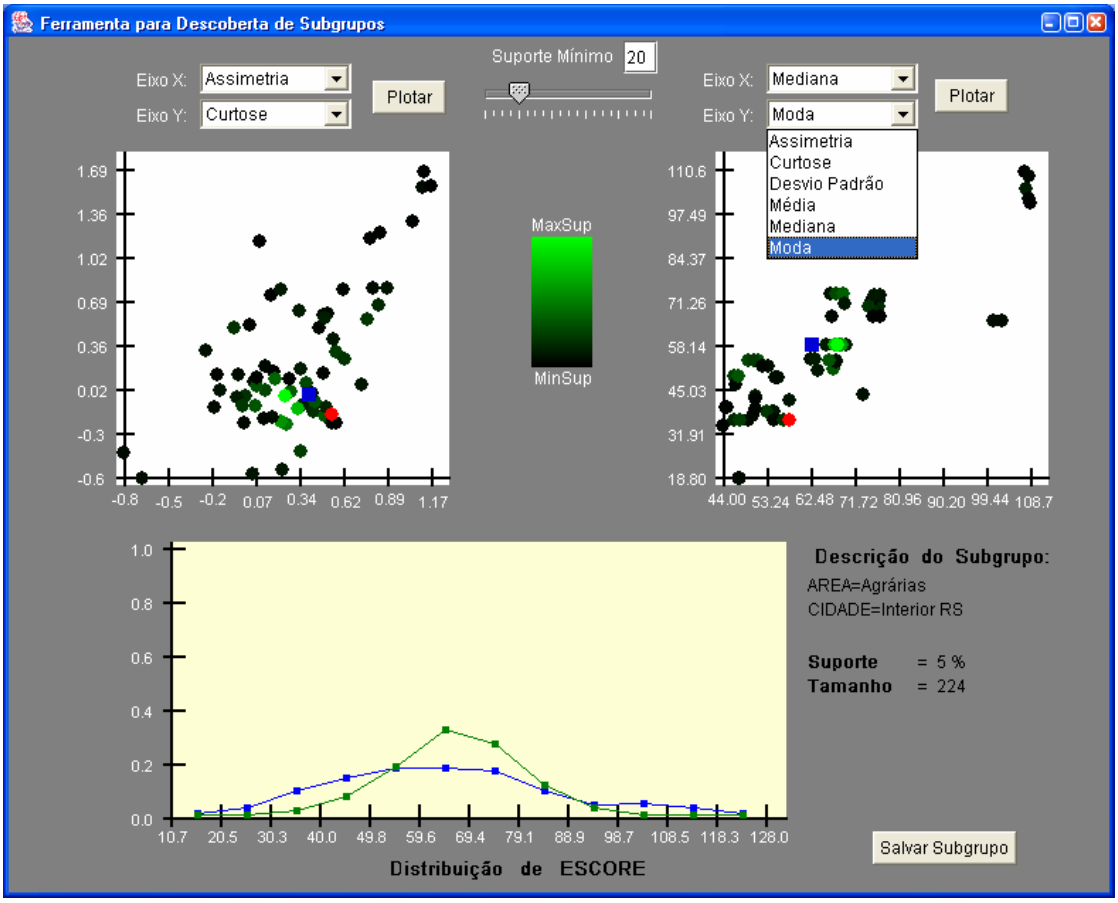

Figura 22 - Interface gráfica da ferramenta.

A interação do analista com os dois gráficos de navegação ocorre através dos campos de seleção das medidas estatísticas que irão compor os eixos cartesianos dos gráficos. Tomando como exemplo o gráfico de navegação presente no lado direito da figura, observa-se o momento em que a Moda está sendo selecionada para compor o eixo y. Após selecionadas as medidas, o analista pode clicar com o mouse no botão “Plotar” e assim os subgrupos são posicionados no gráfico de acordo com as medidas estatísticas definidas para os eixos X e Y. Tal processo pode ser repetido, em ambos os gráficos, quantas vezes o analista desejar.

Os pontos de cor verde presentes nos gráficos representam os subgrupos da população e, o quadrado na cor azul, representa a distribuição *Apriori*. Nota-se que há uma diferença na intensidade da cor dos pontos em verde, esta diferença representa a medida de suporte de cada subgrupo, onde os de tonalidade mais forte apresentam maior suporte. No espaço compreendido entre os dois gráficos, pode-se

observar a existência de um campo de seleção, este campo refere-se ao valor do suporte mínimo estabelecido pelo analista. Cada vez que o analista altera tal valor, exibe-se, nos dois gráficos de navegação, somente os subgrupos que possuírem suporte maior que o valor definido.

Caso o analista clique com o mouse em um subgrupo mostrado em algum dos dois gráficos de navegação, o ponto do gráfico, que representa o subgrupo, passa a ser exibido na cor vermelha em ambos gráficos. Dessa forma, pode-se analisar com mais clareza o subgrupo escolhido, comparando-o mais facilmente com o restante dos subgrupos e a população. Além disso, exibe-se um gráfico (presente na parte inferior da figura) que representa a distribuição da propriedade de interesse do subgrupo (linha em verde), assim como a freqüência relativa de cada individuo que o compõe através de um polígono de freqüências. Ao lado, tem-se a descrição (itens presentes no antecedente da DR). A distribuição *apriori* também é exibida no gráfico (linha em azul), a qual possibilita que o analista realize uma comparação acerca das características de um subgrupo qualquer com as características de toda população, tornando mais fácil a identificação de subgrupos de configuração discrepante ou atípica.

Ao identificar um subgrupo como interessante, o analista pode clicar no botão “Salvar Subgrupo”, pois ao fim da execução da ferramenta, dispõe-se de um arquivo, no formato csv, contendo informações sobre os subgrupos salvos durante a análise.

6 VALIDAÇÃO

Como etapa final deste trabalho, propõe-se a realização de um estudo no qual se possa testar as funcionalidades da ferramenta descrita no capítulo anterior.

