UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

Mineração de Dados Utilizando Aprendizado Não-Supervisionado: um estudo de caso para bancos de dados da saúde

por

MIRIAM LÚCIA CAMPOS SERRA DOMINGUES

Dissertação submetida à avaliação,
como requisito parcial para a obtenção do grau de Mestre
em Ciência da Computação

Prof. Dr. Paulo Martins Engel
Orientador

Porto Alegre, março de 2003

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Domingues, Miriam Lúcia Campos Serra
Mineração de Dados Utilizando Aprendizado Não-Supervisionado: um estudo de caso para bancos de dados da saúde / por Miriam Lúcia Campos Serra Domingues. – Porto Alegre: PPGC da UFRGS, 2003.
127 p.: il.

Dissertação (Mestrado) – Universidade Federal do Rio Grande do Sul. Instituto de Informática. Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2003. Orientador: Engel, Paulo Martins.

1. Inteligência Artificial. 2. Descoberta de Conhecimento em Bases de Dados. 3. Mineração de Dados. I. Engel, Paulo Martins. II. Título.

Agradecimentos

A Deus, em quem busco forças para seguir adiante.

Ao meu orientador, Prof. Dr. Paulo Martins Engel, por todo o apoio que me deu para a realização deste trabalho, pelos ensinamentos, pela dedicação, pela atenção e amizade.

À equipe da SES, pela receptividade e pelas informações prestadas.

Ao meu marido, José Luiz, e aos meus filhos, Rafael e Rodrigo, pelo amor, carinho e incentivo e pela compreensão nos meus períodos de ausência.

À minha mãe, pelo seu imenso amor e pela ajuda em todos os momentos.

Ao meu pai, irmãos, cunhados e sobrinhos que me incentivaram durante a realização do curso.

À Profa. Ana Luiza Lima e ao Dr. Paulo Amorim, ex-diretora e diretor atual do Centro de Ciências da Saúde da UFPA, pela confiança em mim depositada e pelas condições que me deram para realizar o mestrado.

Ao Prof. Antonio Sampaio, pela iniciativa de realizar o MINTERCC.

A todos os amigos do MINTERCC, em especial à Carolina, pela sua amizade, companhia e ajuda nos dias que passamos em Porto Alegre; Gisa, pela amizade e alegrias compartilhadas em Belém e em Porto Alegre; Ana Carla, Raquel e Constância, pelas horas que passamos juntas, compartilhando alegrias e preocupações.

Aos amigos e colegas da UFPA e da UFRGS, pelo apoio e incentivo.

À CAPES, Universidade Federal do Pará e Universidade Federal do Rio Grande do Sul, pela realização do Mestrado Interinstitucional em Ciência da Computação e pelos recursos concedidos.

Sumário

Lista de Abreviaturas

AIH	Autorização de Internação Hospitalar
API	Interface do Programa de Aplicação
AVC	Acidente Vascular Cerebral
BD	Banco de Dados
BMU	Best-Matching Unit
CID	Cadastro Internacional de Doenças
CRISP-DM	CRoss Industry Standard Process Model for Data Mining
CRS	Coordenadoria Regional de Saúde
DCBD	Descoberta de Conhecimento em Bases de Dados
IBM	International Business Machines Corporation
ID	Identificador
IM	Intelligent Miner
KDD	Knowledge Discovery in Databases
MD	Mineração de Dados
MS	Ministério da Saúde
NCC	New Condorcet Criterion
OLAP	On-Line Analytical Processing
SES	Secretaria Estadual de Saúde do Rio Grande do Sul
SIH	Sistema de Informações Hospitalares
SMS	Secretaria Municipal de Saúde do Rio Grande do Sul
SOM	Self-Organizing Maps
SQL	Structured Query Language
SUS	Sistema Único de Saúde
UFRGS	Universidade Federal do Rio Grande do Sul

Lista de Figuras

Lista de Tabelas

Resumo

A mineração de dados constitui o processo de descoberta de conhecimento interessante, com a utilização de métodos e técnicas que permitem analisar grandes conjuntos de dados para a extração de informação previamente desconhecida, válida e que gera ações úteis, de grande ajuda para a tomada de decisões estratégicas.

Dentre as tarefas de mineração de dados, existem aquelas que realizam aprendizado não-supervisionado, o qual é aplicado em bases de dados não-classificados, em que o algoritmo extrai as características dos dados fornecidos e os agrupa em classes. Geralmente, o aprendizado não-supervisionado é aplicado em tarefas de agrupamento, que consistem em agrupar os dados de bancos de dados volumosos, com diferentes tipos de dados em classes ou grupos de objetos que são similares dentro de um mesmo grupo e dissimilares em diferentes grupos desses bancos de dados, de acordo com alguma medida de similaridade. Os agrupamentos são usados como ponto de partida para futuras investigações.

Este trabalho explora, mediante a realização de um estudo de caso, o uso de agrupamento como tarefa de mineração de dados que realiza aprendizado não-supervisionado, para avaliar a adequação desta tecnologia em uma base de dados real da área de saúde. Agrupamento é um tema ativo em pesquisas da área pelo seu potencial de aplicação em problemas práticos.

O cenário da aplicação é o Sistema de Informações Hospitalares do SUS, sob a gestão da Secretaria Estadual de Saúde do Rio Grande do Sul. Mensalmente, o pagamento de um certo número de internações é bloqueado, uma vez que a cobrança de internações hospitalares é submetida a normas do SUS e a critérios técnicos de bloqueio estabelecidos pela Auditoria Médica da SES para verificar a ocorrência de algum tipo de impropriedade na cobrança dos procedimentos realizados nessas internações hospitalares.

A análise de agrupamento foi utilizada para identificar perfis de comportamentos ou tendências nas internações hospitalares e avaliar desvios ou *outliers* em relação a essas tendências e, com isso, descobrir padrões interessantes que auxiliassem na otimização do trabalho dos auditores médicos da SES. Buscou-se ainda compreender as diferentes configurações de parâmetros oferecidos pela ferramenta escolhida para a mineração de dados, o *IBM Intelligent Miner*, e o mapeamento de uma metodologia de mineração de dados, o CRISP-DM, para o contexto específico deste estudo de caso.

Os resultados deste estudo demonstram possibilidades de criação e melhora dos critérios técnicos de bloqueio das internações hospitalares que permitem a otimização do trabalho de auditores médicos da SES. Houve ainda ganhos na compreensão da tecnologia de mineração de dados com a utilização de agrupamento no que se refere ao uso de uma ferramenta e de uma metodologia de mineração de dados, em que erros e acertos evidenciam os cuidados que devem ser tomados em aplicações dessa tecnologia, além de contribuírem para o seu aperfeiçoamento.

Palavras-chave: descoberta de conhecimento, mineração de dados, aprendizado não-supervisionado, agrupamento.

TITLE: "DATA MINING WITH UNSUPERVISED LEARNING: A CASE STUDY FOR A HEALTHCARE DATABASE"

Abstract

Data mining is a discovery process of interesting knowledge. It uses methods and techniques that allow to analyze large amounts of data to extract information previously unknown, valid and that generates useful actions, which are of great help for strategic decision making.

Within data mining tasks, there are those based on unsupervised learning, which are applied to unclassified databases, where the algorithm extracts the characteristics from data provided and clusters them in classes. In general, unsupervised learning is applied to clustering tasks, which consist of splitting large databases, with different types of data in classes or object groups that are similar within the same group and different in different groups. It is made according to some similarity measures. The clusters are used as starting points for future investigations.

The present work explores, through a case study, clustering using as a data mining task that accomplishes an unsupervised learning, in order to evaluate the suitability of this technology in a real healthcare area database. Clustering is an active topic of research in practical problems solving in this area.

The scenery is the Public Hospital Information System (*SUS - Sistema Único de Saúde*), ruled by the Healthcare Department of the Rio Grande do Sul state (*SES - Secretaria Estadual da Saúde*). Monthly, payment of a number of inpatient procedures is stopped, once they are submitted to the SUS rules and to technical criteria. These payments are prevented by the SES Medical Auditors to verify the occurrence of some improper charging procedures carried out in hospitalizations.

In this work, clustering analysis was used to identify behaviors or trends in hospitalizations and to evaluate deviations or outliers with relation to these trends and thus discovering interesting patterns that help to optimize medical auditors' work. We also tried to understand different parameter settings the tool for data mining offered, the *IBM Intelligent Miner*, and mapping of a data mining methodology, CRISP-DM, for the specific context of this case study.

Results have shown that there are possibilities of creation and improvement of technical criteria in hospitalizations that make possible to optimize the work of medical auditors from SES. There were gains in understanding data mining technology with the use of clustering in what concerns the use of data mining tools and methodology, where goals and mistakes make evident what kind of care should be taken when this technology is applied, besides contributing for their refinement.

Keywords: Knowledge discovery, data mining, unsupervised learning, clustering.

1 Introdução

1.1 Motivação

A mineração de dados (MD) ou *data mining* é definida, ultimamente, como um processo de descoberta de padrões[1] em quantidades substanciais de dados, de forma automática ou, na maioria das vezes, semi-automática, para a extração de informação previamente desconhecida, válida e que gera ações úteis, em que os padrões descobertos são significativamente vantajosos para a tomada de decisões estratégicas [CAB 97, WIT 99]. Essas características têm atraído uma boa parte das atenções da indústria da informação, em que MD é apresentada como um resultado da evolução natural da tecnologia da informação [HAN 2001].

Esse processo envolve o uso de diversas técnicas, dentre as quais, as utilizadas na identificação de agrupamentos ou *clustering*. A aplicação de técnicas de agrupamento caracteriza o uso de um método de aprendizado não-supervisionado, em que existe maior autonomia do algoritmo para extrair características dos dados fornecidos, agrupando-os em classes[2] (*clusters*).

As técnicas de identificação de agrupamentos têm motivado muitas pesquisas na área, mas ainda não existem métodos totalmente versáteis e eficientes de agrupamento. Essa área tem atraído grande interesse, não só por seu potencial de aplicação em problemas práticos, mas também por seu relacionamento com a inteligência humana, uma vez que existem evidências de que a formação de conceitos no ser humano e mesmo em animais, durante toda a vida, envolve processos neurais de agrupamento [COS 2001].

O problema de agrupamento consiste no emprego de técnicas para formar, em um banco de dados (BD), grupos de registros similares, que compartilham um certo número de propriedades e, dessa forma, são considerados homogêneos. Os diferentes grupos desse banco de dados apresentam alta heterogeneidade, isto é, são dissimilares entre si de acordo com alguma medida de similaridade.

Assim, o particionamento de BDs volumosos, com diferentes tipos de dados, em sub-populações homogêneas de registros relacionados, resulta na obtenção de padrões de comportamento desses bancos de dados. Um exemplo disso é um banco de dados dividido em dois grupos, em que um deles é rotulado por "*internações hospitalares de hospital de grande porte*, *por AVC agudo*, *em pacientes de 70 anos*" e o outro, "*internações hospitalares por septicemia neonatal, no valor total de R$550,00, com até*

---

[1] Um *padrão* é qualquer entidade física ou simbólica, passível de ser atribuída a uma categoria ou classe, que representa um subconjunto de dados do banco de dados original. Reconhecimento de padrões, em termos gerais, é a ciência que compreende a identificação ou classificação de medidas de informação em categorias, as quais têm a característica de representar entidades ou padrões de informação que apresentam similaridades [VAS 99, CAE 2002].

[2] Classes se referem a padrões gerados de acordo com variáveis, cujo objeto de interesse não é uma descrição detalhada de todas as variáveis, mas sim a consideração de alguns de seus grupos que exibem propriedades específicas. Diz-se que esses padrões constituem uma *classe*, cuja definição varia de acordo com o problema. [GRI 2002].

3 dias de permanência". Esses grupos receberão tratamentos adequados conforme suas características [CAB 97].

Agrupamento é utilizado, freqüentemente, como um dos primeiros passos na análise dos dados feita pela MD para a identificação de grupos de registros relacionados, os quais podem representar classes potenciais, que podem ser usadas como ponto de partida para a exploração de outros relacionamentos. Possibilita a identificação de regiões densas e esparsas que leva à descoberta de padrões de distribuições abrangentes e correlações interessantes entre os atributos de dados. Uma vez identificados os agrupamentos, outros métodos de mineração de dados podem ser aplicados de forma a expressar o significado desses agrupamentos [HAN 2001].

A análise de agrupamentos é usada, largamente, em diversas aplicações, como por exemplo, reconhecimento de padrões, análise de dados, processamento de imagens e pesquisa de mercado. Em MD, os esforços se concentram em achar métodos para essa análise que sejam efetivos e eficientes em grandes bases de dados. Os temas de pesquisa atuais focalizam a escalabilidade dos métodos de agrupamento, a eficácia de métodos para agrupar tipos e formas de dados complexos, técnicas de agrupamento de alta-dimensionalidade e métodos para agrupamento de dados categóricos e numéricos misturados em grandes bases de dados [AGR 98, GRA 98, HAN 2001].

Com este trabalho, pretende-se explorar, pelo emprego de ferramentas de agrupamentos, uma metodologia de utilização das técnicas relacionadas, adequada a uma base de dados real, com grande volume de dados, alta dimensionalidade e tipos de dados diferentes.

O cenário para esta aplicação será o Sistema de Informações Hospitalares do SUS (SIH/SUS), sob a gestão da Secretaria Estadual de Saúde do Rio Grande do Sul (SES), o qual possui uma enorme quantidade de dados sobre a movimentação das internações hospitalares no RS. Mensalmente, o pagamento de um certo número de internações é bloqueado e é preciso verificar se determinados procedimentos são passíveis de serem realizados ou se está ocorrendo algum tipo de impropriedade nessas internações. Busca-se, portanto, a otimização do trabalho dos auditores médicos da SES com a utilização de mineração de dados e técnicas e métodos de agrupamento.

1.2 Objetivos

O objetivo geral deste trabalho é a avaliação de técnicas atuais para a MD, com ênfase na utilização de aprendizado não-supervisionado em um banco de dados complexo do mundo real.

Os objetivos específicos são: avaliar o emprego de técnicas de mineração de dados para a formação de agrupamentos, com a finalidade de identificar as tendências das internações hospitalares representadas pelos padrões médios de procedimentos de interesse e avaliar os desvios em relação a essas médias, com base em impropriedades detectadas no sistema estudado; a utilização de diferentes configurações de parâmetros oferecidas pela ferramenta de MD e uma análise comparativa dos resultados; o mapeamento de uma metodologia de um contexto genérico para um contexto

especializado para a mineração de padrões baseada em agrupamentos, adequada aos dados do SIH/SUS/SES.

1.3 Descrição das Atividades

Este estudo consistiu da realização das seguintes etapas: revisão bibliográfica aprofundada sobre descoberta de conhecimentos em bases de dados (DCBD), mineração de dados e *clustering*; entendimento do domínio da aplicação por intermédio de reuniões com os especialistas da saúde, bem como pela leitura de relatórios e manuais do sistema de internações hospitalares; aquisição dos dados para a análise, os quais foram disponibilizados pela SES; escolha de ferramentas para aplicação das técnicas estudadas; estudo de caso, que consistiu na aplicação de técnicas de MD para a identificação de agrupamento sobre o banco de dados da saúde, com a realização de diversos experimentos, cujos resultados foram analisados e validados pelos especialistas da saúde.

1.4 Organização do Texto

Este trabalho está dividido em seis capítulos:

- Neste capítulo, são apresentados a motivação, os objetivos, a descrição das atividades e a organização do texto.

- No capítulo 2, é apresentado um estudo introdutório sobre DCBD e MD, incluindo diversos itens importantes que devem ser considerados sobre o assunto.

- No capítulo 3, é feita uma revisão geral sobre a descoberta de agrupamentos em MD.

- No capítulo 4, é apresentada uma metodologia para a MD com a utilização de agrupamento.

- No capítulo 5, é apresentado o estudo de caso, com a descrição de todas as etapas do processo de MD para o banco de dados da saúde do SIH/SUS/SES.

- No capítulo 6, são apresentados as conclusões, limitações, contribuições e trabalhos futuros.

2 Descoberta de Conhecimento e Mineração de Dados

Na década de 1980, o grande avanço em tecnologias de hardware dos computadores e de mídias de armazenamento digital facilitou, significativamente, a coleta e armazenamento de dados. Com esses avanços, surgiu também a necessidade de se utilizar novas técnicas e ferramentas que facilitassem a análise de dados e a tomada de decisões, uma vez que os métodos tradicionais utilizados para esse fim apenas criavam relatórios informativos, mas não conseguiam extrair conhecimento importante desses dados. Essas técnicas e ferramentas passaram a ser objetos de estudos de uma área ampla de pesquisas denominada Descoberta do Conhecimento em Banco de Dados (DCBD) ou *Knowledge Discovery in Databases* (KDD) [FAY 96].

A DCBD tornou-se de interesse em diversos campos de pesquisa, tais como aprendizado de máquina, reconhecimento de padrões, banco de dados, estatística, inteligência artificial, aquisição de conhecimento para sistemas especialistas e visualização de dados [FAY 96].

O termo DCBD surgiu em 1989 e substituiu todos os termos antigos que tinham como objetivo encontrar padrões e similaridades em dados brutos, nessa época. Foi adotado rapidamente nas áreas de inteligência artificial e aprendizado de máquina e passou a ser usado para cobrir o processo completo de extração de conhecimento de banco de dados. Nesse contexto, as palavras *mineração de dados* foram usadas para a etapa do processo em que os algoritmos de mineração eram aplicados (Figura 2.1). Esta interpretação foi formalizada durante a Primeira Conferência Internacional em KDD, realizada em Montreal, em 1995 [CAB 97].

2.1 Processo de descoberta de conhecimento

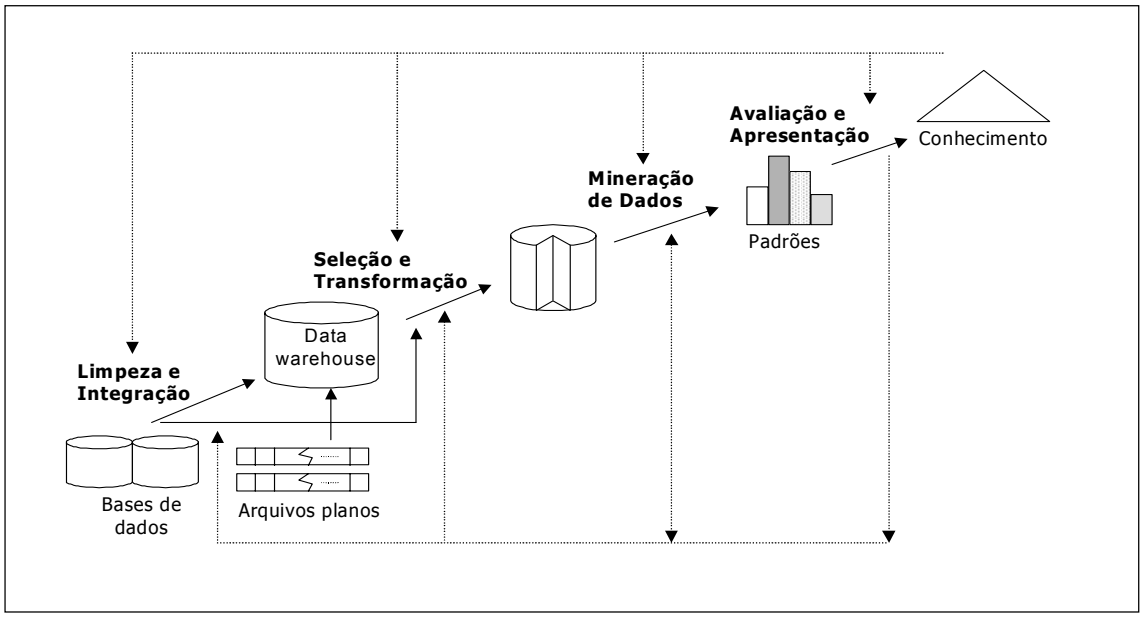


FIGURA 2.1 – Mineração de dados como uma etapa no processo de DCBD.
Fonte: HAN, 2001. p. 6.