Para tanto, fez-se necessário escolher um conjunto de dados para ser fornecido como entrada para a ferramenta, para que, com o auxílio desta, fosse realizada a descoberta de subgrupos. Posteriormente, tendo disponível um conjunto de subgrupos descobertos, pretende-se contribuir para uma melhor compreensão de fatos inerentes ao contexto no qual os dados estão inseridos.

6.1 O Conjunto de Dados Escolhido

Visto que o objetivo neste estudo é o da validação da ferramenta e não o de realizar uma análise estatística completa, buscou-se algum conjunto de dados pequeno e simples, no qual não fosse necessário despender a maior parte do tempo nas etapas iniciais do processo de KDD. Sendo assim, o contexto do conjunto de dados escolhido necessitou ser de simples entendimento, não estando atrelado a uma área específica do conhecimento.

Por outro lado, um aspecto relevante para a escolha foi que o conjunto de dados escolhido permitisse identificar subgrupos interessantes, onde fosse possível confrontar a realidade com resultados obtidos através da ferramenta e, além disso, identificar fatos até então desconhecidos.

Sendo assim, os dados do processo seletivo da Universidade Federal de Pelotas (UFPel), realizado em dezembro de 2005, foram os escolhidos como objeto de análise na validação da ferramenta, pois eles vão ao encontro dos requisitos traçados nesta etapa do trabalho.

6.2 Variáveis em Análise

Antes de iniciar o processo de KDD, faz-se necessário traçar objetivos específicos para a realização do mesmo, pois todo o processo é realizado em função de tais objetivos. Por sua vez, estes estão diretamente relacionados com as variáveis a serem abordadas.

Abordar uma grande quantidade de variáveis tende a tornar os objetivos da análise bastante amplos e que, no caso deste trabalho, não seria conveniente.

Sendo assim, o processo de KDD realizado focou-se na tentativa de identificar relações do desempenho do candidato com informações fornecidas no ato de sua inscrição no processo seletivo, dentre as quais foram escolhidas as seguintes variáveis: idade, turno do curso escolhido, cidade onde reside, área do conhecimento que o curso escolhido faz parte e o escore bruto atingido.

Entretanto, antes de iniciar a descoberta de subgrupos, é necessário aplicar as etapas iniciais do processo de KDD ao conjunto de dados fornecido como entrada.

6.3 Preparação para a Descoberta de Subgrupos

A preparação para a Descoberta de Subgrupos inicia-se com a realização da etapa de pré-processamento. Conforme exposto na seção 2.3.1., nessa etapa faz-se a limpeza e integração dos dados.

Dentre as seis variáveis utilizadas na análise, necessitou-se realizar a limpeza dos dados somente no atributo relativo ao escore bruto do candidato, o qual foi escolhido para ser a propriedade de interesse na descoberta de subgrupos. O valor do escore bruto pode variar de 0 a 140 e é obtido conforme a pontuação do candidato nas provas de questões de múltipla escolha (escore parcial) somada à nota da redação. Esta só é corrigida nas seguintes situações: o escore parcial obtido pelo candidato se enquadrar até a colocação composta pelo número de vagas de seu curso multiplicado por três; o escore parcial obtido pelo candidato for maior ou igual a 60 e o número de candidatos por vaga no seu curso não ultrapassar a seis.

Sendo assim, os candidatos que não tiveram suas redações corrigidas poderiam atingir no máximo 120 pontos no escore bruto. Portanto, foi necessário

excluir da análise os registros referentes a tais candidatos. Com isso, o número de registro do conjunto de entrada reduziu de 13400 para aproximadamente 4000.

Após a realização da limpeza dos dados partiu-se para a etapa de seleção, pois não se fez necessário realizar a integração dos dados, já que estes são oriundos de uma única fonte. Nesta etapa, selecionou-se do conjunto de dados somente os atributos referentes às variáveis mencionadas na seção anterior.

Já na etapa de transformação dos dados, a última antes da etapa de mineração, fez-se necessário modificar os valores dos seguintes atributos: idade, cidade onde reside e área do conhecimento do curso escolhido.

No atributo relativo à idade do candidato, o que se tinha, inicialmente, era sua data de nascimento. Portanto, foi necessário transformar a data de nascimento para a idade do candidato. Além disso, dividiu-se a idade dos candidatos nos seguintes intervalos: igual ou inferior a 17; entre 18 e 20; entre 21 e 23; entre 24 e 27; entre 28 e 30; entre 31 e 35; entre 36 e 42; entre 43 e 50; maiores que 50.

Quanto ao atributo relativo à cidade onde o candidato reside, foram feitas modificações para se ter somente as seguintes cidades:

- Pelotas;
- Porto Alegre;
- Rio Grande;
- Municípios de fora do estado do Rio Grande do Sul;
- Municípios pertencentes a região de Pelotas, a qual, segundo o IBGE (Instituto Brasileiro de Geografia e Estatística), compreende os municípios de Arroio do Padre, Canguçu, Capão do Leão, Cerrito, Cristal, Morro Redondo, Pedro Osório, São Lourenço do Sul e Turuçu;
- Restantes dos municípios pertencentes ao interior do estado do Rio Grande do Sul.

Por fim, restringiu-se o atributo relativo à área do conhecimento do curso escolhido pelo candidato para conter os seguintes valores: letras e artes, ciências humanas, ciências exatas e tecnologia, ciências biológicas, ciências agrárias, direito e medicina. Estes dois últimos foram separados de suas áreas do conhecimento por serem os cursos com o maior número de candidatos.

Cabe salientar que a realização das etapas do KDD mencionadas até aqui, não costuma ser simples e, segundo Bogorny (BORGORI, 2003) consome cerca de

80% dos esforços necessários para concluir todo o processo. Porém, não foi o que ocorreu neste trabalho, devido, principalmente, ao tamanho e à simplicidade do conjunto de dados escolhido.

6.4 Resultados Obtidos

Após a conclusão das etapas mencionas na seção anterior, partiu-se então para as etapas da mineração e pós-processamento, as quais foram realizadas com auxílio da ferramenta desenvolvida e seguindo a metodologia exposta na seção 5.1.

Após a realização dessas duas etapas, obteve-se então alguns fatos e subgrupos que foram considerados interessantes, os quais são descritos ao longo desta seção.

A primeira característica interessante percebida no conjunto de dados em questão foi o caso dos subgrupos compostos por candidatos da medicina, tais subgrupos possuem os menores desvios padrão. Tal fato leva a concluir que os candidatos da medicina, em geral, atingiram um escore bruto semelhante. Na figura abaixo, os referidos subgrupos estão contidos no círculo desenhado em vermelho e o eixo Y do gráfico representa a medida de desvio padrão (Fig. 23).

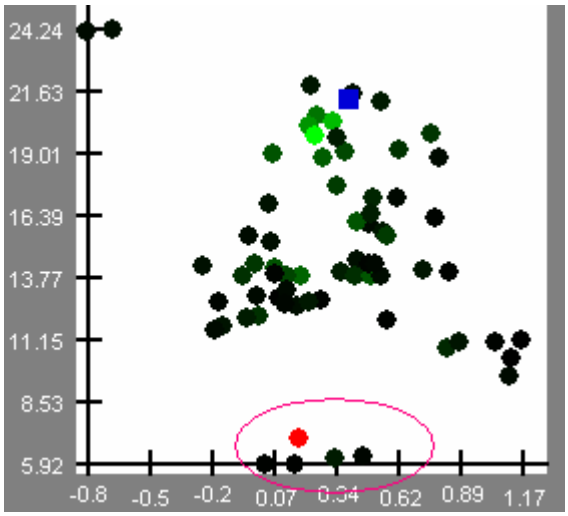

Figura 23 - Subgrupos de candidatos ao curso de medicina.

Dentre os subgrupos referidos anteriormente, pode-se destacar o subgrupo, entitulado S1, que é composto por candidatos ao curso de medicina que residem fora do estado. S1 possui a maior média dentre os subgrupos encontrados, e

também possui assimetria próxima de 0 (0,1). Este valor indica a existência de simetria nos escores dos candidatos, fato que também pode ser confirmado através do gráfico presente na Fig. 24, no qual os valores de maior freqüência são os escores maiores.

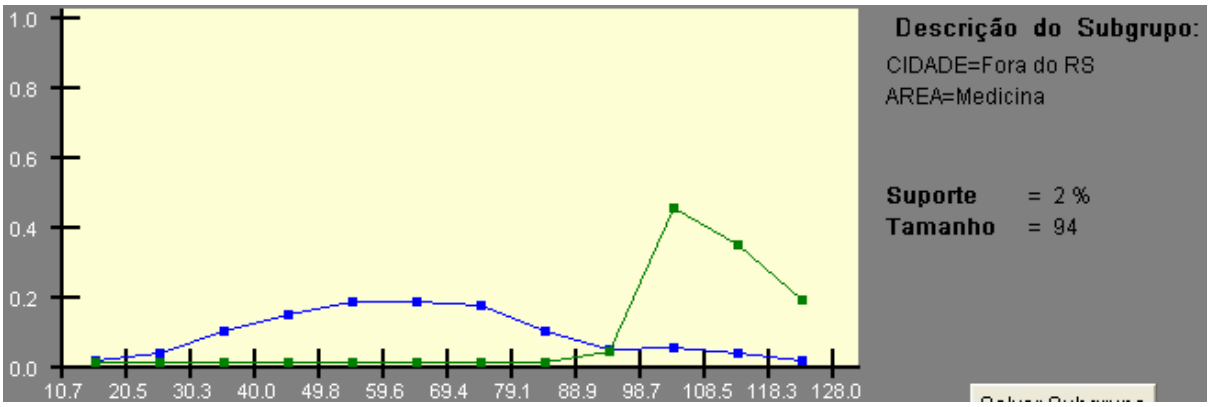

Figura 24 - Subgrupo S1.

Por outro lado, tem-se o subgrupo S2, o qual é composto somente por candidatos que residem em cidades de fora do estado. S2 é o subgrupo com maior desvio padrão, o que leva a concluir que os escores dos candidatos deste subgrupo diferem significativamente entre si, conforme pode ser observado na Fig 25, onde o círculo em vermelho representa o subgrupo em questão.

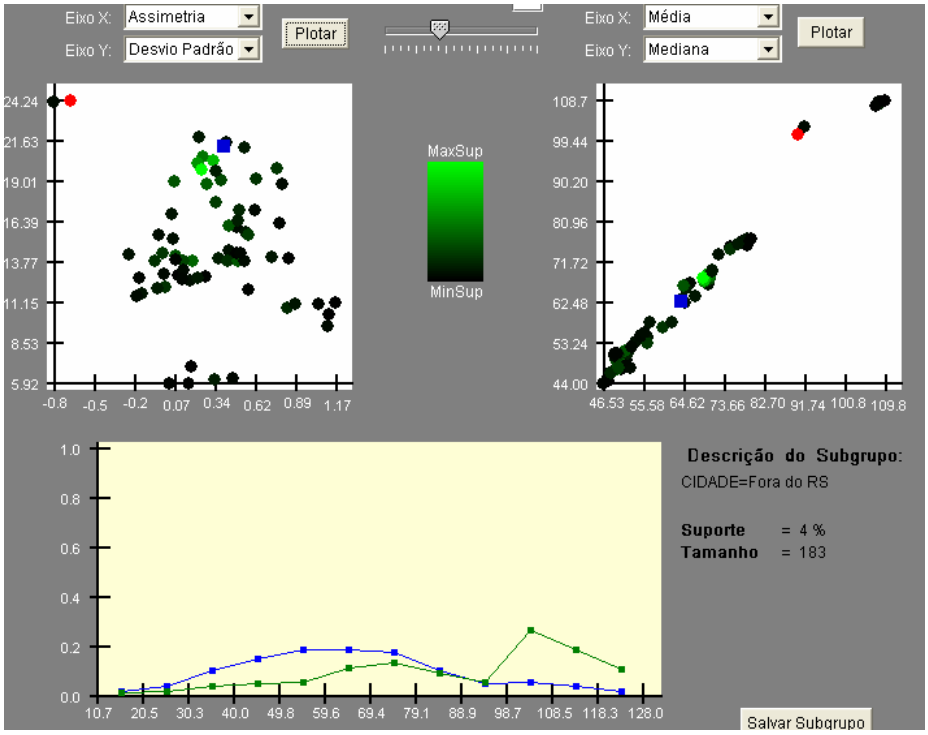

Figura 25 - Subgrupo S2.