O processo de DCBD, que neste trabalho será considerado como sinônimo de processo de MD, conforme se explica na Seção 2.2, consiste em uma seqüência iterativa de etapas, como se observa na Figura 2.1. É um processo iterativo por permitir o retorno entre as diversas etapas, uma vez que qualquer etapa pode produzir novas percepções, as quais permitem melhorar as etapas desenvolvidas previamente. As primeiras etapas formam a base das etapas finais e devem ser bem feitas a fim de evitar problemas que prejudiquem o sucesso da aplicação [CAB 97, CHA 99].

O conhecimento descoberto será avaliado de acordo com os objetivos da aplicação, os quais são determinantes para o projeto inicial e orientam todo o processo de DCBD. As etapas do processo são discriminadas segundo várias abordagens de pesquisadores da área, mas podem ser resumidas em três etapas principais: o pré-processamento, a mineração de dados e o pós-processamento. Estas etapas são descritas brevemente a seguir, segundo Cabena [CAB 97] e Han [HAN 2001]:

2.1.1 Pré-processamento

a) *Determinação dos objetivos da aplicação*: definição clara do problema. É essencial em um projeto de DCBD, o qual, apesar de parecer intuitivo e simples, dificilmente será bem sucedido se for realizado por si só.

b) *Limpeza de dados*: consiste na remoção de ruídos e inconsistência dos dados.

c) *Integração de dados*: etapa em que fontes de dados múltiplas podem ser combinadas.[3]

d) *Seleção de dados*: os dados relevantes para a aplicação de MD são identificados e reunidos, formando um subconjunto do banco de dados.

e) *Transformação de dados*: referente à transformação ou consolidação de dados em formas apropriadas para a mineração, como, por exemplo, pela realização de operações de sumarização ou agregação.[4]

2.1.2 Mineração de dados

f) *Mineração de dados*: processo essencial em que métodos inteligentes são aplicados com a finalidade de extrair padrões dos dados.

2.1.3 Pós-processamento

g) *Avaliação de padrões*: identificação de padrões realmente interessantes que representem conhecimento baseado em algumas medidas de interesses.

---

[3] Na indústria de informação, é comum realizar a limpeza e a integração de dados como uma etapa de pré-processamento, em que os dados resultantes são armazenados em um *data warehouse* [HAN 2001].

[4] "Algumas vezes, a transformação e consolidação de dados são realizadas antes do processo de seleção de dados, particularmente no caso de *data warehousing*" [HAN 2001].

h) *Apresentação do conhecimento*: técnicas de visualização e representação de conhecimento são usadas para apresentar o conhecimento minerado ao usuário.

Apesar dos avanços recentes em tecnologia, estas etapas ainda não são totalmente autônomas e requerem um trabalho intensivo por parte da equipe envolvida no processo. Além disto, demandam pesos diferentes com relação ao tempo e esforço gastos em cada etapa. A preparação de dados, por exemplo, demanda cerca de 60% do esforço total gasto e é crucial para a qualidade final dos resultados, enquanto que a mineração de dados consome cerca de 10% desse esforço. Na fase de pós-processamento, é recomendável a participação de um especialista da área de domínio da aplicação, a fim de solucionar questões específicas que possam influir na análise [CAB 97].

2.2 Mineração de dados

Atualmente, não se refere mais à MD apenas como a uma etapa no processo de DCBD, que envolve a aplicação de métodos para a extração de padrões, mas sim, de uma forma mais abrangente, como um sinônimo de DCBD. Tal fato ocorre devido ao interesse crescente de vendedores em tecnologia da informação e a pressão comercial e popular na área [CAB 97].

Assim, MD passou a ser definida como "*a extração de informação implícita, previamente desconhecida e potencialmente útil dos dados*" [WIT 99].

Uma outra definição, que adota uma visão mais ampla sobre a funcionalidade de mineração de dados é a seguinte [HAN 2001]:

> "*Data Mining é o processo de descoberta de conhecimento interessante de grandes quantidades de dados armazenados tanto em bancos de dados, como em data warehouses ou outros repositórios de informações*".

Com base nessa idéia, a arquitetura de um sistema típico de MD (Figura 2.2) pode ter os seguintes componentes principais, segundo Han [HAN 2001]:

- *Banco de dados, data warehouse*[5]*, ou outro repositório de informação*: diz respeito a um ou a um conjunto de bancos de dados, *data warehouses*, planilhas ou outros tipos de repositórios de informações, sobre os quais podem ser realizadas técnicas de limpeza e integração de dados.

- *Servidor de banco de dados ou servidor de data warehouse*: este servidor é responsável por trazer os dados relevantes, baseado nas requisições do usuário para a mineração de dados.

[5] *Data warehouse* - É um repositório para armazenamento de dados de fontes múltiplas, de longo prazo, organizado de forma a facilitar o gerenciamento de tomada de decisão. Os dados são armazenados sob um esquema unificado e são tipicamente sumarizados. Os sistemas de *data warehouse* possuem algumas capacidades de análises de dados, referidas coletivamente como OLAP (*On-Line Analytical Processing*) [HAN 2001].

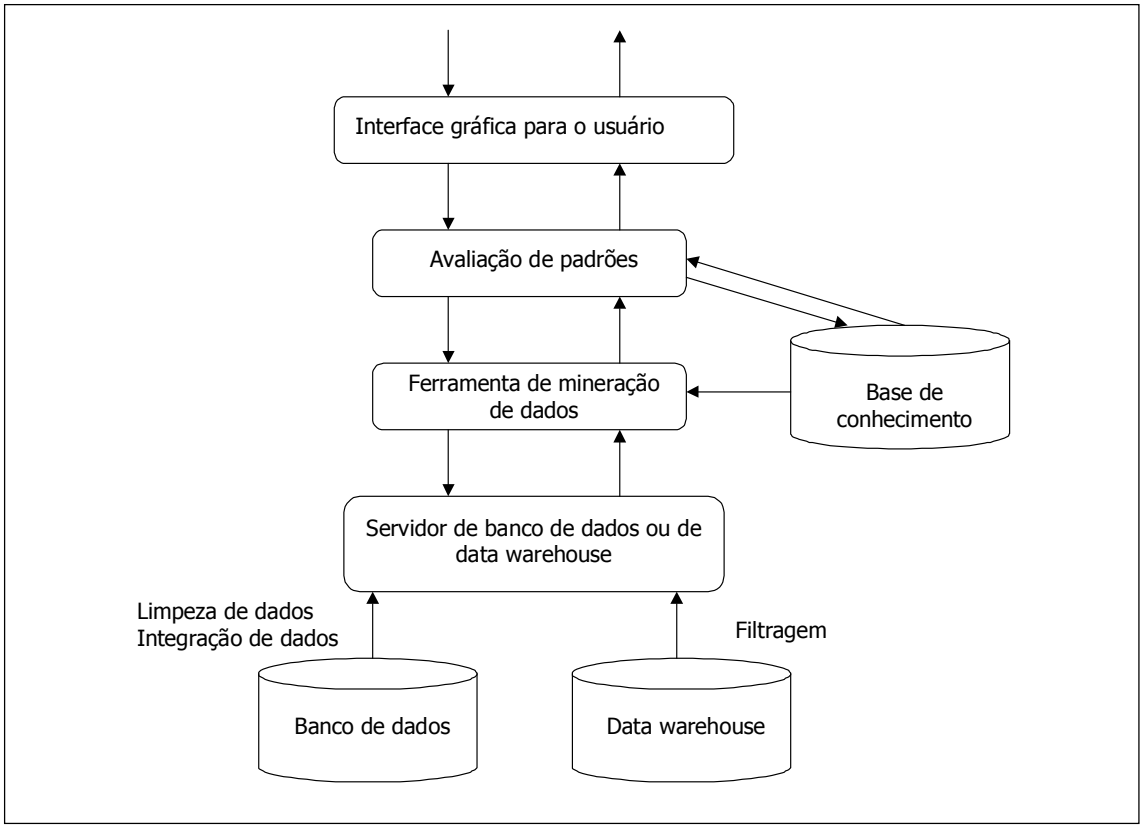


FIGURA 2.2 – Arquitetura de um sistema típico de mineração de dados.
Fonte: HAN, 2001. p. 8.

- *Base de conhecimento*: é o conhecimento do domínio que é utilizado para guiar a pesquisa ou avaliar o interesse dos padrões resultantes. Esse conhecimento pode incluir: hierarquias conceituais, usadas para organizar atributos ou valores de atributos em níveis diferentes de abstração; suposições do usuário, que podem ser usadas como referência para medir o nível de interesse de um padrão baseado em características inesperadas; outros exemplos são o uso de restrições ou limiares de interesse adicionais e metadados que descrevam dados originados de múltiplas fontes heterogêneas.

- *Ferramenta de mineração de dados*: essencial ao sistema de MD. Consiste idealmente de um conjunto de módulos funcionais para tarefas como caracterização, associação, classificação, análise de agrupamentos e análise de evolução e desvio.

- *Módulo de avaliação de padrões*: normalmente emprega medidas de interesse e interage com os módulos de MD de forma a focar a pesquisa em direção a padrões interessantes. Alternativamente, pode estar integrado ao módulo de mineração, dependendo da implementação do método de MD usado.

- *Interface gráfica para o usuário*: módulo responsável pela comunicação entre o usuário e o sistema de MD, pela interação entre esses, através da

especificação de uma tarefa ou consulta de MD. Permite também ao usuário explorar esquemas de banco de dados e *data warehouses* ou estruturas de dados, avaliar os padrões minerados e visualizar os mesmos de diferentes maneiras.

Segundo Han [HAN 2001], nem todos os "sistemas de *data mining*" existentes no mercado se aplicam à mineração de dados verdadeira. Sistemas que não suportam grandes quantidades de dados devem ser categorizados como sistemas de aprendizado de máquina, ferramentas de análise de dados estatística, ou ainda como um protótipo de um sistema experimental. Sistemas que apenas preparam dados ou recuperam informações devem ser categorizados como sistemas de banco de dados, sistemas de recuperação de informações ou sistemas de banco de dados dedutivos.

As técnicas de MD devem ser vistas sob o foco da escalabilidade[6], de forma a satisfazer os requisitos para o relacionamento com grandes conjuntos de dados. Para essa finalidade, novos métodos foram desenvolvidos e métodos estatísticos e matemáticos antigos e bem conhecidos, bem como métodos de redes neurais foram atualizados e retrabalhados para a aplicação em MD [GRA 98].

A mineração de dados trabalha com informação histórica (experiência) para aprender. Um exemplo conhecido é a elaboração de propostas de *marketing* direto para atender parcelas de consumidores com perfis de consumo diferentes. A oferta adequada reduziria o custo e aumentaria as chances de lucro. Para atingir esse objetivo, deve ser construído um modelo preditivo, de forma a ter dados no banco de dados que descrevam o que aconteceu no passado.

A utilização de MD permite a aquisição de conhecimento interessante ou informação de alto nível, que pode ser explorada de ângulos diferentes, e aplicada em tomadas de decisões, controle de processos, gerenciamento de informação e processamento de consultas. Assim, é considerada como uma das fronteiras mais importantes em sistemas de bancos de dados e um dos mais promissores desenvolvimentos interdisciplinares na indústria da informação [HAN 2001].

2.2.1 Tipos de padrões que podem ser minerados

A MD realiza tarefas básicas, comumente, classificadas em duas categorias: *descritivas* e *preditivas*. As descritivas se concentram em encontrar padrões que descrevam os dados, caracterizando as propriedades gerais desses dados em um BD, de forma interpretável pelos seres humanos. As preditivas realizam inferência nos dados correntes para construir modelos, que serão utilizados para predições do comportamento de novos dados [FAY 96, HAN 2001].

Diversas funcionalidades são usadas para especificar os tipos de padrões que podem ser encontrados nas tarefas de MD. Essas funcionalidades se referem a técnicas que podem ser usadas individualmente ou em conjunto para a descoberta de padrões.

---

[6] Um algoritmo escalável é aquele cujo tempo de execução deve crescer linearmente em proporção ao tamanho do banco de dados, considerando-se os recursos do sistema disponíveis tais como memória principal e espaço em disco [HAN 2001].

A seguir, são descritas as funcionalidades de MD e os tipos de padrões que elas podem descobrir, segundo Han [HAN 2001]:

a) *Descrição de Classe/Conceito: Caracterização e Discriminação*: em um BD, os dados podem estar associados a classes ou conceitos. Por exemplo, em uma loja de eletrônicos, *computadores* e *impressoras* são classes de itens para venda, e *grandes* e *pequenos consumidores* são conceitos de clientes. Tais descrições de classes/conceitos são úteis para a sumarização, concisão e precisão de termos e podem ser obtidas por intermédio de:

   (1) *Caracterização de dados*, que é dada pela sumarização das características gerais ou atributos de uma classe alvo de dados, por exemplo, o estudo das características de produtos de *software* cujas vendas cresceram em 10% no ano passado.

   A sumarização cria descrições compactas das características de um subconjunto de dados, permitindo a visualização de sua estrutura de dados. Alguns métodos envolvem a derivação de regras gerais, técnicas de visualização para variáveis múltiplas e a descoberta de relações funcionais entre variáveis [FAY 96].

   (2) *Discriminação de dados*, pela comparação dos atributos gerais dos objetos da classe alvo com os atributos gerais de objetos de uma ou de um conjunto de classes comparativas (classes contrastantes), por exemplo, a comparação de atributos gerais de produtos de *software* cujas vendas cresceram em 10% no ano passado com aqueles cujas vendas decresceram no mínimo 30% durante esse ano.

   As descrições discriminativas utilizam medidas comparativas para distinguir as classes alvo e as classes contrastantes, tais como métodos de análise de relevância dimensional e generalização síncrona para a construção de classes no mesmo nível conceitual incluindo apenas as dimensões mais relevantes. Podem ser expressas em forma de regras discriminativas, como por exemplo, 60% dos produtos de *software* mais vendidos no ano passado custavam menos de R$500,00, enquanto que 80% dos produtos menos vendidos nesse ano custavam mais de R$1.000,00.

   (3) ou ambas as opções.

b) *Análise Associativa*: é a descoberta de *regras associativas* que mostram condições de atributo-valor que ocorrem freqüentemente juntas em um determinado conjunto de dados. São muito utilizadas em cestas de compras ou análise de transações de dados. Por exemplo, no banco de dados relacional da loja de eletrônicos, um sistema de MD encontra regras do tipo:

$$idade(X, \text{“20 ... 29”}) \wedge renda(X, \text{“R\$1000 ... R\$2900”}) \Rightarrow compra(X, \text{“}CD\ player\text{”})$$
$$[suporte = 2\%, confiança = 60\%]$$

em que X é uma variável que representa o cliente. A regra indica que, dos clientes estudados, 2% (suporte) têm de 20 a 29 anos de idade e renda de 1000 a 2900 reais e compraram *CD player*. Há 60% de probabilidade (confiança) de um cliente desse grupo de idade e renda vir a comprar um *CD player*. Este exemplo se refere a uma *regra associativa multidimensional*, em que ocorre a associação de mais de um atributo ou predicado, e cada atributo representa uma dimensão, segundo a terminologia usada em bancos de dados multidimensionais. Um outro exemplo é a determinação de quais itens são comprados juntos freqüentemente na mesma transação. Então, uma regra gerada pode ser:

$$contém(T, \text{"}computador\text{"}) \Rightarrow contém(T, \text{"}software\text{"})$$
$$[suporte = 1\%, confiança = 50\%]$$

em que se uma transação T contém *computador*, há 50% de chance de conter também *software* e 1% de todas as transações contêm ambos. Aqui, um único atributo ou predicado se repete (*contém*) e esta regra é dita *regra associativa unidimensional*. Pode ser escrita na forma "*computador* $\Rightarrow$ *software* [1%, 50%]".

c) *Classificação e Predição*: a *classificação* é o processo de encontrar um conjunto de modelos que descrevem e distinguem classes de dados ou conceitos. Esses modelos são usados para predição de objetos cujas classes são desconhecidas, baseada na análise de um conjunto de dados de treinamento (objetos cujas classes são conhecidas). O modelo gerado pode ser representado sob a forma de regras de classificação (se-então), árvores de decisão, fórmulas matemáticas ou redes neurais. Em determinadas aplicações, é desejável a predição de alguns valores de dados ausentes ou indisponíveis, ao contrário de rótulos de classes. Isto ocorre, normalmente, quando os valores são dados numéricos e é chamado de *predição*. Apesar da predição se referir tanto a valor de dado quanto ao rótulo da classe, ela é normalmente confinada para valor de dado, sendo assim diferente da classificação.

   As tarefas de classificação e predição podem requerer *análise de relevância* para identificar atributos que não contribuem para esses processos, os quais poderão, portanto, ser excluídos.

d) *Análise de Agrupamento*: ao contrário da classificação e predição, que analisam objetos com classes rotuladas, agrupamento ou *clustering* analisa objetos cujos rótulos de classe são desconhecidos. Em dados de treinamento em que os rótulos de classe não estão presentes, esta técnica pode ser utilizada para gerar esses rótulos. Os objetos são agrupados sob o princípio da maximização da similaridade intraclasse e minimização da similaridade interclasse. Significa que os agrupamentos de objetos são formados de maneira que objetos dentro de um agrupamento possuem alta similaridade entre si, mas os objetos de agrupamentos diferentes apresentam alta dissimilaridade. Cada agrupamento formado pode ser visto como uma classe de objetos da qual regras podem ser extraídas. Também é usado para facilitar

a formação de taxonomia, que diz respeito à organização de observações dentro de uma hierarquia de classes que agrupam eventos similares.

e) *Análise de Outlier*: *outliers* são objetos de um banco de dados que não acompanham o comportamento ou modelo dos dados. Muitos métodos de MD descartam os *outliers* como ruído ou exceções. Porém, em aplicações, como por exemplo, detecção de fraudes, os eventos raros podem ser bastante interessantes. Podem ser detectados com testes estatísticos que assumem um modelo de distribuição ou probabilístico para os dados, ou utilizam medidas de distância, em que aqueles objetos mais distantes de qualquer um dos agrupamentos são considerados *outliers*. Também existem métodos baseados em desvios, que os identificam examinando as diferenças das principais características de objetos em um grupo.

f) *Análise de Evolução de Dados*: descreve e modela regularidades ou tendências para objetos cujo comportamento se modifica com o passar do tempo. Apesar de incluir caracterização, discriminação, associação, classificação ou agrupamento de dados relacionados com o tempo, esta análise se caracteriza por incluir a análise de dados de séries temporais, casamento de padrões de seqüência ou periodicidade e análise baseada em similaridade.

2.2.2 Padrões interessantes

Os sistemas de MD podem gerar milhares e até milhões de padrões ou regras. No entanto, apenas uma pequena parte dos padrões gerados será de interesse para os usuários. Para avaliar se os padrões encontrados são realmente interessantes, foram estabelecidos critérios que devem ser considerados.

Esses critérios são [FAY 96, HAN 2001]:

a) *Interpretabilidade*: ou seja, os padrões devem ser facilmente entendidos pelos humanos;

b) *Validade*: devem ser válidos em dados novos ou de teste, com determinado grau de certeza;

c) *Utilidade*: devem ser potencialmente úteis;

d) *Originalidade*: ou seja, se representam novidade.

e) Serão também considerados interessantes *se validarem uma hipótese* que o usuário deseja confirmar.

Com base nessas características, diz-se que um "*padrão interessante representa conhecimento*". Para ajudar na identificação de padrões de interesse, existem medidas de interesse *objetivas* e *subjetivas* [HAN 2001]:

- As *objetivas* são baseadas na estrutura dos padrões descobertos e na sua estatística. Por exemplo, em regras associativas, medidas objetivas são o *suporte* de uma regra, que representa o percentual de transações de um banco de dados transacional que é satisfeito por uma determinada regra, e a *confiança*, que avalia o grau de certeza da associação detectada. Em geral, cada medida de interesse está associada a um limiar estabelecido pelo usuário. Regras abaixo desse limiar podem refletir ruído, exceções, ou casos minoritários e são provavelmente de pouco valor.

- As medidas objetivas precisam ser combinadas com medidas *subjetivas*, que refletem as necessidades e interesses de um usuário em particular. Padrões interessantes sob o foco da objetividade, podem representar conhecimento comum, considerado desinteressante. Assim, as medidas subjetivas de interesse se baseiam em suposições do usuário sobre os dados e encontrarão padrões interessantes se estes forem inesperados para o usuário, ou se oferecerem informações estratégicas que o levem a agir. Quanto a padrões esperados, serão interessantes se confirmarem uma hipótese que se deseja validar ou que se pareça com um palpite do usuário sobre os mesmos.

A descoberta dos padrões interessantes, normalmente, se baseia na utilização de restrições estabelecidas pelo usuário e medidas de interesse que irão focar a pesquisa.

Seria altamente desejável que os sistemas de MD gerassem apenas os padrões interessantes. Porém, isto é considerado um problema de otimização na área, que consiste em ter que pesquisar nos padrões gerados, aqueles verdadeiramente interessantes, fato que ainda constitui um desafio na pesquisa de MD.

2.2.3 Requisitos importantes e desafios em MD

Em mineração de dados, diversos são os requisitos considerados importantes para a obtenção de sucesso em sua aplicação, como os descritos abaixo. Alguns destes itens ainda se encontram em estágio de pesquisa, portanto, representam desafios para sistemas de MD [HAN 2001]:

a) *Metodologia de mineração e questões de interação do usuário*:

- *Mineração de diferentes tipos de conhecimento em banco de dados*: deve ser coberto um largo espectro de tarefas de descoberta de conhecimento e análise de dados, incluindo caracterização, discriminação, associação, classificação, agrupamento, análise de tendências e desvios, e análise de similaridade de dados.

- *Mineração interativa de conhecimento a níveis múltiplos de abstração e incorporação de conhecimento de fundo*: permite ao usuário focar a pesquisa por padrões, provendo e refinando as requisições de MD baseadas nos resultados retornados. A informação referente ao domínio estudado deve ser usada para guiar o processo de descoberta e permitir que os padrões descobertos sejam expressos em termos concisos e em níveis diferentes de abstração.

- *Linguagens de consulta de mineração de dados e mineração de dados ad hoc*: as linguagens de consulta relacionais, como SQL, permitem a construção de consultas específicas para um determinado fim (*ad hoc*). De forma similar, devem ser desenvolvidas linguagens de consultas para MD e o ideal é que venham a ser integradas em linguagens de consulta de banco de dados ou de *data warehouse* e otimizadas para a mineração de dados eficiente e flexível.

- *Apresentação e visualização dos resultados da MD*: o sistema deve apresentar requisitos de expressão do conhecimento descoberto em linguagens de alto nível, representações visuais ou outras formas expressivas que tornem esse conhecimento de fácil entendimento e usável pelos humanos, como por exemplo, árvores, tabelas, regras, gráficos, etc.

- *Suporte a ruídos ou a dados incompletos*: o sistema deve prover métodos de limpeza e análise para o tratamento de ruídos e dados incompletos, os quais comprometem a acurácia dos padrões descobertos. Devem também prover métodos de mineração de *outliers*, para a descoberta e análise de exceções.

- *Avaliação de padrões*: diz respeito ao uso de medidas de interesse para guiar o processo de descoberta e reduzir o espaço de busca.

b) *Itens relacionados ao desempenho*:

- *Eficiência e escalabilidade dos algoritmos de MD*: para a extração de informação de bancos de dados volumosos, os algoritmos de MD devem ser eficientes e escaláveis, isto é, o tempo de execução deve ser predizível e aceitável em grandes bancos de dados. Sob a perspectiva de banco de dados, estes são considerados itens chaves em implementações de sistemas de MD.

- *Algoritmos de mineração paralelos, distribuídos e incrementais*: referentes à divisão de dados em partições, as quais são processadas em paralelo e depois mescladas. Além disso, alguns processos de MD de alto custo geram a necessidade de algoritmos incrementais que incorporem atualizações no banco de dados sem que seja necessária a mineração de todo o banco novamente.

c) *Itens relacionados à diversidade de tipos de banco de dados*:

- *Suporte de tipos de dados relacionais e complexos*: os sistemas de MD devem suportar não apenas os tipos de dados presentes em bancos de dados relacionais e *data warehouses*, mas também dados complexos, tais como hipertexto, multimídia, dados espaciais, temporais ou transacionais.

- *Mineração de informação de banco de dados heterogêneos e sistemas de informações globais*: referente à conexão de múltiplas fontes de dados por meio de redes de computadores, tal como a Internet. A descoberta de conhecimento em fontes estruturadas, semi-estruturadas e não-

estruturadas representa grande desafio para MD. Um exemplo disto é o processo de *Web mining*, em que um simples sistema de consulta de dados é inviável para a extração de informação importante sobre as ações que ocorrem na *Web*.

d) *Proteção da privacidade e segurança da informação em MD*: itens importantes devido ao uso popular crescente de ferramentas de MD e redes de telecomunicações e de computadores, gerando a necessidade de métodos que assegurem a proteção da privacidade e segurança da informação enquanto facilitador do acesso e mineração da informação.

2.3 Mineração de Dados e Estatística

Segundo Cabena [CAB 97], dentre as muitas técnicas usadas na análise tradicional de dados, a estatística é a que mais se aproxima de mineração de dados. Assim, é comum ocorrer o questionamento sobre a diferença entre as duas.

A análise estatística é orientada para verificar e validar hipóteses. A maioria das técnicas estatísticas populares requer o desenvolvimento de uma hipótese prévia. Os estatísticos, tipicamente, têm que desenvolver manualmente equações que casam com as hipóteses [CAB 97, PYL 99].

Em contraste, os algoritmos de MD podem desenvolver essas equações automaticamente. Estes algoritmos são aplicados sobre um conjunto de dados e buscam todas as hipóteses que os dados suportam. Muitas hipóteses são produzidas e a maioria, ou não tem muito significado, ou parece desconectada de algum uso ou valor. Mas outras, se revelam padrões interessantes [CAB 97, PYL 99].

Existe, atualmente, uma atratividade para MD, impulsionada por resultados positivos de pesquisas recentes em aprendizado de máquina, que deram início ao surgimento de novos algoritmos voltados para grandes volumes de dados e para diferentes tipos de dados de entrada aceitáveis em MD.

Contudo, a estatística tem uma função importante e integral na maioria dos ambientes de MD. Apesar dos fatores de distinção mencionados aparentarem favorecer MD mais do que a estatística tradicional, a melhor estratégia é sempre usar estatística e MD como abordagens complementares [CAB 97].

2.4 Considerações

Neste capítulo, a mineração de dados foi apresentada como uma abordagem diferente, que oferece grande acessibilidade para resolver problemas pela análise de dados com um aumento significativo da velocidade e qualidade de tomada de decisões estratégicas.

Porém, existem alguns pontos que devem ser considerados. Um deles é a dependência crítica de dados limpos e bem documentados. Outro, a de que problemas

diferentes exigem técnicas diferentes. Alguns problemas podem ser resolvidos automaticamente, outros, requerem a decisão do minerador e experiência no domínio estudado. Os resultados, apresentados de forma mais acessível, nem sempre são mais fáceis de serem interpretados por um usuário que não possua conhecimento de fundo do problema. Os sistemas de MD ainda são muito dependentes das pessoas envolvidas no processo.

Para os usuários de MD, existe uma expectativa de que os métodos se tornem sistemas verdadeiramente autônomos. Mas, alguns autores, como por exemplo, Pyle [PYL 99], comenta que é um erro pensar que é possível produzir modelos efetivos sem ter conhecimento dos dados e do problema, acreditando que as ferramentas farão todo o trabalho.

O usuário deve conhecer um pouco das inúmeras soluções possíveis e utilizar um ambiente de MD no qual seja capaz de interagir com diferentes tipos de abordagens para obter visão sobre os dados, a fim de que possa obter padrões interessantes.

Estas observações indicam que o uso adequado de técnicas e de metodologias de mineração de dados constitui grande fonte de pesquisa, motivando a busca por novas abordagens de extração de conhecimento em grandes bases de dados, que possam ser empregadas de formas cada vez mais eficientes por aplicações diversas.

No próximo capítulo, será apresentada a tarefa de agrupamento ou *clustering* em mineração de dados, a qual realiza a descrição de dados com a utilização de técnicas de aprendizado não-supervisionado e possui grande potencial para aplicações de MD.

3 Agrupamento ou *Clustering* em Mineração de Dados

Grande parte das teorias, métodos e algoritmos utilizados no processo de MD para a extração de padrões interessantes em grandes bases de dados, são provenientes das áreas de aprendizado de máquina e reconhecimento de padrões.

Do reconhecimento de padrões, por exemplo, provêm duas abordagens principais denominadas *aprendizado supervisionado* e *aprendizado não-supervisionado* [COS 2001]:

- *aprendizado supervisionado* é aquele que requer algum tipo de supervisão do processo que informe sobre exemplos típicos de cada classe, os quais, dentro de um processo de treinamento, permitem identificar a classe de novos objetos. Um exemplo é ensinar alguém que nunca viu maçãs ou laranjas a reconhecê-las, mostrando-lhe várias frutas e identificando cada uma; depois disso, esse alguém provavelmente será capaz de classificar novas amostras desses frutos.

- *aprendizado não-supervisionado* não possui exemplos e o número típico de classes é desconhecido. Nesse caso, é preciso identificar como os objetos podem ser agrupados em classes, sempre com base em atributos dos mesmos. O sistema de reconhecimento deverá deduzir o número de classes e quais objetos pertencem a quais classes. No exemplo acima, as frutas seriam mostradas sem identificação, devendo a pessoa descobrir os dois tipos de frutas (classes).

O processo não-supervisionado é bem mais difícil que o supervisionado, já que o sistema tem de definir sozinho o número de classes e os atributos típicos. Os métodos desse processo, também chamado de *clustering*, tendem a apresentar bons resultados quando as classes (nuvens de pontos) estão bem separadas no espaço de atributos, e apresentam dificuldades se elas estiverem sobrepostas (quando os objetos são semelhantes). A aplicação de reconhecimento de padrões juntamente com bases de dados constitui mineração de dados [COS 2001].

Assim, em MD, a descoberta de agrupamentos é uma tarefa descritiva, que utiliza técnicas de aprendizagem não-supervisionada. A identificação das características intrínsecas dos dados permite a descrição de cada agrupamento por meio de um padrão protótipo. É um processo iterativo e interativo, em que o usuário normalmente modifica parâmetros e reapresenta os dados até encontrar uma configuração satisfatória de agrupamentos.

Esse processo, ao contrário do processo de classificação, diante da análise de um conjunto de objetos, em que o rótulo da classe de cada objeto é desconhecido, consiste em agrupar dados em classes ou grupos de objetos, os quais, quando comparados uns com os outros dentro de um mesmo grupo, possuem alta similaridade, e, quando comparados com os objetos de um outro grupo, são bastante dissimilares. As dissimilaridades são avaliadas com base nos valores dos atributos que descrevem os objetos. Para essa finalidade, normalmente são usadas medidas de distâncias.

Um bom agrupamento produz segmentos de alta qualidade, em que a similaridade intraclasse é alta e a interclasse é baixa. A qualidade desse resultado também depende da métrica de similaridade usada pelo método e de sua implementação, além de sua habilidade em descobrir algum ou todos os padrões escondidos [LOP 99].

Segundo Witten, existem formas diferentes de se expressar os resultados dos agrupamentos. Os grupos identificados podem ser: *exclusivos*, de forma que suas instâncias pertencem apenas a um grupo (Figura 3.1a); *sobrepostos*, em que uma instância pode pertencer a diversos grupos (Figura 3.1b); *probabilísticos*, caso em que uma instância pertence a cada grupo com alguma probabilidade (Figura 3.1c); e ainda, *hierárquicos*, em que é feita uma divisão a grosso modo de instâncias em grupos maiores, os quais são refinados, se decompondo em grupos menores (Figura 3.1d) [WIT 99].

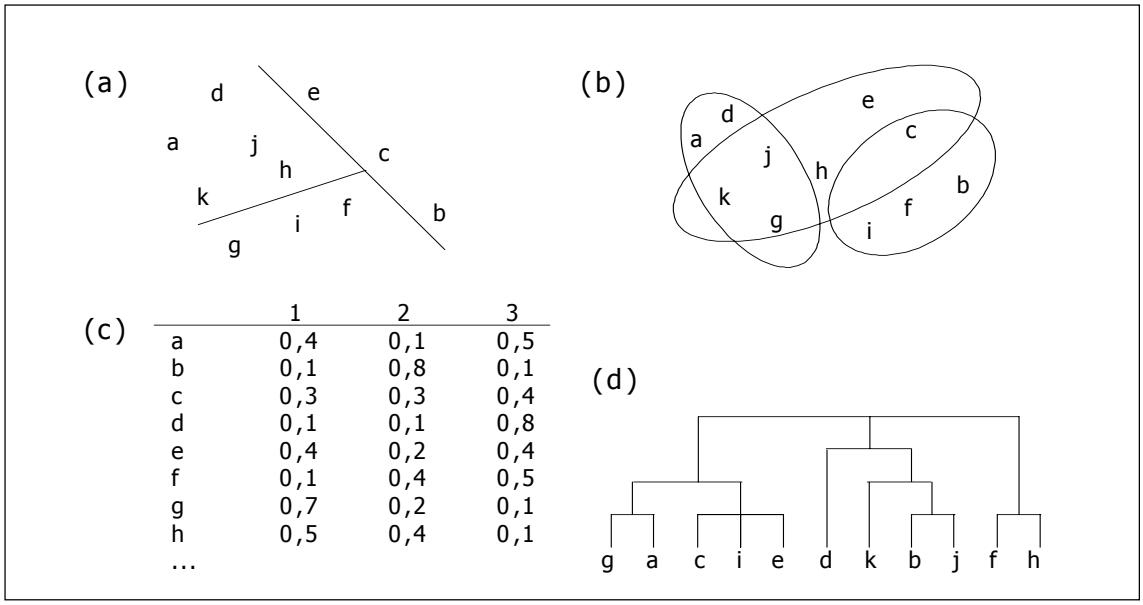


FIGURA 3.1 – Diferentes formas de representação de agrupamentos.
Fonte: WIT, 1999. p. 75.

Clustering tem sido um problema amplamente estudado em uma variedade de domínios de aplicações incluindo mineração de dados, redes neurais, inteligência artificial, análise de dados, biologia, aprendizado de máquina, reconhecimento de padrões, análise de dados, estatística, processamento de imagens, pesquisa de mercado, compressão, quantização vetorial e outras [ALS 98, BRA 98, HAN 2001].

A maioria das ferramentas de agrupamento trabalha em razão de um número pré-definido de grupos especificado por um usuário. Isso requer um conhecimento detalhado do domínio, tornando assim a tarefa de MD menos atrativa. Existem, entretanto, tecnologias mais sofisticadas que são capazes de procurar por diferentes possibilidades de quantidades de grupos e avaliar cada configuração de acordo com a sua importância [LOP 99]. Um exemplo é o algoritmo demográfico, que será visto na Seção 3.6.1, o qual determina automaticamente o número de agrupamentos a ser gerado.

3.1 Aplicações típicas de agrupamentos

As técnicas de agrupamentos são úteis em diversos tipos de aplicações, como se observa em alguns exemplos mostrados a seguir [HAN 2001]:

a) Em *negócios*, podem ajudar comerciantes a descobrir grupos distintos em suas bases de clientes e caracterizar grupos baseados em padrões de compras.

b) Em *biologia*, podem ser usadas para derivar taxonomias de plantas e animais, categorizar genes com funcionalidades similares e ter uma visão dentro de estruturas inerentes em populações.

c) Na *identificação de áreas de uso de terra* similar em um banco de dados de observação terrestre ou na identificação de grupos de casas em uma cidade de acordo com o tipo de casa, valor e localização geográfica.

d) *Classificação de documentos* na Web para a descoberta de informação.

e) *Como uma função de MD*, na qualidade de uma ferramenta independente para se ter visão de distribuição de dados, para se observar as características de cada agrupamento e para focar em um conjunto particular de grupos para análise posterior.

f) Alternativamente, a clusterização pode servir *como uma etapa de pré-processamento* para outros algoritmos, tais como caracterização e classificação, que irão então operar sobre os agrupamentos detectados.

A utilização de agrupamento em grandes BDs é um tema ativo na pesquisa de MD que objetiva métodos escaláveis, eficazes para agrupar tipos diferentes e formas de dados complexas, além de conjuntos de dados de alta-dimensionalidade [GRA 98, HAN 2001].

Agrupamento, no entanto, já vem sendo estudado por muitos anos em outras áreas. Como um ramo da estatística, por exemplo, a análise de agrupamentos tem sido estudada extensivamente, focalizando sobretudo a análise baseada em distância. Ferramentas para esse tipo de análise baseadas em métodos tais como *k-médias*, *k-medoids* e muitos outros foram também construídas em diversos *softwares* ou sistemas de análise estatística, como *S-Plus*, *SPSS* e *SAS* [HAN 2001].

Em aprendizado de máquina, os algoritmos de agrupamento realizam aprendizado não-supervisionado e utilizam o conceito de agrupamento conceitual, o qual consiste de dois componentes: (1) descobre as classes apropriadas e (2) forma descrições para cada agrupamento, como em classificação. A esse conceito se aplica a linha-guia de esforço para alta similaridade intraclasse e baixa similaridade interclasse [WIT 99, HAN 2001].

3.2 Requisitos para um bom agrupamento

Diante de uma aplicação de MD, a escolha de algoritmos de agrupamento para um determinado problema é uma tarefa árdua. Suas abordagens são provenientes de uma variedade de disciplinas e, em função disso, o uso de diferentes terminologias e definições básicas, dificultam a escolha de uma abordagem relevante para um determinado problema [FAS 99].

Para a escolha de um algoritmo, no entanto, é fundamental observar se este atende a diversos requisitos especiais próprios de *clustering*, desejáveis para a formação de bons agrupamentos, alguns dos quais serão listados abaixo, segundo Agrawal [AGR 98], Fasulo [FAS 99] e Han [HAN 2001]:

a) *Escalabilidade* e *usabilidade*: as técnicas devem apresentar rapidez e escalabilidade para agrupar dados, de forma a suportar todo o conjunto de dados. Além disso, devem ser insensíveis à ordem de entrada dos registros de dados apresentados ao sistema. Devem, ainda, ser capazes de detectar agrupamentos de formas arbitrárias, visto que a maioria desses algoritmos determina agrupamentos baseados em medidas de distância euclidianas, que tendem a encontrar agrupamentos esféricos com tamanhos e densidades similares.

b) *Habilidade de trabalhar com diferentes tipos de atributos*: os algoritmos devem ser capazes de agrupar diversos tipos de dados, tais como aqueles baseados em intervalos, numéricos, binários, categóricos, nominais, ordinais ou ainda, misturas desses tipos de dados.

c) *Requisitos mínimos do domínio do conhecimento para determinar os parâmetros de entrada*: muitos algoritmos de agrupamentos requerem dos usuários certos parâmetros de entrada na análise de agrupamentos (tais como o número de grupos desejados). É desejável que esses algoritmos ofereçam a determinação de parâmetros de forma automática, eliminando situações que afligem os usuários e tornam a qualidade de agrupamento de difícil controle.

d) *Tratamento efetivo e alta dimensionalidade*: os bancos de dados, normalmente, apresentam várias dimensões ou atributos, enquanto que muitos algoritmos de agrupamentos conseguem apenas tratar dados envolvendo duas ou três dimensões. Portanto, é um desafio agrupar objetos de dados em um espaço de alta dimensionalidade, especialmente considerando que tais dados podem estar bem esparsos e altamente distorcidos. Além disso, funções de distância aplicadas sobre dados com ruídos, dados excedentes, ausentes, desconhecidos ou errôneos, serão ineficientes, e esses algoritmos devem ser capazes de tratar esses tipos de problemas.

e) *Interpretabilidade dos resultados*: os usuários esperam resultados de agrupamento que sejam interpretáveis, compreensíveis e usáveis. Devem estar ligados a uma semântica específica para interpretações e aplicações. É importante estudar como o objetivo de uma aplicação pode influenciar a seleção de métodos de agrupamentos.