Ainda através da Fig. 25, percebe-se uma característica relevante do conjunto de dados em questão, onde, no gráfico presente no canto superior direito da figura, observa-se que a média e a mediana de todos os subgrupos são praticamente iguais, como exceção dos subgrupos que representam candidatos de fora do estado.

Sabendo que S1 compreende 94 candidatos e S2 compreende 183, conclui-se que mais da metade dos candidatos de fora do estado prestam vestibular para medicina. A diferença entre a média e a mediana de S2, assim como seu alto desvio padrão, ocorrem devido a alta freqüência de escores elevados no subgrupo S1.

Ainda em relação aos candidatos ao curso de medicina, percebeu-se através da Fig. 26, a discrepância dos escores destes candidatos em relação ao resto da população. Observa-se que os subgrupos compostos por candidatos da medicina se encontram significativamente mais à direita, no eixo que representa a média, que o restante dos subgrupos.

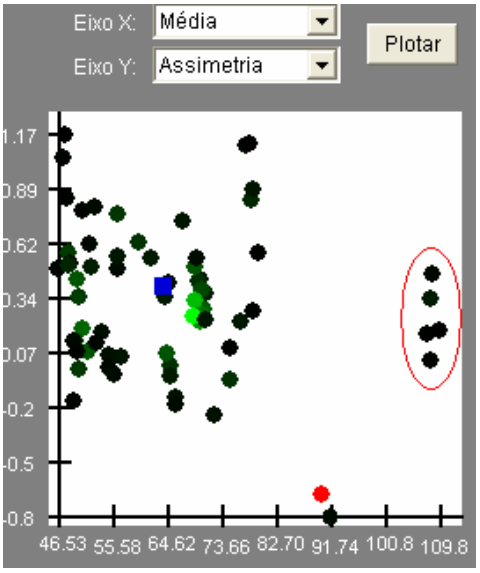

Figura 26 - Subgrupos de candidatos
ao curso de medicina.

Em contra partida, tem-se os subgrupos com escore bruto menor, como é o caso do subgrupo S3, o qual é composto por candidatos a cursos noturnos da área de letras e artes e apresenta uma média de 46,6. Visto que S3 possui desvio padrão baixo (11,1), conclui-se que os escores brutos dos candidatos deste subgrupo não diferem significativamente entre si, fato este que também pode ser observado na Fig. 27.

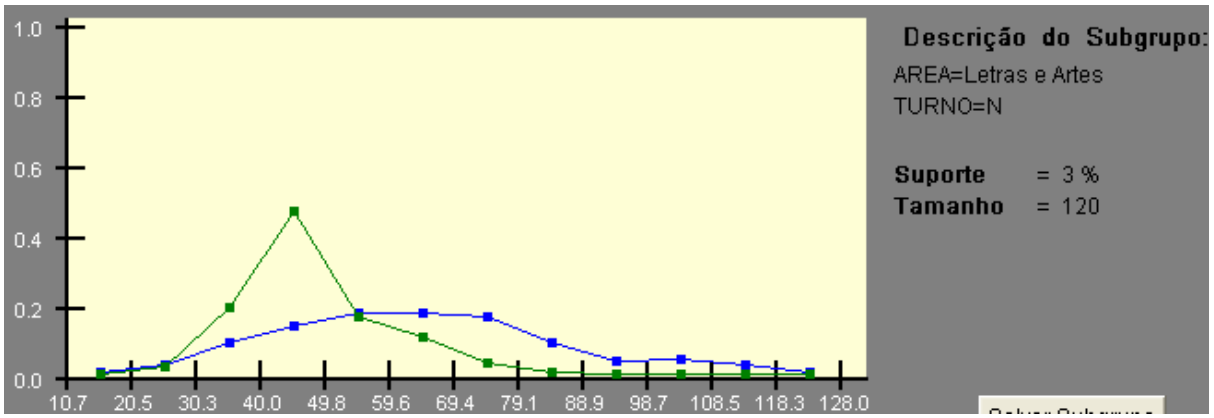

Figura 27 - Subgrupo S3.

Sabendo-se que S3 possui 120 elementos e que um outro subgrupo com a mesma descrição de S3, porém com candidatos residentes na cidade de Pelotas,

possui 98 elementos, conclui-se que a maior procura para cursos noturnos da área de letras é de candidatos que residem em Pelotas.

Observando-se os subgrupos S4, composto por candidatos a cursos noturnos e de tamanho igual a 771, e S5, com a mesma descrição de S4, porém com candidatos de Pelotas e de tamanho igual a 555, conclui-se que a procura por cursos noturnos ocorre predominante por candidatos que residem em Pelotas.

A distribuição de S4 é exibida na Fig. 28, na qual percebe-se que os elementos do subgrupo possuem freqüências mais altas de escores baixos .

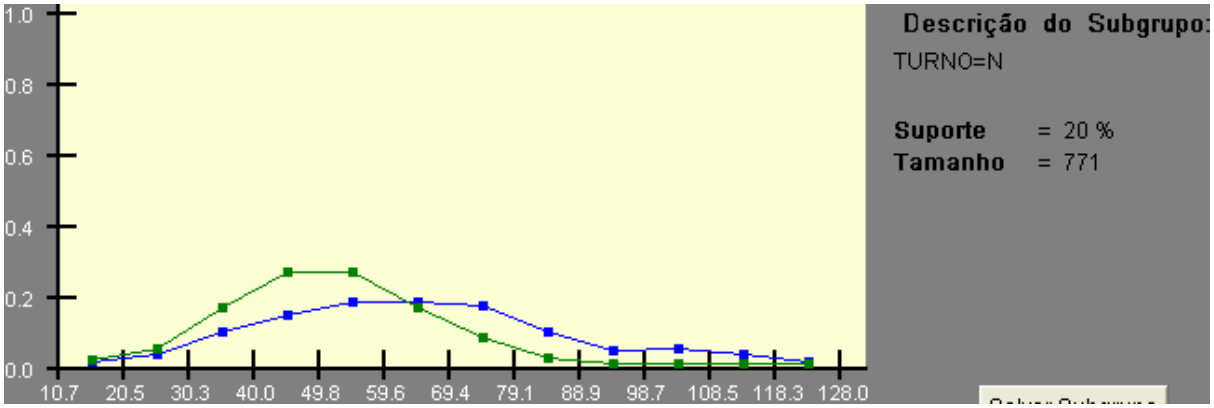

Figura 28 - Subgrupo S4.