3.3 Similaridade/Dissimilaridade

A qualidade de um agrupamento é dada em função da métrica de similaridade dos objetos desse agrupamento. A definição de similaridade ou dissimilaridade entre objetos depende do tipo de dado considerado e que tipo de similaridade se está procurando e é geralmente expressa em termos de uma função de distância $d(i,j)$, cuja métrica deve satisfazer as seguintes condições [HAN 2001]:

a) $d(i,j) \geq 0$: a distância é um número não negativo.

b) $d(i,i) = 0$: a distância de um objeto para ele mesmo é 0.

c) $d(i,j) = d(j,i)$: a distância é uma função simétrica.

d) $d(i,j) \leq d(i,h) + d(h,j)$: a distância de um objeto i para um objeto j no espaço não é maior do que o caminho entre eles passando por qualquer outro objeto h (desigualdade triangular).

3.4 Estruturas de dados na análise de agrupamento

Os algoritmos de agrupamento operam tipicamente em uma das duas estruturas de dados apresentadas a seguir.

Supondo-se que um conjunto de dados a ser agrupado contenha n objetos que representam pessoas, casas, documentos, países e outros, tem-se [HAN 2001]:

- *Matriz de dados* (ou estrutura objeto-por-variável): esta representa n objetos, tais como pessoa, com p variáveis (também chamadas dimensões ou atributos), tais como idade, altura, peso, sexo, raça, etc. A estrutura está na forma de uma tabela relacional, ou matriz n-por-p (n objetos x p variáveis):

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix} \qquad 3.1$$

- *Matriz de dissimilaridade* (ou estrutura objeto-por-objeto): esta armazena uma coleção de distâncias que estão disponíveis para todos os pares de n objetos. São freqüentemente representadas por uma tabela n-por-n:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix} \qquad 3.2$$

em que $d(i,j)$ é a diferença ou dissimilaridade entre objetos i e j. Em geral, $d(i,j)$ é um número não-negativo que está próximo de 0 quando objetos i e j são altamente similares ou "próximos" uns dos outros, e se tornam maiores a medida que diferem. Desde que $d(i,j) = d(j,i)$ e $d(i,i)=0$, tem-se a matriz (3.2).

As linhas e colunas da matriz de dados representam entidades diferentes, enquanto que as linhas e colunas da matriz de dissimilaridade representam a mesma entidade.

Alguns algoritmos de agrupamento, como por exemplo o algoritmo demográfico que será apresentado mais adiante, operam sobre matrizes de dissimilaridade. Se os dados forem apresentados na forma de uma matriz de dados, podem ser primeiro transformados em uma matriz de dissimilaridade antes de se aplicar tais algoritmos de agrupamento [HAN 2001].

3.5 Tipos de dados na análise de agrupamento

Em um conjunto de dados, uma variável é um depósito que contém todas as medidas de uma característica particular de um objeto específico. Diferentes variáveis constituem diferentes tipos de dados, que requerem diferentes tipos de medidas. Na análise de agrupamentos, é preciso considerar essas diferenças de medidas, as quais têm um impacto majoritário, tanto na preparação, como na modelagem dos dados [PYL 99].

Serão relacionados a seguir, os tipos de dados mais comuns segundo Grabmeier [GRA 98], Pyle [PYL 99] e Han [HAN 2001]:

3.5.1 Variáveis binárias

Uma variável binária tem apenas dois estados: 0, que significa que a variável está ausente e 1, que significa que ela está presente.

O cálculo de dissimilaridades em dados binários é obtido com a utilização de métodos específicos. Uma forma é computar uma matriz de dissimilaridade dos dados binários apresentados. Se todas as variáveis binárias forem pensadas como tendo o mesmo peso, ter-se-á uma tabela de contingência 2-por-2 (Tabela 3.1), na qual:

a) q é o número de variáveis iguais a 1 para ambos os objetos i e j;

b) r é o número de variáveis que são iguais a 1 para o objeto i, mas são iguais a 0 para o objeto j;

c) s é o número de variáveis que são iguais a 0 para o objeto i, mas iguais a 1 para o objeto j;

d) e t é o número de variáveis que são iguais a 0 para ambos os objetos i e j.

e) O número total de variáveis é p, em que $p = q + r + s + t$.

TABELA 3.1 – Tabela de contingência para as variáveis binárias.

		Objeto *j*		
		1	0	*Soma*
Objeto *i*	1	*Q*	*r*	*q* + *r*
	0	*S*	*t*	*s* + *t*
	Soma	*q* + *s*	*r* + *t*	*P*

Fonte: Han 2001. p. 341.

3.5.2 Variáveis escaladas

- **Medidas de escala nominais**

São dadas por valores nominais ou rótulos de identificação. Apenas nomeiam objetos, independentemente da ordem ou valores de medidas, como, por exemplo, nome de pessoas.

- **Medidas de escala categóricas**

Nomeiam grupos de objetos, não entidades individuais. São rótulos arbitrários para diferentes grupos, que não informam sobre o tamanho ou tipo de diferença. Denotam que existe uma diferença de espécie ou tipo, mas não são capazes de quantificar essa diferença. A escala usada lista todas as categorias dentro das quais o valor recai. Um exemplo é o Código de Endereço Postal (CEP) de localidades.

Também chamadas q*ualitativas* ou *modais*, são consideradas uma generalização das variáveis binárias, em que podem ter mais do que dois estados. Assim, o número de estados de uma variável categórica é *M*, que pode ser denotado por letras, símbolos ou um conjunto de inteiros tais como 1, 2, ..., *M*. Esses inteiros não representam ordenação, apenas diferenciam as categorias desses valores. A dissimilaridade entre dois objetos *i* e *j* pode ser computada com a *abordagem de casamento simples*:

$$d(i,j) = \frac{p - m}{p}, \qquad 3.3$$

em que *m* é o número de casamentos (isto é, o número de variáveis para as quais *i* e *j* estão no mesmo estado), e *p* é o número total de variáveis. Podem ser designados pesos para aumentar o efeito de *m* ou para designar peso maior aos casamentos em que as variáveis têm um grande número de estados.

Às vezes, é conveniente ou necessário discretizar uma variável quantitativa para receber uma variável qualitativa. Existem várias possibilidades de se fazer isto. Uma delas consiste em aproximar a densidade de uma variável quantitativa por uma função degrau *h*, chamada *histograma*. Duas possibilidades de discretização serão mostradas a seguir:

(1) escolher intervalos de larguras iguais para o intervalo r [$r_{\min}$, $r_{\max}$] de uma variável *f* que descreve propriedades de um objeto i_f – a qual é restrita a um

subconjunto C $\subset O$, em que O é o conjunto de dados – com n objetos, cardinalidade # e q seções de igual largura $l = \frac{r_{\max} - r_{\min}}{q}$ – chamadas de *buckets* ou intervalos – em que h é constante e definida como

$$h_k = \frac{\#\{i \in C \mid r_{\min} + (k-1)l \leq i_f < r_{\min} + kl\}}{l \# C} \qquad 3.4$$

para k intervalos.

(2) usar intervalos contendo aproximadamente o mesmo número de objetos.

- **Medidas de escala ordinais**

Nesse caso, existe uma ordem significativa na listagem dos rótulos, mas não existe uma medida do quão diferente uma categoria é da outra. Se A, B e C são valores de uma variável de escala ordinal, em que A>B e B>C, então A>C.

As variáveis ordinais podem ser *discretas* ou *contínuas*. As discretas são semelhantes às categóricas, exceto que os M estados do valor ordinal são ordenados em uma seqüência significativa, que registram avaliações subjetivas de qualidades que não podem ser medidas objetivamente. Por exemplo, escalas profissionais enumeradas em uma ordem seqüencial de assistente, associado e pleno. As contínuas correspondem a um conjunto de dados contínuos em uma escala desconhecida, em que a ordenação relativa dos valores é essencial, mas sua magnitude atual não é. Por exemplo, uma escala relativa em um determinado esporte (ouro, prata, bronze).

As variáveis ordinais podem ser obtidas, também, da discretização de quantidades escaladas em intervalos, pela divisão em valores de intervalos em um número finito de classes. Então, supondo uma variável ordinal f com M_f estados, estas podem ser mapeadas para faixas de intervalos. Os estados ordenados definem as faixas 1, ..., M_f.

Se f é uma variável de um conjunto de variáveis ordinais descrevendo n objetos, a computação da dissimilaridade com respeito a f envolve as seguintes etapas:

1) O valor de f para o i-gésimo objeto é x_{if}, e f tem M_f estados ordenados, representando o intervalo 1, ..., M_f. Substituir cada x_{if} por seu intervalo correspondente $r_{if} \in \{1, ..., M_f\}$.

2) Uma vez que cada variável ordinal pode ter vários estados diferentes, costuma ser necessário fazer o mapeamento do intervalo de cada variável para [0,0, 1,0] de forma que cada variável tenha peso igual. Isto pode ser obtido pela substituição do intervalo r_{if} do i-gésimo objeto na f-gésima variável por

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}. \qquad 3.5$$

3) A dissimilaridade pode ser computada usando qualquer uma das medidas de distância citadas no item descrito a seguir, para variáveis escaladas em intervalos, usando z_{if} para representar o valor f para o i-gésimo objeto.

- **Medidas de escala em intervalos**

Apresentam informações sobre a ordenação dos valores e também sobre as diferenças em tamanho entre os valores. Trazem consigo um meio de indicar a distância que separa o valor medido. São também denominadas *quantitativas, reais* ou *contínuo numéricas*. Alguns exemplos típicos são peso e altura, latitude e longitude, temperaturas em uma escala real, renda, idade, etc.

A unidade de medida utilizada pode afetar a análise do agrupamento. Por exemplo, mudar metros para polegadas, para alturas, ou quilogramas para libras, para pesos, pode levar a uma estrutura de agrupamentos bem diferente.

Para evitar a dependência na escolha de unidades de medidas, os dados devem ser padronizados com a utilização de medidas de padronização que tentam dar a todas as variáveis um peso igual. Isto é útil em casos em que não se tem nenhum conhecimento prévio sobre os dados ou quando se deseja atribuir maior peso a um certo conjunto de variáveis do que a outros, como por exemplo, ao se agrupar candidatos a jogador de basquete, em que se deseje dar maior peso à variável altura.

Uma opção para a padronização de medidas é converter as medidas originais em variáveis sem unidades. Dadas medidas para uma variável f, isto pode ser feito da seguinte forma:

1. Calcular o desvio absoluto médio s_f:

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|), \qquad 3.6$$

em que x_{1f}, ..., x_{nf} são n medidas de f e m_f é o valor médio de f, isto é: $m_f = \frac{1}{n}(x_{1f} + x_{2f} + \ldots + x_{nf})$.

2. Calcular a medida padronizada, ou *z-score*:

$$z_{if} = \frac{x_{if} - m_f}{s_f} . \qquad 3.7$$

A vantagem do uso do desvio absoluto médio é que os *z-scores* de *outliers* não se tornam tão pequenos, permitindo a detecção de *outliers*.

Após a padronização, ou sem a padronização, em certas aplicações, a dissimilaridade (ou similaridade) entre os objetos descritos por variáveis escaladas em intervalos é tipicamente computada com base na distância entre cada par de objetos.

Algumas medidas de distância populares são:

- Distância de Minkowski:

$$d(i,j)=\sqrt[q]{|x_{i1}-x_{j1}|^q+|x_{i2}-x_{j2}|^q+\ldots+|x_{ip}-x_{jp}|^q}, \quad 3.8$$

em que $i = (x_{i1}, x_{i2}, ..., x_{ip})$ e $j = (x_{j1}, x_{j2}, ..., x_{jp})$ são dois objetos p-dimensionais e q é um inteiro positivo.

- Distância de Manhattan, se $q = 1$:

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\ldots+|x_{ip}-x_{jp}|, \quad 3.9$$

- Distância euclidiana, se $q = 2$:

$$d(i,j)=\sqrt{|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\ldots+|x_{ip}-x_{jp}|^2}, \quad 3.10$$

- Se para cada variável é determinado um peso de acordo com a sua importância percebida, a distância euclidiana ponderada pode ser computada como (ponderação pode também ser aplicada para as distâncias de Manhattan e de Minkowski):

$$d(i,j)=\sqrt{w_1|x_{i1}-x_{j1}|^2+w_2|x_{i2}-x_{j2}|^2+\ldots+w_p|x_{ip}-x_{jp}|^2}. \quad 3.11$$

- **Medidas de escala funcionais**

Essas medidas podem ser lineares, como no caso de variáveis quantitativas de valores e tamanhos iguais. Por exemplo, a escala de uma conta bancária, que inicia em um ponto 0, indicando que não há dinheiro depositado. Com relação a essa conta, valores significativos podem ser expressos como R$10,00, indicando que este é o dobro de R$5,00, e R$100,00 é o dobro de R$50,00. Em qualquer posição na escala, para quaisquer valores, a razão é uma medida significativa de propriedades da escala.

Ou podem ser também variáveis que fazem uma medida positiva em uma escala não-linear, tal como uma escala exponencial, segundo uma fórmula aproximada

$$Ae^{Bt} \text{ ou } Ae^{-Bt}, \quad 3.12$$

em que A e B são constantes positivas. Por exemplo, o crescimento de uma população de bactérias.

3.5.3 Variáveis Mistas

Muitos bancos de dados do mundo real podem conter todos os tipos de variáveis daqueles já citados aqui. Os objetos são descritos, portanto, por uma mistura de tipos de variáveis.

Algumas abordagens que podem ser utilizadas para o cálculo da dissimilaridade de variáveis mistas são:

1) Agrupar cada tipo de variável, realizando uma análise de agrupamento separada para cada tipo. Esta opção só é viável se as análises derivarem resultados compatíveis. Contudo, em aplicações reais, é pouco provável que isso aconteça.
2) Uma abordagem preferível é processar todos os tipos de variáveis juntos, realizando uma análise de agrupamento única, com o emprego de técnicas que combinem as diferentes variáveis em uma escala comum do intervalo [0,0, 1,0].

   Supondo, então, um conjunto de dados que contém p variáveis de tipos mistos. A dissimilaridade $d(i,j)$ entre os objetos i e j é definida como:

$$d(i,j) = \frac{\Sigma_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\Sigma_{f=1}^{p} \delta_{ij}^{(f)}}, \qquad 3.13$$

   em que o indicador $\delta_{ij}^{(f)} = 0$ se tanto (1) x_{if} ou x_{jf} está ausente (isto é, não há medida da variável f para o objeto i ou objeto j), quanto (2) $x_{if} = x_{jf} = 0$ e a variável f é binária assimétrica[7]; senão, $\delta_{ij}^{(f)} = 1$. A contribuição da variável f para a dissimilaridade entre i e j, $d_{ij}^{(f)}$ é dependente computacionalmente conforme o seu tipo:

   - Se f é binária ou nominal: $d_{ij}^{(f)} = 0$ se $x_{if} = x_{jf}$, senão $d_{ij}^{(f)} = 1$.
   - Se f é de escala de intervalo: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$, em que h se estende sobre todos os objetos não-ausentes para a variável f.
   - Se f é ordinal ou de escala funcional: computar os intervalos r_{if} e $z_{if} = \frac{r_{if} - 1}{M_f - 1}$, e tratar z_{if} como de escala de intervalo.

3.6 Principais métodos de agrupamentos

Diante dos inúmeros algoritmos de agrupamentos descritos na literatura, a escolha de um determinado algoritmo depende tanto do tipo de dado que vai ser avaliado, quanto do objetivo da aplicação. Em geral, na análise de agrupamento como ferramenta descritiva ou exploratória, é recomendável tentar diversos algoritmos sobre os mesmos dados para ver o que estes irão revelar.

Segundo Han [HAN 2001], os métodos de agrupamento podem ser categorizados em: métodos de particionamento, hierárquicos, baseados em densidade, baseados em grade e baseados em modelo.

[7] Uma variável é binária assimétrica se os resultados de seus estados não são de igual importância. Por convenção, o resultado mais importante deve ser codificado como 1 e o outro como 0 [HAN 2001].

Estes métodos serão resumidos a seguir. Os métodos *k-médias*, o *algoritmo demográfico* e os *mapas auto-organizáveis*, serão apresentados mais detalhadamente. O primeiro, por ser um método bastante conhecido da área, e os dois últimos, por terem sido empregados na aplicação deste trabalho.

3.6.1 Métodos de particionamento

Um método de particionamento constrói k partições de dados em um conjunto de dados de n objetos ou tuplas de dados. Cada partição representa um agrupamento e $k \leq n$. Portanto, os dados são classificados em k grupos, os quais juntos satisfazem os seguintes requisitos: (1) cada grupo deve conter no mínimo um objeto e (2) cada objeto deve pertencer a exatamente um grupo. Observa-se que o segundo requisito pode ser relaxado em algumas técnicas de particionamento fuzzy[8].

Dado o número k de partições a serem construídas, um método de particionamento cria uma partição inicial e depois utiliza uma técnica de realocação iterativa que tenta melhorar o particionamento movendo objetos de um grupo para outro. O particionamento costuma ser considerado bom se os objetos de um mesmo grupo estão "próximos" ou relacionados uns aos outros, e os objetos de diferentes grupos estão "distantes" ou são bastante diferentes.

Para se alcançar otimização global neste tipo de agrupamento, seria necessário a enumeração exaustiva de todas as possibilidades de partições. Em vez disto, muitas aplicações adotam métodos heurísticos populares, como por exemplo o algoritmo *k*-médias, em que cada grupo é representado pelo valor médio dos objetos no grupo. Os métodos heurísticos funcionam bem para encontrar grupos de forma esférica em bancos de dados de tamanho pequeno a médio. Para encontrar grupos de forma complexa e para grandes bancos de dados, estes métodos precisam ser estendidos, dando origem assim a diversas variantes de seus algoritmos.

- **Algoritmo *k*-médias**

O método *k*-médias e suas variantes é um dos mais conhecidos métodos de particionamento. Utiliza uma técnica baseada em centróide, em que a similaridade é medida com base no valor médio dos objetos em um grupo, o qual pode ser visto como o *centro de gravidade* do grupo.

O algoritmo *k*-médias é um dos mais conhecidos dentre os algoritmos de *clustering* e sua teoria e implementação têm sido amplamente mencionadas na literatura estatística nos últimos 20 anos [PEL 99]. O *k*-médias é definido sobre dados numéricos (valores contínuos), visto que requer a habilidade de computar a média.

O método *k*-médias se processa da seguinte forma: primeiro, é feita a seleção aleatória de k dos objetos, cada um dos quais representa inicialmente a média ou centro de um agrupamento. Para cada um desses k agrupamentos, são designados os objetos

[8] Técnicas de agrupamento *fuzzy* se referem àquelas em que um objeto pode ter um grau de pertinência em cada agrupamento [DUD 97].

mais similares, baseados na distância entre eles e a média do agrupamento. Então, uma nova média é computada para cada agrupamento, e o processo se repete até a convergência de uma função critério. Um critério tipicamente usado é o critério de *erro-quadrático* definido como:

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in Ci} |p - m_i|^2, \qquad 3.14$$

em que E é a soma do erro-quadrático para todos os objetos no banco de dados, p é o ponto no espaço que representa um determinado objeto e m_i é a média do agrupamento C_i (p e m_i são ambos multidimensionais). Este critério tenta deixar os k agrupamentos resultantes o mais compactos e separados possível. Na Figura 3.2, observa-se o processo resumido, segundo Han [HAN 2001].

Algoritmo: ***k*-médias**, para particionamento baseado no valor médio dos objetos no agrupamento.

Entrada: O número de agrupamentos *k* e um banco de dados contendo *n* objetos.

Saída: Um conjunto de k agrupamentos que minimizam o critério de erro-quadrático.

Método:

(1) **escolher** arbitrariamente k objetos como os centros dos agrupamentos iniciais;

(2) **repetir**

(3) (re)designar cada objeto para o agrupamento ao qual o objeto é mais similar, baseado no valor médio dos objetos no agrupamento;

(4) atualizar as média dos agrupamentos, i.e., calcular o valor médio dos objetos para cada agrupamento;

(5) **até** que não haja mudanças;

FIGURA 3.2 – O algoritmo *k*-médias.
Fonte: HAN, 2001. p. 349.

Um exemplo deste algoritmo pode ser visualizado na Figura 3.3, na qual se deseja agrupar um conjunto de objetos em 2 grupos em função dos atributos peso e altura. A média de cada agrupamento é marcada por um " □ ".

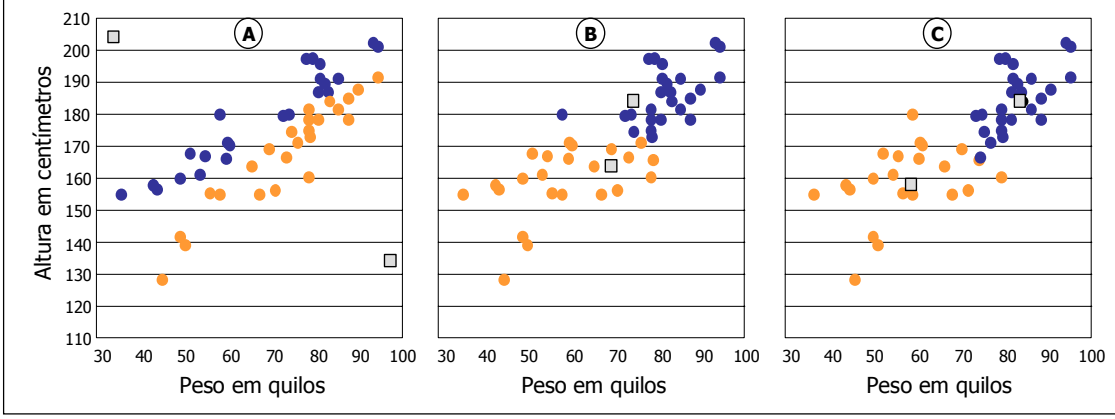


FIGURA 3.3 – Agrupamento de um conjunto de objetos baseado no método *k*-médias.
Fonte: COS, 2001. p. 29.

O método *k*-médias funciona bem quando os agrupamentos formam nuvens compactas e bem separadas umas das outras. É relativamente escalável e eficiente no processamento de grandes quantidades de dados devido à sua complexidade computacional, definida como $O(nkt)$, em que n é o número total de objetos, k, o número de agrupamentos e t, o número de iterações. Normalmente, $k << n$ e $t << n$.

Algumas dificuldades que apresenta são: (1) o fato de que este método só pode ser aplicado quando a média de um agrupamento é definida, o que não é o caso de aplicações que envolvem dados categóricos; (2) a necessidade do usuário ter que especificar o número k de agrupamentos; (3) não é adequado para a descoberta de grupos com formas não convexas ou de tamanhos bem diferentes e (4) é sensível a ruídos e *outliers*, visto que um pequeno número de pontos de dados pode afetar o valor médio.

- **Algoritmo demográfico**

O conceito fundamental do agrupamento demográfico é a construção dos agrupamentos pela comparação de cada objeto com todos os agrupamentos criados pela execução da mineração de dados. O algoritmo atribui o objeto a um agrupamento pela maximização da diferença entre os pontos a favor e contra a localização de um registro [CAB 97].

A técnica de agrupamento demográfico se baseia em um princípio de voto simples, chamado *Condorcet*, para medir a distância entre os objetos de entrada e assim designá-los para agrupamentos específicos [CAB 97, IBM 99].

Segundo Cabena [CAB 97], este processo se realiza da seguinte maneira:

- Inicialmente, os pares de objetos são comparados pelos valores de seus atributos individuais. Se um par de objetos tem o mesmo valor para o mesmo atributo, este obtém um voto positivo, senão, este obtém um voto negativo. Ao final, a pontuação total é calculada como a diferença dos pontos a favor e contra, posicionando o objeto em um determinado agrupamento.

- Um objeto será designado para um outro agrupamento se a pontuação total for maior do que as pontuações totais caso o objeto tivesse sido designado para qualquer um dos outros agrupamentos. Se as pontuações totais apresentam-se negativas, o objeto torna-se um candidato para ser posicionado em seu próprio agrupamento.

- Existe um número de passagens sobre o conjunto de dados, ao qual cada objeto deve ser revisto para uma redesignação potencial para um agrupamento diferente. Os agrupamentos e seus centros são então atualizados continuamente a cada passagem, até que ou o número máximo de passagens seja alcançado, ou o número máximo de agrupamentos seja alcançado e os centros dos agrupamentos não mudem, significativamente, ao serem medidos por uma margem determinada pelo usuário.

- Uma vez que estejam formados agrupamentos grandes o suficiente, a decisão para posicionar um próximo objeto em um agrupamento não é feita pela comparação de cada valor de atributo deste objeto com valores de atributos de todos os objetos em todos os agrupamentos, e sim, pela comparação dos valores de atributo deste objeto contra as distribuições dos valores de cada atributo dos agrupamentos que já se encontram formados.

As vantagens do agrupamento demográfico são: (1) a sua habilidade para determinar automaticamente o número de agrupamentos a ser gerado, (2) a clareza do particionamento resultante de grandes conjuntos de dados, e ainda, (3) a técnica provê ordenação rápida e natural de bancos de dados bastante volumosos.

Em contraste com o agrupamento neural (vide Seção 3.6.5), que é mais adequado para dados numéricos, o agrupamento demográfico é adequado, particularmente, para dados categóricos. Contudo, variáveis não-categóricas também podem ser tratadas, mas nestes casos, o analista deve discretizar os valores que serão usados pelo algoritmo na determinação da similaridade ou dissimilaridade de duas variáveis. Valores dentro da uma faixa de discretização registram um voto em favor da igualdade, enquanto que valores fora da faixa registram um voto contra a igualdade. A medida de similaridade é então não apenas um simples valor binário (0,1), mas varia de 0 a 1. Zero indica valores distantes, 1 indica valores idênticos, e 0,5 indica que os valores estão separados exatamente pelo valor de tolerância [CAB 97]. Na Figura 3.4, observa-se o processo resumido, com base em Grabmeier [GRA 98].

Algoritmo: **demográfico**, para particionamento baseado em um critério de agrupamento (critério *Condorcet*).

Entrada: Um conjunto de dados O,
um critério c : {agrupamento C} $\rightarrow [s_{min}, s_{max}]$.

Método:

(1) **estabelecer** $C = \varnothing$;

(2) **iteragir** sobre todos os objetos x em O :

(2.1) iteragir sobre todos os k agrupamentos já construídos C $\in$ C, colocar x em C e atualizar $c(C)$ sob esta modificação potencial;

(2.2) considerar a construção de um novo agrupamento $\{x\}$, consistindo exclusivamente de x e potencialmente colocar $\{x\}$ em C e atualizar $c(C)$;

(2.3) escolher, dentre as $t + 1$ possibilidades o agrupamento que obtiver o maior valor de $c(C)$, sendo t o número de agrupamentos.

(3) **repetir** até $n(t + 1)$ casos, sendo n o número de objetos.

Saída: Retornar ao agrupamento C.

FIGURA 3.4 – O algoritmo demográfico.
Fonte: Baseado em GRA, 1998. p. 48-49.

– O Critério *Condorcet*

O algoritmo demográfico é baseado no *New Condorcet Criterion* (NCC) de Michaud (1997) [MIC 97], o qual foi inspirado pelo trabalho de Condorcet (1743-1794)

na busca de um meio apropriado para agregar votos (posições dos candidatos) em uma eleição [SOF 2002].

O NCC mede as concordâncias intraclasses, bem como as discordâncias interclasses e as combina de tal forma que partições que têm pequenas distâncias intraclasses e grandes distâncias interclasses terão maior medida. Tal medida pode ser expressa como [MIC 97, SOF 2002]:

$$G(P) = \sum_{k=1}^{p} \sum_{i \in L_k} \left(\sum_{j \in L_k ; j \neq i} \left(m - d_{ij} \right) + \sum_{j \notin L_k} d_{ij} \right) \qquad 3.15$$

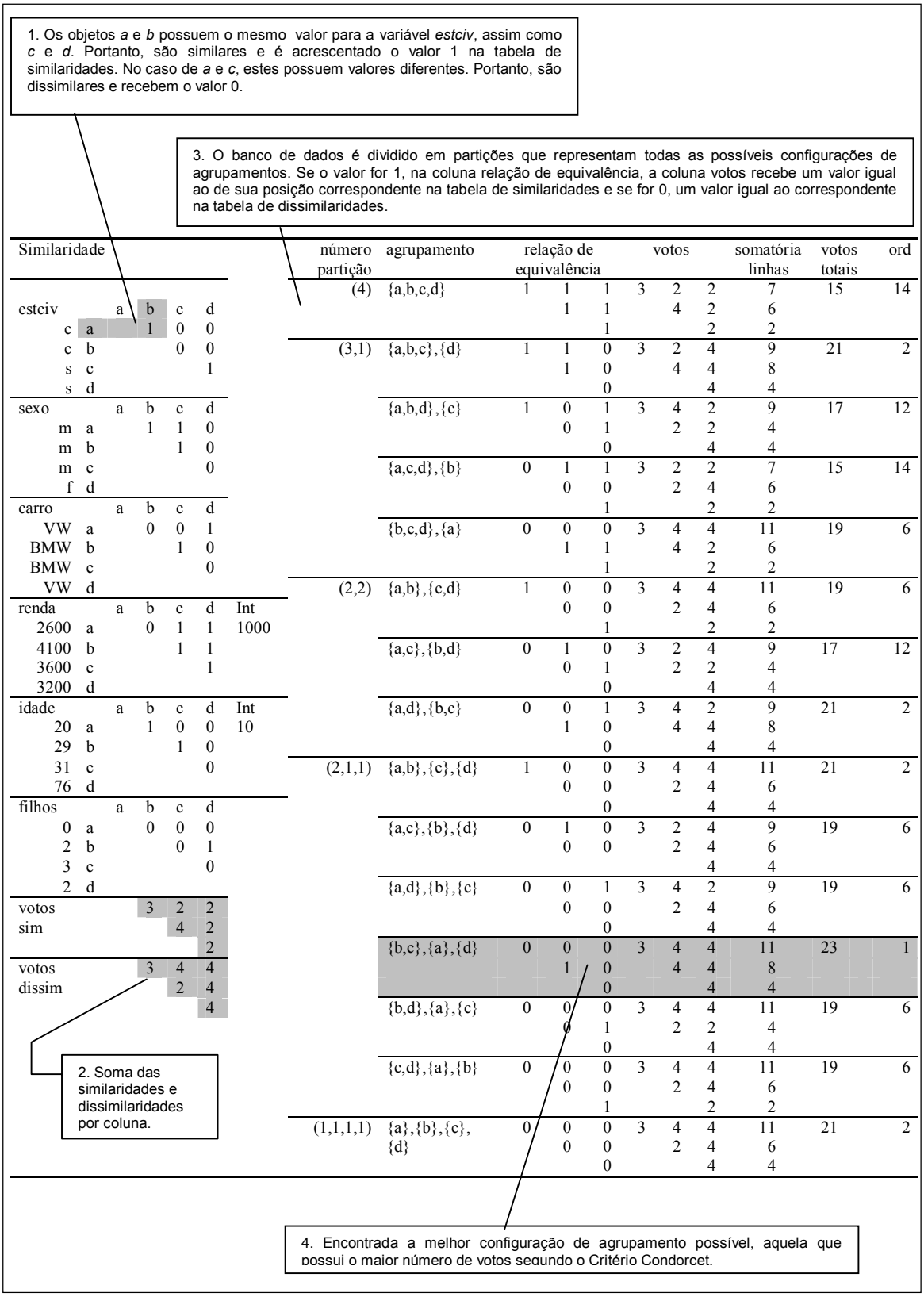
em que o índice da soma à esquerda se refere às classes p (agrupamentos), enquanto que as somas internas dizem respeito às observações que estão dentro e fora de algum agrupamento L_k. Os fatores da soma são as distâncias d_{ij}.

Para variáveis categóricas, a distância d entre duas observações i e j é o número de variáveis para as quais as duas observações empregam valores diferentes, isto é, o número de dissimilaridades entre duas observações. Se m variáveis são medidas para cada observação, então segue que $m - d_{ij}$ é justamente o oposto de d_{ij}: o número de similaridades entre observações i e j.

G (P) calcula para uma determinada partição P, a soma de todas as similaridades intraclasses e todas as dissimilaridades interclasses. Diferentes partições podem então ser ordenadas de acordo com o seu valor de G (P) – quanto mais alto o G (P), melhor a partição. Com a interpretação de concordâncias intraclasses a discordâncias interclasses como votos para uma determinada partição, a conexão para o trabalho *Condorcet* se torna aparente: o vencedor entre todos os candidatos (partições) da eleição (a análise de agrupamento) é aquele que recebe a maioria dos votos (o valor mais alto de G (P)) [SOF 2002]. Se houver variáveis numéricas, elas devem ser discretizadas de forma que o NCC possa ser aplicado.

Assim se explica a estratégia para encontar a partição ótima (isto é, aquela com o valor mais alto de G (P)). São consideradas todas as partições cujo número de agrupamentos p é menor do que o número máximo de agrupamentos especificados pelo usuário. Por exemplo, se o usuário quiser encontrar a melhor partição dentre aquelas com, no máximo, 3 agrupamentos, então o algoritmo calculará G (P) para todas as partições com um, dois ou três agrupamentos.

A Figura 3.5 mostra um exemplo de quatro objetos *a*, *b*, *c*, e *d* descritos por seis variáveis: *estciv*, *sexo*, *carro*, *renda*, *idade* e *filhos*, cujos valores são dados nas colunas à esquerda. As variáveis *renda* e *idade* são quantitativas e uma diferença limite de 1000 e 10, respectivamente, é utilizada como medida simplificada de similaridade. As outras variáveis são qualitativas [GRA 98].



Similaridade							número partição	agrupamento	relação de equivalência			votos			somatória linhas	votos totais	ord
							(4)	{a,b,c,d}	1	1	1	3	2	2	7	15	14
estciv		a	b	c	d					1	1		4	2	6		
c	a		1	0	0						1			2	2		
c	b			0	0		(3,1)	{a,b,c},{d}	1	1	0	3	2	4	9	21	2
s	c				1					1	0		4	4	8		
s	d										0			4	4		
sexo		a	b	c	d			{a,b,d},{c}	1	0	1	3	4	2	9	17	12
m	a		1	1	0					0	1		2	2	4		
m	b			1	0						0			4	4		
m	c				0			{a,c,d},{b}	0	1	1	3	2	2	7	15	14
f	d									0	0		2	4	6		
carro		a	b	c	d						1			2	2		
VW	a		0	0	1			{b,c,d},{a}	0	0	0	3	4	4	11	19	6
BMW	b			1	0					1	1		4	2	6		
BMW	c				0						1			2	2		
VW	d						(2,2)	{a,b},{c,d}	1	0	0	3	4	4	11	19	6
renda		a	b	c	d	Int				0	0		2	4	6		
2600	a		0	1	1	1000					1			2	2		
4100	b			1	1			{a,c},{b,d}	0	1	0	3	2	4	9	17	12
3600	c				1					0	1		2	2	4		
3200	d										0			4	4		
idade		a	b	c	d	Int		{a,d},{b,c}	0	0	1	3	4	2	9	21	2
20	a		1	0	0	10				1	0		4	4	8		
29	b			1	0						0			4	4		
31	c				0		(2,1,1)	{a,b},{c},{d}	1	0	0	3	4	4	11	21	2
76	d									0	0		2	4	6		
filhos		a	b	c	d						0			4	4		
0	a		0	0	0			{a,c},{b},{d}	0	1	0	3	2	4	9	19	6
2	b			0	1					0	0		2	4	6		
3	c				0									4	4		
2	d							{a,d},{b},{c}	0	0	1	3	4	2	9	19	6
votos		3	2	2						0	0		2	4	6		
sim			4	2							0			4	4		
				2				{b,c},{a},{d}	0	0	0	3	4	4	11	23	1
votos		3	4	4						1	0		4	4	8		
dissim			2	4							0			4	4		
				4				{b,d},{a},{c}	0	0	0	3	4	4	11	19	6
										0	1		2	2	4		
											0			4	4		
								{c,d},{a},{b}	0	0	0	3	4	4	11	19	6
										0	0		2	4	6		
											1			2	2		
							(1,1,1,1)	{a},{b},{c},{d}	0	0	0	3	4	4	11	21	2
										0	0		2	4	6		
											0			4	4		

FIGURA 3.5 – Exemplo do processo de votação Condorcet.
Fonte: GRA, 98. p. 64.

3.6.2 Métodos hierárquicos

Estes métodos criam uma decomposição hierárquica de um determinado conjunto de objetos de um banco de dados. Um método hierárquico pode ser classificado como *aglomerativo* ou *divisivo*, baseado em como a decomposição hierárquica é formada.

A abordagem *aglomerativa*, também chamada *bottom-up*, inicia com cada objeto formando um grupo separado. Ela mescla, sucessivamente, os objetos ou grupos próximos uns aos outros, até que todos os grupos tenham sido mesclados em um único grupo (o grupo mais alto da hierarquia), ou até que atenda a uma condição de terminação. A abordagem *divisiva*, também chamada *top-down*, inicia com todos os objetos em um mesmo grupo e a cada iteração sucessiva, um grupo é dividido em grupos menores, até que eventualmente cada objeto esteja em um único grupo ou que atenda a uma condição de terminação.

Esses métodos, apesar de simples, apresentam dificuldades relacionadas à seleção de pontos de mesclagem ou divisão. Isso é crítico, pois uma vez que um agrupamento é criado, o processo, na próxima etapa, será aplicado sobre o novo agrupamento gerado. Portanto, uma desvantagem desses métodos é que, uma vez que uma etapa seja feita (mesclagem ou divisão), ela não pode ser desfeita. Essa rigidez é útil por levar a um menor custo computacional, visto que não há a preocupação sobre um número combinatório de diferentes escolhas. Mas, resulta em um problema maior, que é a impossibilidade de corrigir decisões errôneas. Além disso, a escalabilidade fica comprometida, uma vez que a decisão de mesclar ou dividir requer o exame e avaliação de um bom número de objetos ou agrupamentos.

Existem várias abordagens citadas na literatura para melhorar a qualidade dos agrupamentos hierárquicos, dentre elas cita-se: (1) fazer uma análise cuidadosa das "ligações" de objetos a cada partição hierárquica, tais como em CURE (Guha *apud* [HAN 2001] p. 356) e Chameleon (Karypis *apud* [HAN 2001] p. 361), ou (2) integrar aglomeração hierárquica e realocação iterativa usando primeiro um algoritmo aglomerativo hierárquico e então refinando o resultado usando realocação iterativa, como em BIRCH (Zhang *apud* [HAN 2001] p. 357).

3.6.3 Métodos baseados em densidade

A maioria dos métodos de particionamento agrupa objetos baseados na distância entre eles. Tais métodos encontram apenas grupos de forma esférica e encontram dificuldade para descobrir grupos de formas arbitrárias. Para resolver esse problema, foram criados métodos de agrupamentos baseados na noção de *densidade*. Tipicamente, os agrupamentos são considerados como regiões densas de objetos no espaço de dados que estão separados por regiões de baixa densidade (representando ruídos). A idéia geral desses métodos é aumentar o grupo dado, inicialmente, à medida que a densidade (número de objetos ou pontos de dados) na vizinhança exceda algum limiar, isto é, para cada ponto em um dado agrupamento, a vizinhança em um determinado raio tem que conter ao menos um número mínimo de pontos. Tal método pode ser usado para filtrar ruídos (*outliers*) e descobrir grupos de formas arbitrárias.

Portanto, um método de agrupamento baseado em densidade é um conjunto de objetos conectados por densidade, que é máxima, no que se refere à alcançabilidade da densidade. Todo o objeto que não estiver contido em qualquer agrupamento será considerado como ruído.

Exemplos típicos desses métodos são (1) o DBSCAN (Ester *apud* [HAN 2001] p. 363.) que faz um agrupamento crescer de acordo com a densidade dos objetos da vizinhança; (2) o DENCLUE (Hinneburg e Keim *apud* [HAN 2001] p.366) que faz o crescimento baseado em alguma função de densidade e (3) o OPTICS (Ankerst *apud* [HAN 2001] p. 365), que gera uma ordenação aumentada da estrutura de agrupamento dos dados para a análise automática e interativa.