Se analisado um outro subgrupo, S5, composto por candidatos a cursos noturnos, porém com idades entre 24 e 27 anos, observa-se que os escores brutos destes candidatos são ainda menores que os do subgrupo S4, pois a média deste é igual a 48,2, enquanto que a desse é igual a 46,9.

Tomando como referência a idade dos candidatos e analisando a Fig. 29, pode-se observar que subgrupos de candidatos com idade entre 28 e 30 anos possuem a média do escore bruto mais baixa que a dos candidatos com idade entre 24 e 27, os quais, por sua vez, possuem média mais baixa que a dos candidatos com idade entre 21 e 23 anos, que, por sua vez, possuem média mais baixa que os entre 18 e 20 anos. Os pontos em vermelho na Fig. 29 representam tais subgrupos, quanto maior for a faixa etária, mais abaixo no eixo Y se encontrará o subgrupo.

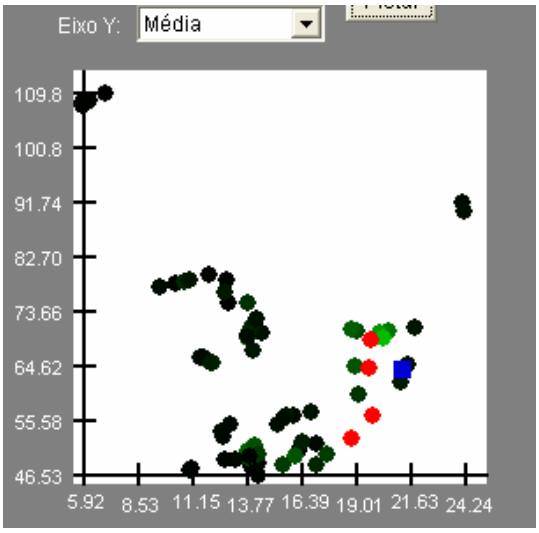

Figura 29 - Subgrupos compostos pela faixa etária dos candidatos.

Tomando como referência a cidade onde os candidatos residem, foi possível concluir que o subgrupo, S6, composto por candidatos oriundos de municípios da região de pelotas é o que possui menor média se comparado ao de outras regiões. Tal subgrupo possui média igual a 55,58, enquanto que toda a população possui uma média igual a 64. Além disso, S6 possui um desvio padrão (15,00) mais baixo que a maioria dos subgrupos, portanto, conclui-se que os escores dos candidatos deste subgrupo não diferem significativamente entre si. A distribuição de S6 é exibida na Fig.30.

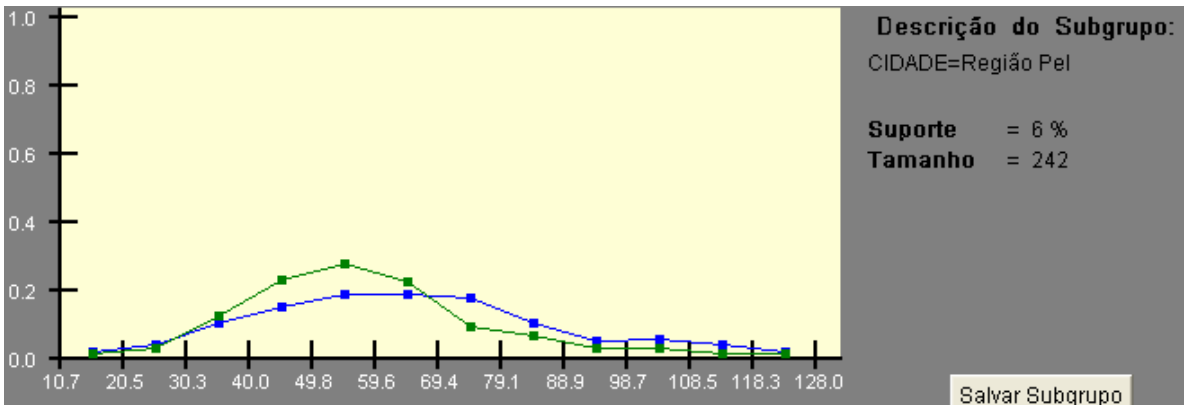

Figura 30 - Subgrupo S6.

Uma outra característica interessante que pôde ser observada no conjunto de dados está relacionada aos subgrupos compostos por candidatos que optaram

por cursos de ciências agrárias. Nestes subgrupos, os escores dos candidatos são quase simétricos em relação à média, pois conforme a Fig. 31, tais subgrupos se encontram em uma mesma região do gráfico, na qual tem-se medidas de assimetria próximas de zero.

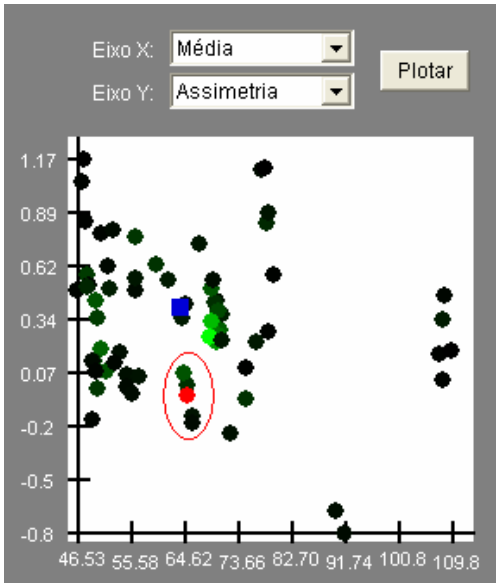

Figura 31 - Subgrupos das ciências agrárias.

A distribuição e descrição do subgrupo selecionado, na figura acima, são exibidas na Fig. 32.

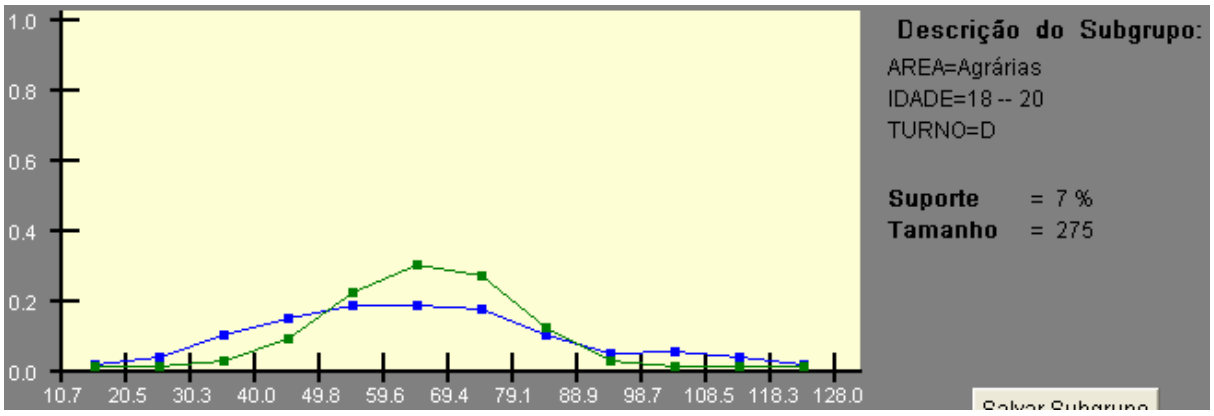

Figura 32 - Distribuição de um subgrupo de candidatos das ciências agrárias.

6.5 Considerações sobre a Utilização da Ferramenta

Através dos resultados obtidos no estudo descrito neste capítulo, provou-se que foi possível utilizar a ferramenta para a realização da descoberta dos subgrupos.

Além disso, pôde-se escolher uma variável numérica para ser a propriedade de interesse, o que permite uma descrição mais detalhada e a identificação de características inerentes à distribuição da propriedade de interesse em uma população. A grande maioria das ferramentas para a descoberta de subgrupos utilizam somente variáveis categóricas, fato que limita uma análise mais detalhada. Por este motivo, as técnicas de visualização utilizadas no pós-processamento de regras de associação são mais eficientes que as técnicas de visualização utilizadas na descoberta de subgrupos.

Por outro lado, por ser uma ferramenta de descoberta de subgrupos, permite-se identificar, com maior facilidade, grupos da população que possuem alguma característica discrepante em relação à mesma. Nela, toda a análise é focada na identificação de grupos com características não similares à toda população. Tal funcionalidade não está presente em outras ferramentas que, na maioria dos casos, fornecem um panorama geral da população.

7 CONCLUSÕES

A vantagem de se ter uma implementação híbrida, abrangendo princípios de regras de associação e descoberta de subgrupos, é a possibilidade de poder aplicar, na visualização de subgrupos, técnicas de visualização utilizadas no pós-processamento de regras de associação. A utilização de tais técnicas proporciona maior facilidade na interpretação de modelos que possuam um elevado número de regras de associação, fator que desencoraja o uso destas na etapa de mineração de dados.

Valendo-se da metodologia exposta na seção 5.1, a ferramenta desenvolvida neste trabalho apresenta como contribuição inovadora alguns aspectos relacionados à interação proporcionada durante a análise. Dentre estes, pode-se citar a filtragem dos subgrupos exibidos de acordo com a medida de suporte definida pelo analista, em tempo de execução, e também a maneira dinâmica com os gráficos de navegação se compõem, onde analista pode escolher, em tempo de execução, as medidas estatísticas que formam os eixos cartesianos dos mesmos.

Fazendo uso das funcionalidades da linguagem de programação Java pôde-se obter uma ferramenta portável e com uma interface gráfica interativa. Através desta, possibilita-se ao analista participar de um processo no qual ele é capaz de encontrar subgrupos com características interessantes quando confrontados com o resto da população, pois, dessa forma, ele pode perceber e/ou compreender melhor a ocorrência de tais valores.

Após o amadurecimento e consolidação da metodologia que a ferramenta desenvolvida utiliza, espera-se contribuir no sentido desta poder ser uma funcionalidade extra disponível em algum um sistema mais amplo de KDD, no qual seja possível direcionar o processo de KDD para a descoberta de subgrupos.

Em relação ao estudo realizado no capítulo 6, foi possível encontrar subgrupos com características significativamente distintas em relação à toda

população. Através dos resultados obtidos no estudo, espera-se poder contribuir, de alguma forma, para uma melhor compreensão de fatores que influenciam o rendimento de certos subgrupos de candidatos no processo seletivo. Porém, é notório que o conjunto de dados utilizado pode ainda ser amplamente explorado. Dessa forma, propõe-se, como trabalho futuro, a realização de uma análise estatística completa em tais dados, aonde se possa ter o envolvimento de especialistas de diferentes áreas do conhecimento.

Ainda como trabalho futuro, sugeri-se a continuação do desenvolvimento da ferramenta para que outras etapas do KDD, além das de mineração e pós-processamento, possam ser incorporadas. Um primeiro passo neste sentido poderia ser em relação à etapa de transformação dos dados, onde se permitiria uma maior flexibilidade na composição dos dados do arquivo de entrada e, inclusive, possibilitando outros formatos de arquivo além do CSV. Um outro passo poderia ser em relação à etapa de representação do conhecimento, pois no momento em que se dispõe dos subgrupos salvos pelo analista, poderia ser gerado, automaticamente, relatórios estatísticos descrevendo os subgrupos salvos.

REFERÊNCIAS

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A., Mining association rules between sets of itens in large databases. ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 1993, **Anais do**... p. 207-216.

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules. INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 1994, Santiago. **Anais do...** Santiago: 1994. p. 487-499.