3.6.4 Métodos baseados em grade

Esses métodos quantizam o espaço de objetos em um número finito de células que formam uma grade estrutural. Todas as operações de agrupamento são realizadas na estrutura da grade (ou seja, no espaço quantizado). A maior vantagem dessa abordagem é o baixo tempo de processamento, o qual é tipicamente independente do número de objetos do banco de dados e dependente apenas do número de células em cada dimensão no espaço quantizado.

Um exemplo típico deste método é o STING (Wang *apud* [HAN 2001] p. 370), o qual explora informação estatística armazenada nas células da grade. Outros dois exemplos, ambos baseados em grade e em densidade, são CLIQUE [AGR 98], que representa uma abordagem para agrupamento em espaço de alta dimensionalidade de dados e WaveCluster (Sheikholeslami *apud* [HAN 2001] p. 372), que agrupa objetos usando o método de transformada *wavelet*.

3.6.5 Métodos baseados em modelos

Esses métodos criam um modelo hipotético para cada um dos agrupamentos e encontram o melhor ajuste dos dados para o modelo apresentado. Um algoritmo desse tipo deve alocar grupos construindo uma função de densidade que reflete a distribuição espacial dos pontos de dados. Também leva a uma maneira de determinar automaticamente o número de grupos baseado em estatísticas padrões, em que "ruídos" ou *outliers* são levados em conta, produzindo assim métodos de *clustering* robustos.

Existem duas abordagens principais para os métodos baseados em modelos, que são a *abordagem estatística* e *abordagem de redes neurais*.

- **Abordagem estatística**

Em aprendizado de máquina, existe o conceito de *agrupamento conceitual* que consiste em, dado um conjunto de objetos não rotulados, produzir um esquema de classificação sobre os objetos. Diferente do agrupamento convencional, que identifica primariamente grupos de objetos semelhantes, o agrupamento conceitual vai um nível

além, encontrando também descrições características para cada grupo, em que cada um destes representa um conceito ou classe.

Assim, o agrupamento conceitual é um processo de duas etapas: (1) é realizado o agrupamento e (2) é feita a caracterização. A qualidade do agrupamento não é apenas uma função dos objetos individuais, em vez disso, incorpora fatores tais como a generalidade e simplicidade das descrições conceituais derivadas. A maioria dos métodos de agrupamento conceitual adota uma abordagem estatística que usa medidas probabilísticas para a determinação dos conceitos ou agrupamentos.

Um exemplo do método conceitual incremental é o COBWEB proposto por Fisher (Fisher *apud* [HAN 2001] p. 377).

- **Abordagem de redes neurais**

Nos últimos anos, alguns progressos marcantes em redes neurais se relacionam a problemas de *clustering*. A simulação de uma rede de neurônios ocorre de forma semelhante à modelagem de processos no cérebro humano, em que sinais de entrada são transferidos pelos *axiônios*[9] por meio de conexões químicas chamadas *sinapses*[10], entre os neurônios. Cada neurônio envia novos sinais de saída se obtiver suficiente ativação de entrada [GRA 98]. Em uma rede neural, os nós (neurônios) são arranjados em camadas com conexões ponderadas (sinapses) entre elas. Sua estrutura pode conter várias camadas.

A utilização de redes neurais para a formação de agrupamentos representa cada agrupamento como um "*protótipo*", o qual não precisa corresponder, necessariamente, a um exemplo de dados ou objeto particular. Os novos objetos são atribuídos ao agrupamento cujo protótipo é o mais similar baseado em alguma medida de distância e seus atributos podem ser preditos com base nos atributos do protótipo daquele agrupamento.

Duas abordagens importantes desses métodos são *aprendizado competitivo* e *mapas auto-organizáveis*.

1) Aprendizado competitivo

O *aprendizado competitivo* envolve uma arquitetura hierárquica de muitas unidades ou neurônios artificiais que realizam uma competição do tipo "*o vencedor leva tudo*" para o objeto que está sendo correntemente apresentado ao sistema.

É um processo adaptativo no qual os neurônios, em uma rede neural, se tornam sensíveis gradualmente a categorias diferentes de entrada, formando conjuntos de exemplares em um domínio específico do espaço de entrada. Ocorre uma espécie de divisão de trabalho na rede neural, em que diferentes neurônios se especializam para representar diferentes tipos de entradas. A especialização é forçada pela competição

[9] O axônio é uma projeção filamentar do neurônio que transmite informação na forma de pulsos elétricos para as várias partes do sistema nervoso e do organismo [ENG 2001].

[10] Sinapses são estruturas de contato para a recepção da informação que chega por meio de um pulso nervoso do axônio [ENG 2001].

entre os neurônios: quando uma entrada x incide, o neurônio que é capaz de melhor representá-la vence a competição e tem permissão para aprendê-la cada vez melhor [KAS 97].

A Figura 3.6 mostra um exemplo simples com apenas uma camada de entrada, uma camada de saída e um neurônio de saída para cada agrupamento. É dado o número t de agrupamentos. Cada uma das m variáveis corresponde a um neurônio de entrada. A rede consiste de todas as conexões m x t representadas na figura, com $m = 4$ e $t = 3$, que mostra apenas as conexões correspondentes aos agrupamentos 1 e 3. Os neurônios de saída são modelados também por elementos $w \in \mathbb{R}^m$ respectivamente $[0,1]^m$. Nesse modelo, a extremidade da conexão de entrada de uma variável i é rotulada com um peso w_i denominado *peso sináptico*. Os objetos de entrada x pertencentes a uma população O, que serão agrupados, são considerados como sinais de entrada x_i para o neurônio de entrada i. O sinal de ativação para o neurônio de saída modelado por w é definido pelo produto escalar $x^t w = \sum_{k=1}^{n} x_k w_k$ de x e w. O neurônio de saída, que representa o agrupamento no qual será posicionado o objeto x, será aquele que obtiver o maior sinal $x^t w$ [GRA 98].

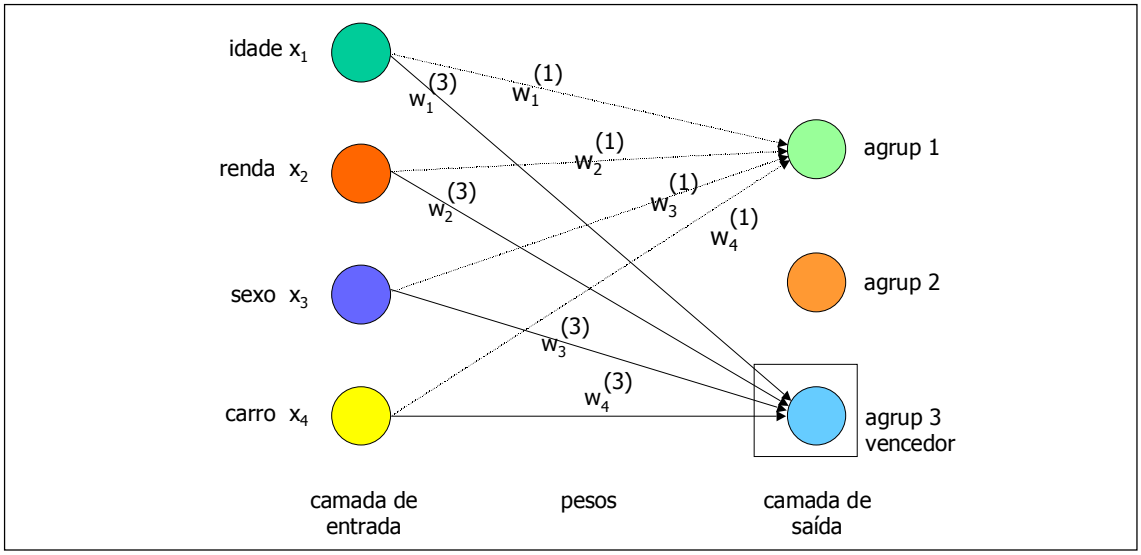


FIGURA 3.6 – Aprendizado competitivo em uma rede neural.
Fonte: GRA, 98, p. 56.

Considerando-se os pesos como a definição de um protótipo, então novos objetos são designados ao grupo do protótipo mais próximo. O número de grupos e o número de unidades por grupo são parâmetros de entrada.

Ao fim da clusterização, cada grupo pode ser pensado como um novo "atributo" que detecta alguma regularidade nos objetos. Assim, os grupos resultantes podem ser vistos como um mapeamento de atributos de baixo nível para atributos de alto nível.

2) Mapas auto-organizáveis ou *self-organizing feature maps* (SOMs)

Os mapas auto-organizáveis assumem que existe uma ordenação entre os objetos de entrada, e que as unidades (neurônios) irão eventualmente assumir esta estrutura no

espaço. A organização desses neurônios é dita como formadora de um mapa de características. Os neurônios estão localizados em uma grade discreta, que é o mapa auto-organizável. É uma generalização do aprendizado competitivo. Se o neurônio vencedor e também seus vizinhos na grade têm permissão para aprender, os neurônios vizinhos irão gradualmente se especializar para representar entradas similares, e as representações se tornarão ordenadas na grade do mapa [KAS 97].

O SOM, proposto por T. Kohonen (1981), realiza agrupamento com várias unidades competitivas para o objeto corrente. A unidade cujo vetor de peso está mais próximo ao objeto corrente se torna a unidade ativa ou vencedor. Assim, os pesos da unidade vencedora, bem como aqueles de seus vizinhos mais próximos, são ajustados de forma a se aproximarem do objeto de entrada [HAN 2001]. A Figura 3.7 mostra um exemplo desse ajuste [GRA 98].

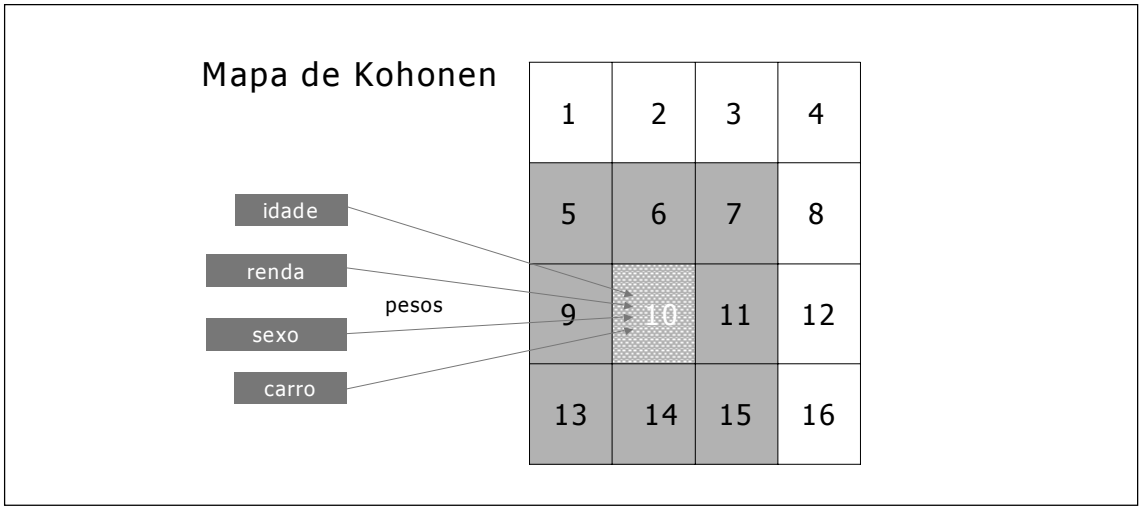


FIGURA 3.7 – Mapa auto-organizável de Kohonen. Neste exemplo, os pesos da unidade vencedora (10) são atualizados, bem como os de seus vizinhos 5, 6, 7, 11, 13, 14 e 15.
Fonte: GRA, 98, p. 57.

– Algoritmo básico

O algoritmo básico do SOM será descrito a seguir, segundo Vesanto [VES 2000]:

Um SOM é formado de M neurônios localizados em uma grade regular, geralmente unidimensional ou bidimensional. Cada neurônio i do SOM é representado por um peso n-dimesional ou vetor de referência $m_i = [m_{i1}, ..., m_{in}]$, em que n é igual à dimensão dos vetores de entrada. As grades podem apresentar dimensões mais elevadas, geralmente pouco usadas, já que sua visualização é um pouco mais problemática.

Os neurônios do mapa são conectados aos neurônios adjacentes por uma relação de vizinhança ditando a estrutura do mapa. Vizinhos imediatos, os neurônios que são adjacentes pertencem à vizinhança-1 $N_{i,1}$ do neurônio i. No caso bidimensional, os neurônios do mapa podem ser arranjados em grades retangulares ou hexagonais, com vizinhanças de diferentes tamanhos, como mostrado na Figura 3.8. O número de

neurônios determina a granularidade do mapa resultante, que afeta a acurácia e a capacidade de generalização do SOM.

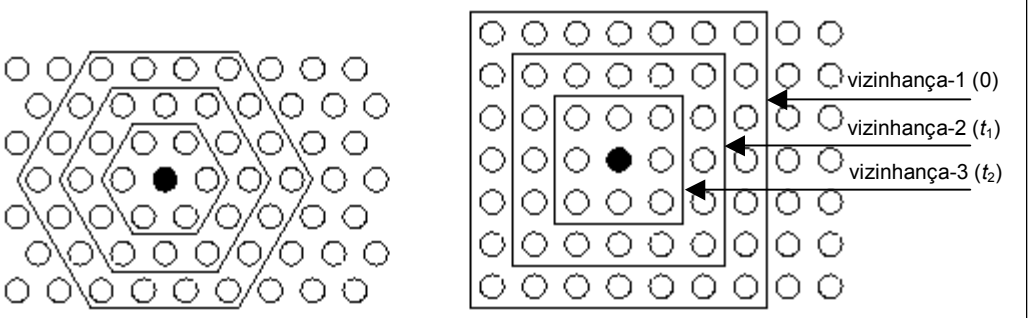


FIGURA 3.8 – Vizinhança (tamanhos 1, 2 e 3) do neurônio *i* (ponto escuro ao centro). A vizinhança inicia larga e decresce de tamanho, lentamente, ao passar do tempo, em que $0<t_1<t_2$.
Fonte: VES 97, p. 4.

Antes da fase de treinamento, os vetores de pesos recebem valores iniciais. O SOM é robusto com respeito à inicialização, mas apropriadamente executado, permite ao algoritmo convergir mais rapidamente para uma boa solução. A inicialização pode ser feita com o uso de um dos três procedimentos abaixo:

- inicialização *randômica*, na qual os vetores de pesos são inicializados com pequenos valores randômicos;

- inicialização *de um exemplo*, na qual os vetores de pesos são inicializados com amostras extraídas do conjunto de dados de entrada;

- inicialização *linear*, na qual os vetores de pesos são inicializados ao longo do subespaço linear atravessado pelos dois principais autovetores do conjunto de dados de entrada.

Em cada etapa de treinamento, um vetor exemplo x do conjunto de dados de entrada é escolhido randomicamente e uma medida de similaridade é calculada entre ele e todos os vetores de pesos do mapa. A BMU (*Best-Matching Unit* ou unidade com melhor casamento), denotada como c, é a unidade cujo vetor de peso tem a maior similaridade com o exemplo de entrada x. A similaridade normalmente é definida por meio de medida de distância, tipicamente distância euclidiana. Formalmente, a BMU é definida como o neurônio para o qual

$$|| x - m_c|| = \min_i \{|| x - m_i||\}, \qquad 3.16$$

em que || . || é a medida da distância.

Depois de achar a BMU, os vetores de pesos do SOM são atualizados. Os vetores de pesos da BMU e seus vizinhos topológicos são movidos para perto do vetor de

entrada no espaço de entrada. Este procedimento de adaptação expande a BMU e seus vizinhos topológicos em direção ao vetor exemplo, como ilustrado a Figura 3.9, em que o vetor de entrada dado à rede é marcado por um *x*.

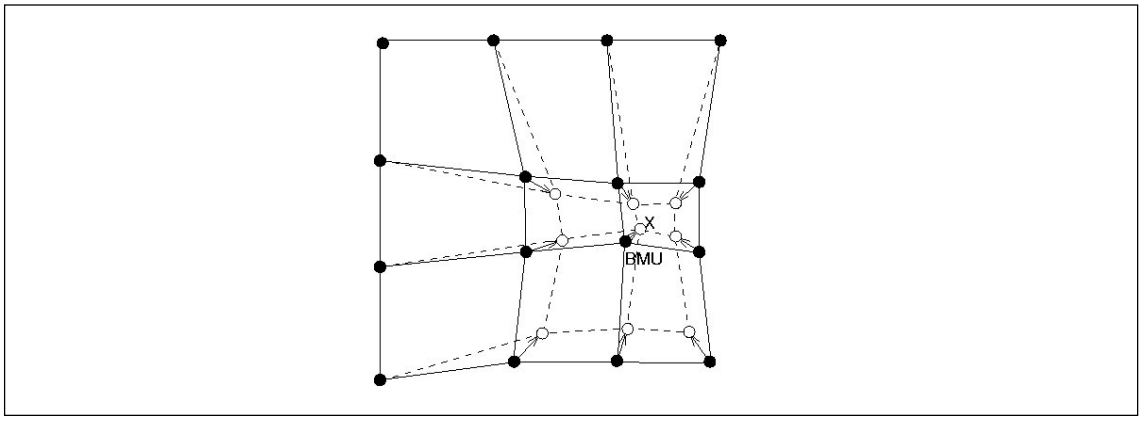


FIGURA 3.9 - Atualização da BMU e seus vizinhos em direção ao exemplo de entrada *x*. As linhas sólidas e pontilhadas correspondem à situação antes e depois da atualização, respectivamente.
Fonte: VES 97, p. 5.

A regra de atualização do SOM para o vetor de peso da unidade *i* é:

$$m_i(t+1) = m_i(t) + h_{ci}(t)\,[x(t) - m_i(t)], \qquad 3.17$$

em que *t* denota tempo. O $x(t)$ é o vetor de entrada randomicamente extraído do conjunto de dados de entrada no tempo *t* e $h_{ci}(t)$ o núcleo de vizinhança em volta da unidade vencedora *c* no tempo *t*. O núcleo de vizinhança é uma função não-crescente do tempo e da distância da unidade *i* para a unidade vencedora *c*. Ele define a região de influência que o exemplo de entrada tem no SOM. O núcleo é formado de duas partes: a função de vizinhança $h(d, t)$ e a função da taxa de aprendizado $\alpha(t)$:

$$h_{ci}(t) = h(\|r_c - r_i\|, t)\alpha(t), \qquad 3.18$$

em que r_i é a localização da unidade *i* na grade do mapa e r_c a localização da unidade vencedora.

A função de vizinhança pode ser, por exemplo, gaussiana: $exp\left(-\frac{\|r_c - r_i\|^2}{2\sigma^2(t)}\right)$, que dá resultados um pouco melhores, mas é computacionalmente mais pesada. Usualmente o raio de vizinhança é maior no início e é decrescido linearmente para 1 durante o treinamento.

A taxa de aprendizado $\alpha(t)$ é uma função de tempo decrescente. Duas formas comumente usadas são uma função linear e uma função inversamente proporcional ao tempo: $\alpha(t) = \frac{A}{t+B}$, em que *A* e *B* são constantes selecionadas de forma adequada. O uso do último tipo de função assegura que todos os exemplos de entrada têm aproximadamente igual influência no resultado do treinamento.

O treinamento costuma ser executado em duas fases. Na primeira fase, são usados um valor inicial de alfa e raio de vizinhança relativamente grandes. Na segunda

fase, o valor de alfa e o raio da vizinhança são pequenos desde o início. Este procedimento corresponde a primeiro ordenar o SOM para, aproximadamente, o mesmo espaço que o dado de entrada, e depois ir convergindo suavemente o mapa. Se o procedimento de inicialização linear for usado, a primeira fase de treinamento pode ser pulada.

Na formação de agrupamentos com redes neurais, os dados de entrada devem estar normalizados em um intervalo [0,1] que permite que estes sejam comparáveis. No caso de valores categóricos, estes devem ser codificados para valores numéricos para apresentação à rede neural.

3.7 Análise de *outliers*

Outliers são objetos de um banco de dados que não estão de acordo com o comportamento geral ou modelo dos dados. São objetos totalmente diferentes ou inconsistentes com o restante dos dados (Barnett e Lewis *apud* [KNO 2002] p. 3).

Segundo Knorr [KNO 2002], dependendo da distribuição dos pontos de dados, uma ou mais das seguintes entidades caracteriza um *outlier*:

- Um valor extremo ou relativamente extremo;
- Um "contaminante", que é uma observação de alguma outra distribuição (possivelmente desconhecida);
- Um valor de dado legítimo, mas surpreendente ou inesperado.
- Um valor de dado que foi medido ou gravado incorretamente.

A idade de uma pessoa mostrada como 999, por exemplo, pode ter sido causada por um valor padrão de um programa para uma idade não apresentada. Outro exemplo é o alto salário de um gerente de uma companhia, que pode ser considerado um *outlier* dentre os salários dos outros empregados da firma [HAN 2001].

Muitos algoritmos de MD tentam minimizar a influência de *outliers* ou eliminá-los. Isto pode resultar na perda de informação escondida importante, em particular, na detecção de fraudes, em que eles podem indicar atividades fraudulentas. Desta forma, a mineração de *outlier*, que consiste na detecção e análise de *outliers* é uma tarefa de MD interessante [HAN 2001]. O ideal é que sejam identificados e que sejam investigados por um especialista do domínio da aplicação [KNO 2002].

Os *outliers* podem ser detectados com a análise de agrupamentos. Por exemplo, podem constituir um grupo com alguns poucos registros que requer, então, investigação. Ou ainda, um grupo com muitos registros de um determinado valor, acima do normal.

Podem ser descritos da seguinte maneira: dado um conjunto de n pontos de dados ou objetos, e k, o número esperado de *outliers*, encontrar os k objetos máximos que são consideravelmente dissimilares, excepcionais ou inconsistentes com relação ao restante dos dados. Pode ser visto como dois subproblemas: (1) definir que tipo de

dados pode ser considerado como inconsistente em um determinado conjunto de dados, e (2) encontrar um método eficiente para minerar os *outliers* assim definidos [HAN 2001].

Dentre as técnicas mais conhecidas estão a detecção de *outlier* baseada em estatística, a detecção de *outlier* baseada em distância e detecção de *outlier* baseada em desvio. O detalhamento dessas técnicas se encontra em [HAN 2001].

3.8 Considerações

Sobre a categorização dos métodos de agrupamento, a literatura relata que alguns algoritmos integram as idéias de vários métodos, de forma que nem sempre é possível classificá-los como pertencente a uma única categoria desses métodos. Além disso, algumas aplicações podem ter critérios de clusterização que requerem a integração de diversas técnicas de agrupamento.

Com relação à utilidade da análise de agrupamentos nos diversos tipos de aplicações, observa-se que é difícil comparar os méritos de um método sobre o outro sem a análise do efeito de tal método sobre a aplicação.

Os diversos métodos apresentados foram bem sucedidos em algum tipo de aplicação e seus resultados dependem da escolha dos atributos e de parâmetros, tais como pesos atribuídos a esses atributos, o número de agrupamentos desejado e diversos outros parâmetros solicitados pelo método, conforme será mostrado neste trabalho.

O próximo capítulo apresenta a metodologia de MD que será utilizada no estudo de caso.

4 Metodologia

Este capítulo apresenta a metodologia utilizada para a mineração de dados no estudo de caso analisado neste trabalho. O objetivo é explorar o uso de técnicas de aprendizado não-supervisionado em MD com a utilização de métodos de agrupamento de dados e avaliar a adequação destes para a descoberta de padrões médios de comportamentos de interesse e identificação de irregularidades em uma base de dados real da área de saúde.

A ferramenta escolhida para a criação dos modelos de mineração foi o *software IBM DB2 Intelligent Miner for Data* © (IM).

Para a execução do processo de MD, procurou-se uma metodologia que facilitasse a sua realização. Diversos modelos de metodologia são mencionados na literatura, e apesar de possuírem nomes diferentes, a idéia central é basicamente a mesma: primeiro o problema e os dados são definidos, depois os dados são preparados e os modelos são construídos e avaliados. Finalmente, o conhecimento é consolidado e implementado para solucionar o problema. A maioria dessas metodologias se destina a organizar o processo de DM o qual é realizado, tradicionalmente, em etapas ordenadas conforme mostra a Figura 4.1 apresentada em [FAY 96].

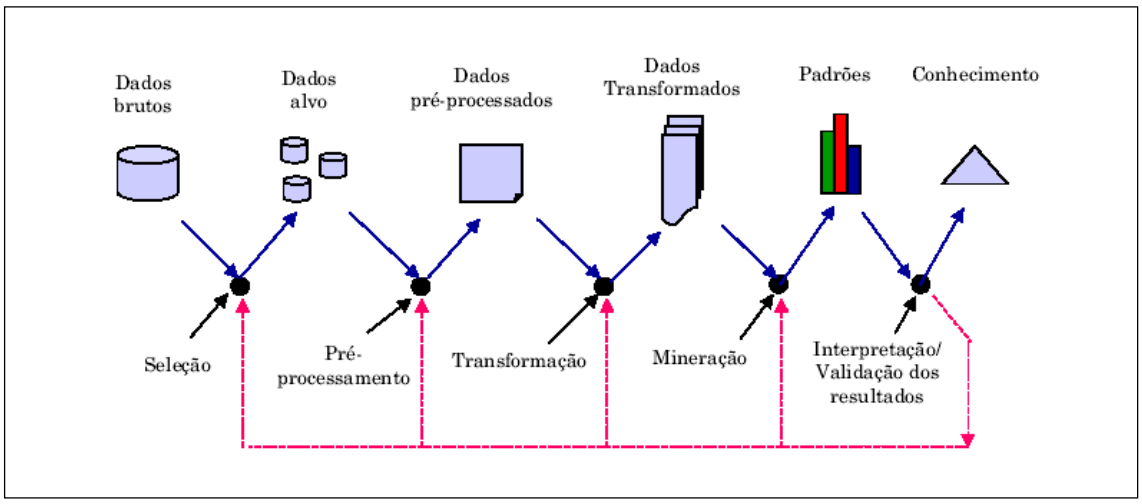


FIGURA 4.1 – Processo de mineração de dados.
Fonte: FAY 96, p. 10.

A metodologia formalizada pelo *Cross Industry Process Model for Data Mining* (CRISP_DM) [CHA 99], utilizada neste trabalho, organiza o processo de DM mostrado na Figura 4.1 em um modelo de processo hierárquico que parte de um conjunto de tarefas mais gerais para um conjunto de tarefas mais específicas, discriminadas em quatro níveis de abstração [CHA 99]:

a) no topo da hierarquia, o processo de MD é organizado em *fases;*

b) as fases, por sua vez, são constituídas por diversas *tarefas genéricas*, que formam o segundo nível da hierarquia;

c) o terceiro nível, de *tarefas especializadas*, envolve a descrição de como as ações das tarefas genéricas são aplicadas em situações específicas. Por exemplo, uma tarefa genérica do segundo nível é a limpeza de dados. No terceiro nível, essa tarefa seria descrita em diferentes situações, tais como limpeza de valores numéricos ou de valores categóricos;

d) o quarto nível, de *instâncias do processo*, é um registro das ações, decisões e resultados da mineração de dados de uma aplicação em particular.

A metodologia CRISP-DM permite o mapeamento de modelos genéricos para modelos especializados com a utilização de *Contextos de Mineração de Dados*, que possuem quatro dimensões diferentes [CHA 99]:

1) O *domínio da aplicação*, que é a área específica da aplicação.

2) O *tipo de problema de MD*, que descreve as classes específicas de objetivos que são tratadas pelo projeto de MD.

3) O *aspecto técnico*, que cobre questões específicas em MD, as quais descrevem diferentes desafios (técnicos) que costumam ocorrer durante a MD.

4) As *ferramentas e técnicas*, que especificam qual(is) ferramenta(s) de MD e/ou técnicas são aplicadas durante o projeto de MD.

Segundo esse mapeameto, o contexto de MD para o problema de agrupamentos e sua aplicação em banco de dados da saúde pode ser representado conforme a Tabela 4.1:

TABELA 4.1 – Dimensões dos contextos de MD para o problema da análise de agrupamentos em dados da saúde.

Dimensões do Contexto de Mineração de Dados			
Domínio da aplicação	**Tipo de problema de MD**	**Aspecto técnico**	**Ferramenta e técnicas**
Descoberta de padrões médios de procedimentos de interesse	Análise de agrupamentos	Agrupamento de registros similares	IBM Intelligent Miner – Pesquisa de Agrupamento Demográfico ou Pesquisa de Agrupamento Neural
Detecção de desvios ou irregularidades	Análise de agrupamentos	Análise de *Outliers* Visualização dos agrupamentos	IBM Intelligent Miner – Pesquisa de Agrupamento Demográfico ou Pesquisa de Agrupamento Neural

Fonte: Baseado em [CHA 99]. p. 3.

A metodologia CRISP-DM descreve o ciclo de vida de um projeto de MD em uma seqüência de seis fases. Esta seqüência não é estrita e costuma apresentar avanços e retornos entre as diferentes fases. Estas serão descritas a seguir, procurando-se direcioná-las, quando necessário, para o contexto da análise de agrupamentos [CHA 99].

4.1 Fases da metodologia

4.1.1 Compreensão do domínio da aplicação

A primeira fase se refere à *compreensão do domínio da aplicação*. Abrange o entendimento dos objetivos do projeto e requisitos, sob uma perspectiva da aplicação, transformando esse conhecimento em uma definição do problema de MD e gerando um projeto preliminar para alcançar os objetivos.

Nesta fase, é essencial a participação dos analistas com conhecimento do domínio da aplicação e dos analistas de dados, que juntos identificarão os pontos relevantes para a aplicação. É preciso chegar a uma definição clara do problema para que a aplicação seja bem sucedida.

Ao final desta fase, deve ser feita a escolha de ferramentas e técnicas que serão utilizadas para a clusterização, levando-se em conta os tipos de dados permitidos para os atributos de entrada, a métrica de similaridade utilizada e a forma como são visualizados os resultados para a análise. Conforme a ferramenta escolhida, será necessário colocar os dados de entrada em um formato específico. Essa escolha é importante no processo, uma vez que influenciará o projeto inteiro de MD.

É preciso considerar também a utilização da ferramenta escolhida pelos usuários do sistema. Se as implementações não forem usáveis, todo o projeto de MD será inútil.

4.1.2 Compreensão dos dados

A próxima fase é a *compreensão dos dados*, que inicia com a coleta de dados, prossegue com a familiarização destes, com a identificação de problemas de qualidade dos dados, com a descoberta dos primeiros aspectos sobre os dados ou a detecção de subconjuntos interessantes como ponto de partida do processo.

Como a fase seguinte é a mais problemática do processo de MD, visto que na maioria das vezes os dados não foram modelados para a mineração, resultando em dados inconsistentes, com ruídos ou ausentes, faz-se necessária, nesta fase, uma revisão geral na estrutura de dados e algumas medidas de suas qualidades.

Para variáveis categóricas, por exemplo, ferramentas que oferecem gráficos setoriais podem, rapidamente, mostrar a contribuição de cada valor para essas variáveis e ajudar a identificar valores inválidos, ausentes ou com desvios. Para variáveis quantitativas, medidas estatísticas como máximo e mínimo, média, moda, mediana e

outras, são meios potentes para determinar a presença de dados inválidos ou com desvio [CAB 97].

A Estatística Bivariada é uma das funções que a ferramenta *Intelligent Miner* (IM) utilizada neste trabalho, oferece para a exploração e análise de dados. A Figura 4.2 mostra um exemplo dessa função aplicada sobre o objeto de dados INTERNAÇÕES 2000, em que são mostradas as estatísticas dos atributos das internações bloqueadas, ou seja, aquelas sob suspeita de apresentar impropriedades. Maiores detalhes sobre o IM são apresentados na Seção 5.1.4 do Capítulo 5.

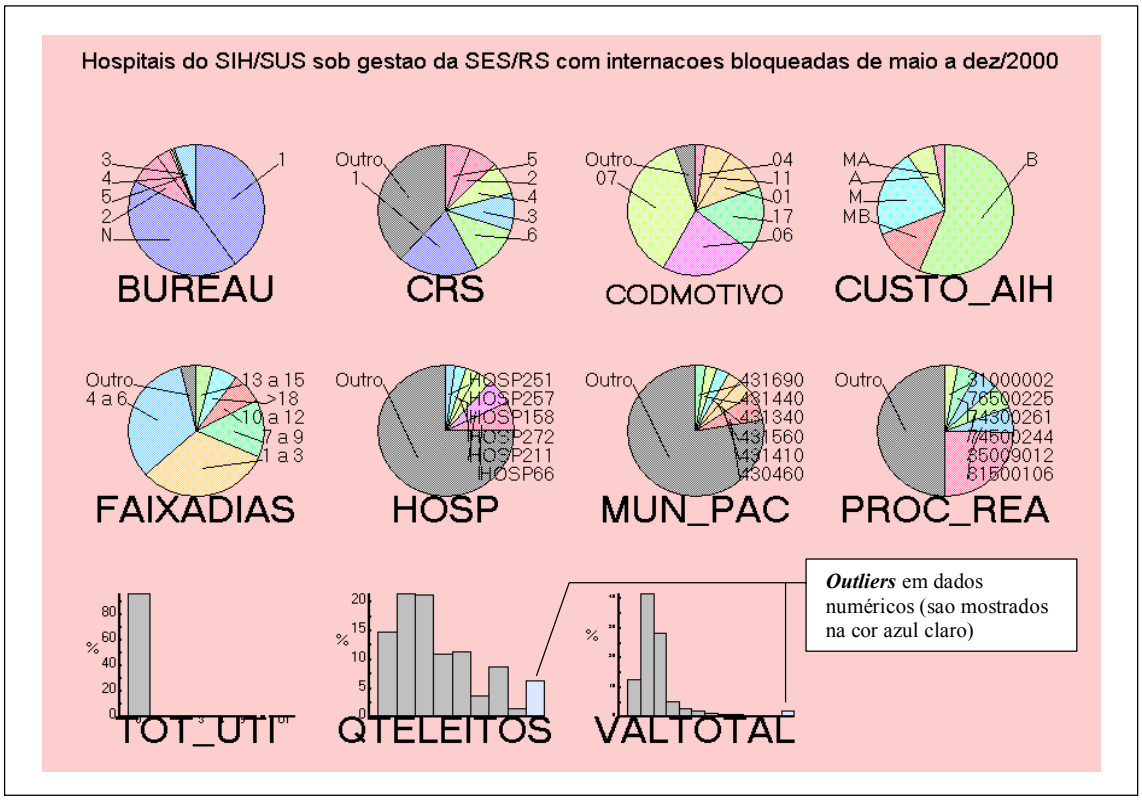


FIGURA 4.2 – Exemplo de um resultado da função Estatística Bivariada do IM.

O gráfico estatístico gerado mostra os atributos categóricos em setores e os atributos numéricos em histogramas. Cada uma dessas representações pode ser expandida para que possa ser melhor observada.

Por exemplo, a Figura 4.3 exibe os detalhes para TOT_UTI (total de dias de internação em UTI) e VALTOT (valor total da AIH). O primeiro gráfico mostra uma coluna que indica que 95,82% dos registros têm valores iguais a zero, significando que foi descoberto apenas um pequeno número de impropriedades em internações com UTI. Mostra ainda, no eixo horizontal, os rótulos OT (*outliers*), com uma pequena quantidade de internações com duração de muitos dias, MV (*missing values* ou valores ausentes) e IV (valores inválidos), que no caso, não ocorreram. No segundo gráfico, vê-se que a maioria das internações sustadas tem custo entre R$ 200,00 e R$ 400,00. A coluna mais à direita com a cor diferente (azul claro) indica que valores acima de R$2.400,00 são bastante elevados em relação aos outros (*outliers*). Observações como estas devem ser consideradas em uma aplicação de MD.

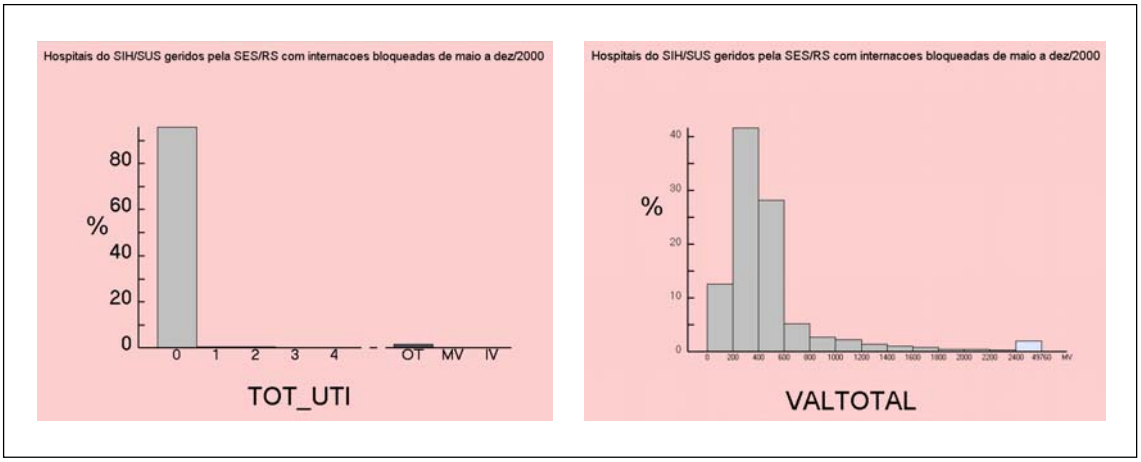


FIGURA 4.3 – Exemplo da função Estatística Bivariada do IM: detalhe para campos numéricos.

A Estatística Bivariada do IM gera também relatórios estatísticos com informações tais como: tamanho absoluto e tamanho relativo do conjunto de dados; valores modais de cada atributo; valores mínimos e máximos, média e desvio padrão dos campos numéricos.

Outras considerações da fase de compreensão dos dados se referem à hierarquia do conjunto de dados: no caso de conjuntos de dados múltiplos, como estes estão relacionados entre si? Que tipos de agregações de dados serão necessários? Quais os diferentes valores das variáveis e suas distribuições? Que valores são típicos e quais são erros? Novamente nesta fase os especialistas do domínio poderão informar sobre o que deve ser procurado, o que deve ser ignorado, que tipo de soluções eles gostariam de ter como resposta [VES 2000].

4.1.3 Preparação de dados

A seguir, vem a fase de *preparação de dados*, que envolve todas as atividades para a construção de um conjunto de dados final, extraído dos conjuntos de dados iniciais, que irá subsidiar a ferramenta de modelagem. As tarefas dessa fase costumam ser realizadas diversas vezes e não seguem uma ordem pré-determinada.

Esta é a fase mais exaustiva do processo e consome aproximadamente 60% dos esforços do processo de MD. Inclui a seleção de atributos, registros e tabelas, limpeza de dados, que consiste na remoção de ruídos e inconsistência dos dados; integração de dados, etapa em que fontes de dados múltiplas podem ser combinadas; transformação de dados, referente à transformação ou consolidação de dados em formas apropriadas para a mineração, como, por exemplo, pela realização de operações de sumarização ou agregação.

As técnicas de agrupamento diferem das outras técnicas de MD por apresentarem algoritmos sensíveis a atributos redundantes e irrelevantes. Assim, é desejável que a ferramenta escolhida apresente algoritmos direcionados a ignorar subconjuntos de atributos que descrevem cada instância ou a designar pesos para cada variável [CAB 97].

Muitas vezes, os dados apresentam valores errados que devem ser corrigidos ou retirados do conjunto de dados. Esses valores são detectados pelo algoritmo de MD como *outliers* e seu impacto na modelagem pode ser considerável. Os *outliers*, em valores simbólicos, podem ser detectados por apresentarem poucas instâncias em relação às centenas de instâncias de outros valores. Esses registros não devem, entretanto, ser retirados do BD, visto que a informação das outras variáveis será perdida [VES 2000].