AZEVEDO, Paulo J. Caren - A Java Based Apriori Implementation for Classification Purposes, Tecnical Report, Universidade do Minho, Portugal, 2003

BORGORI, Vânia. **Algoritmos e Ferramentas de Descoberta de Conhecimento em Bancos de Dados Geográficos**. 2003. Dissertação de Mestrado (Programa de Pós-Graduação em Computação) - Universidade Federal do Rio Grande do Sul, Porto Alegre

BRAUNER, Daniela F. **O Processo de Descoberta de Conhecimento em Banco de Dados:** Um Estudo de Caso Sobre os Dados da UFPel. 2003. Monografia de Conclusão de Curso - Instituto de Matemática e Física, Universidade Federal de Pelotas, Pelotas.

BRANDÃO, Antonio J. S.; MOREIRA, Edson S. Agentes Móveis e Sistemas de Gerenciamento. In: II WORKSHOP EM SEGURANÇA DE SISTEMAS COMPUTACIONAIS. 2002, Búzios. **Anais do...** Búzios: 2002. p. 49-56.

BRIN, S.; MOTWANI, R.; ULLMAN, J. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In: ACM INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA. **Anais do...**1997.

CHAPMAN, Pete; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHARER, C.; WIRTH, R. CRIPS-DM 1.0 Step-by-step data mining guide. 2000. Disponível em: < WWW em http://www.crisp-dm.org> Acesso em: 25 nov. 2005.

DEMŠAR, J.; ZUPAN, B.; LEBAN, G. Orange: From Experimental Machine Learning to Interactive Data Mining. In: 8TH EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES. 2004, Pisa, **Anais do...**Pisa: 2004.

DOMINGUES, M. Aurélio. **Generalização de Regras de Associação**. 2004. Dissertação de Mestrado - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **From Data Mining to Knowledge Discovery**: An Overview. Advances. In Knowledge Discovery and Data Mining. Menlo Park: AAAI Press: 1996. p.11-34.

FRAWLEY, W. J., PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. **Knowledge discovery in databases: An overview**. AI Magazine v.13, n.3, p. 57-70, 1992.

GERSHON, N.; EICK, S.G.; Card, S. **Information Visualization**. vol. 5, n^0. 2 ACM Interactions, 1998. p. 9-15.

GONÇALVES, Léren P. F. **Mineração de Dados em Supermercados: O Caso do Supermercado "Tal"**. 1999. Dissertação de Mestrado (Programa de Pós-Graduação em Administração) - Universidade Federal do Rio Grande do Sul, Porto Alegre.

HAN, J.; FU, Y.; WANG, W.; KOPERSKI, K.; ZAIANE, O. DMQL: A Data Mining Query Language for Relational Databases. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 1996.

HAN, J. P.; Yin, Y. - Mining Frequent Patterns without candidate generation. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2000. **Anais do...** 2000.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. Morgan Kaufmann, 2001.

HAN, Jianchao; CERCONE, Nick; HU, Xiaohua. **An interactive visualization system for mining association rules**. Data mining, rough sets and granular computing, 2002, p.145 -165.

HARRISON, Thomas H. **Intranet Data Warehouse.** São Paulo: Bekerley Brasil, 1998.

HETZLER, B.; HARRIS, W. M.; HAVRE, S.; WHITENY, P. Visualizing the Full Spectrum of Document Relationships, In: FIFTH INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION CONFERENCE, San Francisco, 1998. **Anais do...**San Francisco: 1998.

HIPP, J., C. MANGOLD; U. GÜNTZER; G. NAKHAEIZADEH. Efficient Rule Retrieval and Postponed Restrict Operations for Association Rule Mining. In: 6TH PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2002, Taipei. **Anais do...** Taipei: 2002, p. 52-65.

JAKARTA Commons. Disponível em <http://jakarta.apache.org/commons/> Acesso em: 15 abr. 2006.

JORGE, A. Hierarchical clustering for thematic browsing and summarization of large sets of association rules. In: SIAM SDM 2004, CONFERENCE ON DATA MINING, , 2004, Orlando. **Anais do...** Orlando: SIAM Press, 2004.

JORGE, A.M., AZEVEDO, P.J., PEREIRA F., Distribution Rules with Numerical Properties of Interest. In: 10TH EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES, 2006, Berlin. **Anais do...** Berlin: LNAI Springer-Verlag, 2006

JORGE, A. M., PEREIRA, F, AZEVEDO, P.J. Visual Interactive Subgroup Discovery with numerical properties of interes, In PROCEEDINGS OF DISCOVERY SCIENCE 06, Barcelona, 2006, **Anais do....** Barcelona: LNAI, Springer-Verlag, 2006

KAVSEK, Branko; LAVRAC, Nada; JOVANOSKI, V. APRIORI-SD: Adapting Association rule learning to subgroup discovery. IN: THE FIFTH INTERNATIONAL SYMPOSIUM ON INTELLIGENT DATA ANALYSIS, IDA 2003, Berlim. **Anais do...** Berlim, 2003.

KAVSEK, Branko; LAVRAC, Nada. Analysis of example weighting in subgroup discovery by comparison of three algorithms on a real-life data set. In: EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES. **Anais do...** Pisa, 2004.

KDDNUGGETS, Software for Data Mining and Knowledge Discovery – Disponível em: <http://www.kdnuggets.com/software/index.html>. Acesso em: 5 mai. 2006.

KLÖSGEN, W. Exploration of simulation experiments by discovery. In: AAAI-94 WORKSHOP ON KNOWLEDGE DISCOVERY IN DATABASES. 1994. **Anais do...**AAAI Press, 1994. p. 251-262.

KLÖSGEN, Willi. Applications and Research Problems of Subgroup Mining. In: 11TH INTERNATIONAL SYMPOSIUM ON FOUNDATIONS OF INTELLIGENT SYSTEMS, 1999, Warsaw. **Anais do...** Warsaw: 1999. p.1-15.

KLÖSGEN, W. ; MAY, M. Census Data mining - an application. In: THE 6TH EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES (PKDD), 2002, Helsinki. **Anais do...** Helsinki: 2002

KRALJ, Petra; LAVRAC, Nada; ZUPAN, Blaz; GRAMBERGER, Dragan. Experimental Comparison of Three Subgroup Discovery Algorithms: Analysing Brain Ischemia Data. In: 8TH INTERNATIONAL MULTICONFERENCE INFORMATION SOCIETY, JOZEF STEFAN INSTITUTE, 2005, Ljubljana. **Anais do...** Ljubljana: 2005. p. 220-223.

LAVRAC, N.; KAVSEK, B.; FLACH, P.; TODOROVSKI, L. Subgroup discovery with CN2-SD. **Journal of Machine Learning Research**, 2004.

LIU, B., HSU, W. E MA, Y. Pruning and Summarizing the Discovered Associations. In: 5TH ACM INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 1999. **Anais do....** 1999.

MANNILA, H.; TOIVONEN, H.; VERKAMO, A.I. - Efficient Algorithms for Discovering Association Rules. In: AAAI WORKSHOP ON KNOWLEDGE DISCOVERY IN DATABASES, EDS. USAMA M. FAYYAD AND RAMASAMY UTHURUSAMY, 1994, Washington. **Anais do...**Washington: 1994. p. 181-192.

MELANA, Edson Augusto. **Pós-Processamento de Regras de Associação**. 2004. Teste de Doutorado - Instituto de Ciências Matemáticas e Computação, Universidade de São Paulo, São Carlos.

MEO, R.; PSAILA, G.; CERI, S. A new SQL-like operator for mining association rules. In: 22ND INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 1996. **Anais do...** 1996. p. 122-123.

MORZY, T.; M. Zakrzewicz. SQL-Like Language For Database Mining. In: 1ST INTERNATIONAL CONFERENCE ON ADVANCES IN DATABASES AND INFORMATION SYSTEMS, 1997, Petersburg. **Anais do...** Petersburg: 1997. p. 311-317.

NEVES, João P. **Ambiente de Pós-processamento para Regras de Associação. Porto: UP**, 2003. Dissertação de Mestrado - Faculdade de Economia, Universidade do Porto, Porto.

ONG, Kian-Huat; ONG, Kok-Leongg; NG, Wee-Keong; LIM, Ee-Peng. CrystalClear: Active Visualization of Association Rules. ICDM2002 WORKSHOP - INTERNATIONAL WORKSHOP ON ACTIVE MINING. 2002, Maebashi. **Anais do...**, Maebashi: 2002.

OSJava – Open Sourced Java. Disponível em: < http://www.osjava.org/. > Acesso em: 13 abr. 2006.

PARK, J. S.; CHEN, M.; YU, P. S. Using a Hash-Based Method with Transaction Trimming for Mining Association rules. In: IEEE Transactions on Knowledge and Data Engineering, Vol. 9, N^{o} 5, 1997.

PEREIRA, F. **Descoberta de subgrupos com regras de associação.** Dissertação (Mestrado em Análise de Dados e Sistemas de Apoio à Decisão) - Faculdade de Economia do Porto, Universidade do Porto. Disponível em: (www.fep.up.pt/cursos/mestrados/madsad). Acesso em: 12 jan. 2006

PITONI, Rafael M. **Mineração de Regras de Associação nos Canais de Informação do Direto.** 2002. Monografia de Conclusão de Curso - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

REZENDE, S. O.; PUGLIESE, J. B.; MELANDA, E. A.; PAULA, M. F. Mineração de Dados. In: Sistemas Inteligentes: Fundamentos e Aplicações. . Volume 1. Barueri: Editora Manole, 2003. p. 307-335.

ROMÃO, W., FREITAS, A. A., PACHECO, R. C. S. Uma Revisão de Abordagens Genético-Difusas para Descoberta de Conhecimento em Banco de Dados. In: Acta Scientiarum, v.22, n.5, 2000. p.1347 – 1349,.

SANTOS, RODRIGO G. **Utilização de Técnicas de Data Mining na Busca de Conhecimento na Web. Pelotas: UFPel**, 2000. Monografia de Conclusão de Curso - Instituto de Matemática e Física, Universidade Federal de Pelotas, Pelotas.

SEIFERT, Jeffrey W. Data Mining: an overview. Congressional Research Service (CRS) Report RL31798. Washington, EUA, 2005.

SRIKANT, R; AGRAWAL, R. MINING. Generalized association rules. Future Generation ComputerSystems, v.13, p.161–180, 1997.

SRIKANT, R. Association rules: Past, present and future. In: INTERNATIONAL WORKSHOP ON CONCEPT LATTICE-BASED THEORY, METHODS AND TOOLS FOR KNOWLEDGE DISCOVERY IN DATABASES, 2001. **Anais do...** 2001.

TOIVONEN, H., M. Klemettinen, P. Ronkainen, K. Hätönen e H. Mannila. Pruning and Grouping Discovered Association Rules, In: MLNET WORKSHOP ON STATISTICS, MACHINE LEARNING AND DISCOVERY IN DATABASES. **Anais do...** 1995.

WEKA, Waikato Environment for Knowledge Analysis. Disponível em: < http://www.cs.waikato.ac.nz/ml/weka/ >. Acesso em: 14 fev. 2006.

WEBER, I. Pruning strategies for discovery of generalized and quantitative association rules. In: WORKSHOP ON KNOWLEDGE DISCOVERY AND DATA MINING, 1998. **Anais do...** 1998.

WONG, Pak Chung; WHITNEY, Paul; THOMAS, Jim. Visualizing Association Rules for Text Mining. In: IEEE INFORMATION VISUALIZATION, 1999. **Anais do...** 1999.

ZAKI, M. J.; PARTHASARATHY, S.; LI, W. - New Algorithms for Fast Discovery of Association Rules. In PROC. OF THE 3TH ACM INTL. CONFERENCE. ON KNOWLEDGE DISCOVERY AND DATA MINING, 1997. **Anais do...** 1997.

ZHANG, H. **Mining And Visualization Of Association Rules Over Relational DBMSs**. 2000. Tese de doutorado - Department of Computer and Information Science and Engineering, The University of Florida, Florida.