Na fase de preparação de dados, é importante formalizar o processo, de tal maneira que este possa ser facilmente repetido ou desfeito, se possível. Este cuidado é necessário, uma vez que a natureza iterativa da MD requer a reexecução e reconfiguração, além da aplicação para novos modelos de dados [VES 2000].

4.1.4 Modelagem

A fase de m*odelagem* inclui a seleção e aplicação das técnicas de modelagem, bem como sua parametrização. Algumas técnicas possuem requisitos específicos de formatação de dados e, normalmente, ocorre o retorno para a fase de preparação de dados.

Neste estudo, essa etapa consiste em criar modelos de MD com o emprego da ferramenta escolhida, para aplicar um ou mais métodos de agrupamentos sobre os dados previamente selecionados para a aplicação.

Diversos parâmetros são informados ao sistema. Esses parâmetros podem ser configurados várias vezes, até que se encontre um resultado satisfatório. Uma vez obtidos os padrões médios dos procedimentos de interesse, a detecção de desvios pode ser feita pela análise de *outliers* e também pela visualização dos resultados dos agrupamentos.

A fase de modelagem será melhor entendida com o exemplo mostrado a seguir, que resume os principais passos de sua execução com a ferramenta *Intelligent Miner*.

- **A função de pesquisa Agrupamento do IM**

O IM apresenta duas técnicas de formação de agrupamentos que consistem nas funções de pesquisa: agrupamento demográfico e agrupamento neural.

1) Agrupamento demográfico

O agrupamento demográfico determina automaticamente o número de agrupamentos a ser gerado. As semelhanças entre os registros são determinadas pela comparação dos valores de seus campos. Os agrupamentos são então definidos para que o Critério Condorcet seja maximizado. Este critério, descrito no Capítulo 3, Seção 3.6.1, "*é a soma de todas as semelhanças nos registros de pares do mesmo agrupamento menos a soma de todas as semelhanças no registro de pares de agrupamentos diferentes*" [IBM 99].

Com o algoritmo demográfico, as medidas de distância recaem em uma faixa que vai de 0 a 1. Uma partição perfeita atingirá um valor de Condorcet global e valor de Condorcet para similaridade intraclasse igual a 1 e para cada similaridade interclasse igual a 0 [SOF 2002].

Neste exemplo, o objetivo é encontrar o porte dos hospitais sob a gestão da SES/RS, com base em sua quantidade de leitos.

Passo 1: Definição dos dados

O primeiro passo é definir um objeto de dados do IM que aponte para o arquivo plano **hospit.txt**, que é o arquivo de dados de hospitais residente no servidor do IM. Na janela principal do IM (Figura 4.4) deve ser escolhido o ícone CRIAR DADOS, que ao ser acionado, abrirá a janela do Assistente de Dados. Na página DEFINIÇÕES, desse assistente, deve ser informado o formato dos dados, neste caso, arquivo plano, e o nome do objeto de dados, que neste exemplo será HOSPITAIS. Na página ARQUIVOS PLANOS, será informado o nome e a localização do arquivo plano.

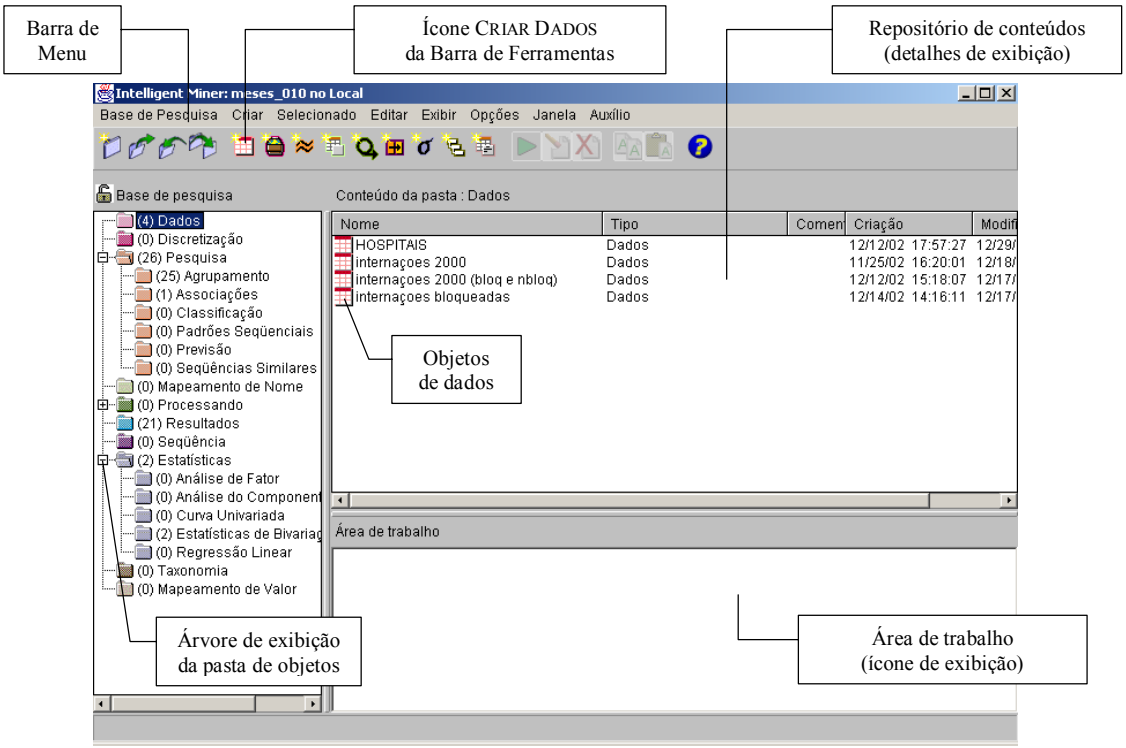


FIGURA 4.4 – A janela principal do Intelligent Miner.

Na página PARÂMETROS DE CAMPO, as propriedades de HOSPITAIS contidas nos dados devem ser especificadas, isto é, seus tipos de dados e as colunas dos arquivos planos que elas ocupam (Figura 4.5). A última página do Assistente de Dados é a página RESUMO, que resume os parâmetros definidos para o objeto de dados que está sendo criado. A opção ENCERRAR salva o objeto de dados na Base de Pesquisa e levará o usuário de volta à janela principal do IM.

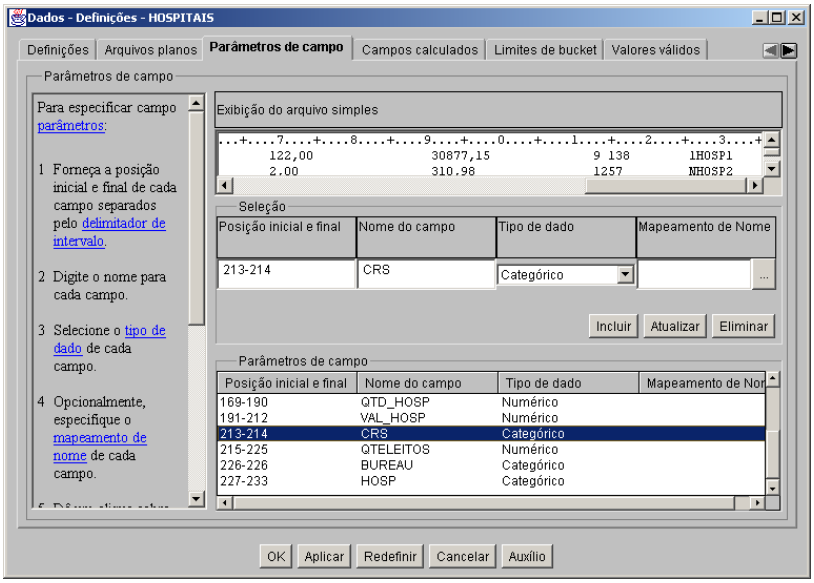


FIGURA 4.5 – Parâmetros de campo.

Passo 2: Criação de um modelo

Neste passo, deve ser escolhido o ícone CRIAR PESQUISA, na janela principal do IM, e na página DEFINIÇÕES E FUNÇÕES DE PESQUISA do Assistente de Pesquisa será selecionado AGRUPAMENTO DEMOGRÁFICO, informando-se o nome do objeto de definições de pesquisa PORTE DOS HOSPITAIS (Figura 4.6).

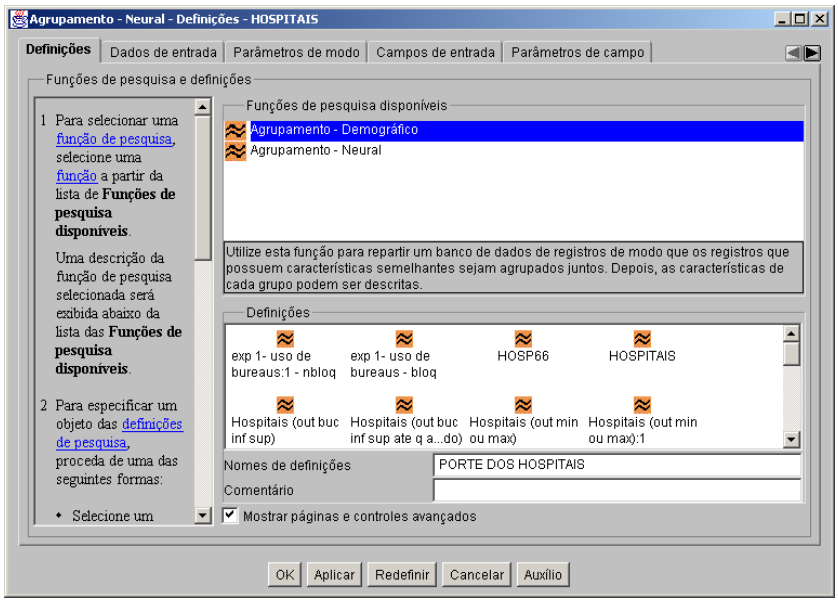


FIGURA 4.6 – Objeto de definições de pesquisa.

Na página DADOS DE ENTRADA do assistente será selecionado o objeto de dados HOSPITAIS. A próxima página solicita os parâmetros de modo, cujas opções são: modo de agrupamento[11] e modo de aplicação[12]. Neste caso, será selecionado o modo de

[11] Modo de agrupamento: neste modo, a função de pesquisa tenta formar um agrupamento satisfatório e cria o número especificado de agrupamentos [IBM 99].

agrupamento. Depois disso, são requisitados outros parâmetros ao usuário para controlar os resultados, que são [IBM 99]:

- *Número máximo de passagens* sobre os dados de entrada: um número limitado de passagens reduz o tempo de processamento, mas também reduz a acurácia. Um número mais alto aumenta a qualidade dos agrupamentos, mas aumenta também o custo de desempenho. Normalmente, duas ou três passagens são suficientes. O valor padrão do IM é de 2 passagens.

- *Número máximo de agrupamentos* a ser gerado pelo algoritmo: um número pequeno aumenta o desempenho, mas limita a homogeneidade dos agrupamentos e a acurácia da solução completa. Um número alto aumentará o tempo de execução. Esse parâmetro é opcional e caso não especificado, o algoritmo determina automaticamente um número ótimo de agrupamentos a ser gerado. O valor padrão do IM é de 9 agrupamentos máximos.

- *Melhora na precisão*: representa a porcentagem de melhora do agrupamento a cada passagem pelos dados. É usado como um critério de interrupção. Se a melhoria real for menor do que o valor especificado, não ocorre mais nehuma passagem. Quanto menor o valor, mais preciso é o agrupamento. O valor padrão do IM é de 2%.

- *Limite de semelhança*: limita os valores aceitos como o melhor ajuste de um agrupamento. Por exemplo, um limite de semelhança definido em 0,25, os registros com 25% de valores de campo idênticos serão provavelmente atribuídos ao mesmo agrupamento. Para obter um número maior de agrupamentos, o valor do limite de semelhança deve ser aumentado. O valor padrão do IM é de 0,5.

A página CAMPOS DE ENTRADA do Assistente de Pesquisa lista os campos disponíveis do objeto de dados. Devem então ser especificados os campos ativos e os campos suplementares desse objeto de definições de pesquisa.

Os *campos ativos* são usados como critério para determinar se os registros são ou não semelhantes.

Os *campos suplementares* têm suas estatísticas incluídas no resultado, mas estas não são usadas para determinar semelhanças entre os registros.

Os campos de um agrupamento são ordenados por importância e se os campos suplementares aparecerem entre os campos ativos, significa que influenciaram a criação do agrupamento talvez mais do que aqueles selecionados para campos ativos.

O objeto de definições de pesquisa permite também a configuração de parâmetros avançados, tais como:

- *Parâmetros de campo*, em que a *medição de um campo* oferece um peso maior ou menor a certos campos ativos durante o processo de agrupamento.

[12] Modo de aplicação: a função de pesquisa atribui IDs do agrupamento aos registros de dados, em um modelo que tenha sido criado previamente em modo de agrupamento [IBM 99].

O valor padrão do IM é 1 para todos os campos, indicando que todos têm pesos iguais.

- *Parâmetros de campo adicionais*, que permitem a configuração de uma medida de semelhança, para dados numéricos, que varia de 0 a 1. Valores próximos a 0 indicam valores distantes e valores próximos a 1 indicam valores idênticos. Uma medida de semelhança igual a 0,5 reflete valores separados por uma medida de distância.

- *Definições de similaridade*, que possibilitam, através do mapeamento de valor, uma correspondência entre itens nos dados de entrada e em uma tabela de intervalo.

Neste exemplo, foram mantidos os valores padrões fornecidos pelo IM. A Tabela 4.2 exibe as definições que resumem os parâmetros básicos e avançados da função de pesquisa agrupamento demográfico para este exemplo:

TABELA 4.2 – Definições de parâmetros para a função de pesquisa Agrupamento Demográfico.

Página do Assistente	**Parâmetro**	**Valor**
Função de pesquisa	Nome Comentário Função de pesquisa	PORTE DOS HOSPITAIS Agrupamento demográfico
Dados de entrada	Dados de entrada	HOSPITAIS
Parâmetros de modo (valores padrões)	Modo de uso No. de passagens máx No. de agrup máx Melhora na precisão	Modo de agrupamento 2 9 2
Campos de entrada	Ativos Suplementares	QTELEITOS (quant. de leitos) BUREAU (cód. do *bureau*) CRS (cód. da Coord. Reg. de Saúde)
Parâmetros de campo (valores padrões)	Peso de campo Medição de valor Compensar	1 Nenhum Não
Parâmetros de campo adicionais (valores padrões)	Modificar a unidade Modificar o fator de distância	Sem valores Sem valores
Tratamento de *outlier*	Tratamento de *outlier*	Substituir *outlier* por mín ou máx
Matriz de semelhança	Seleção de campos ativos Mapeamento de valor	Sem valores Sem valores
Campos de saída		Não criar saída
Resultados	Nome dos resultados Comentário	PORTE DOS HOSPITAIS

Na última página do assistente, é especificado o nome do objeto de resultados. O IM executa o objeto de definições e exibe um indicador de progresso para monitorar o *status* da função de pesquisa, o qual exibe a data e hora de início, o tempo de execução, as etapas da pesquisa, o número de passagens sobre os dados, o número de agrupamentos criados e o valor do Condorcet (Figura 4.7).

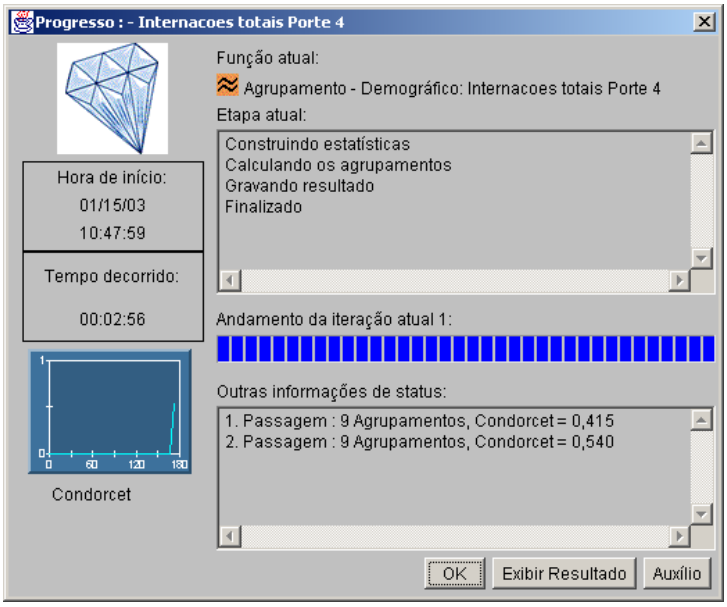


FIGURA 4.7 – Indicador de progresso para monitorar o status da função pesquisa do IM.

A Figura 4.7 mostra o indicador, não deste exemplo, que apresentava poucos dados e o indicador não mostrou as etapas da pesquisa, mas de uma outra pesquisa, sobre um conjunto mais volumoso de dados. A finalidade é chamar a atenção para as etapas realizadas durante a mineração: primeiro são construídas as estatísticas dos atributos, seus valores modais formam o centro do conjunto exemplo que será o ponto de partida para a criação dos agrupamentos na primeira passagem sobre os dados. A cada passagem do algoritmo pelos dados de entrada, os centros são ajustados para que a qualidade do modelo de agrupamento total seja alcançada, conforme se observa com a melhora do valor do Condorcet.

Ao final da execução, o IM exibe o objeto de resultados gerado por esse objeto de definições (Figura 4.8).

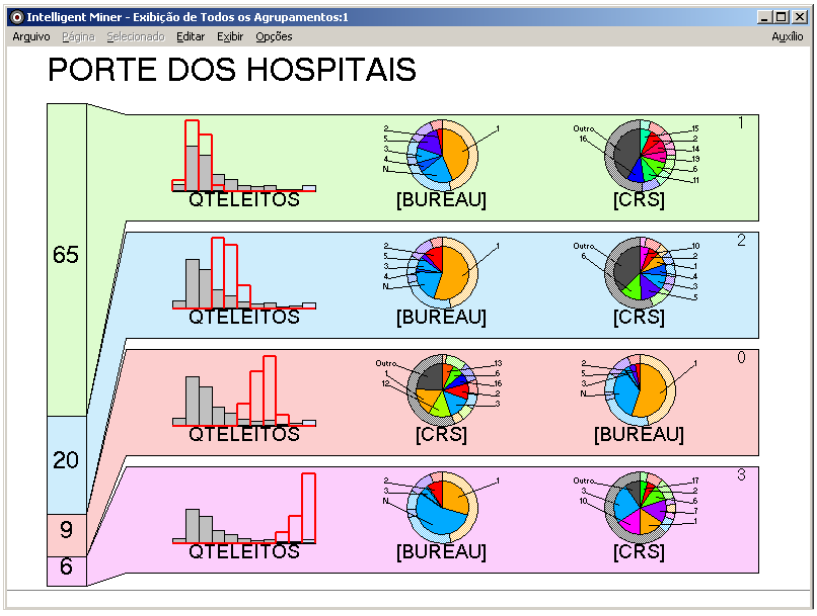


FIGURA 4.8 – Resultado do agrupamento demográfico gerado pelo IM.

Passo 3: Análise dos resultados

No objeto de resultados gráfico mostrado na Figura 4.8, as linhas múltiplas do gráfico mostram os grupos descritos no resultado. A tela mostra 4 linhas, cada qual representando um dos 4 grupos identificados pela execução da mineração. Dentro de cada grupo, os gráficos setoriais representam os campos usados, em que os campos que mais influíram sobre a formação do grupo são apresentados à esquerda. Os números do lado esquerdo representam o tamanho do grupo em porcentagem, por exemplo, o grupo de cima representa 65% dos dados. Os números do lado direito representam a ID (identificação) do grupo.

O Agrupamento 1, que aparece na faixa mais elevada da Figura 4.8, foi expandido, resultando no gráfico da Figura 4.9, que exibe um histograma (QTELEITOS) e dois gráficos setoriais (BUREAU e CRS).

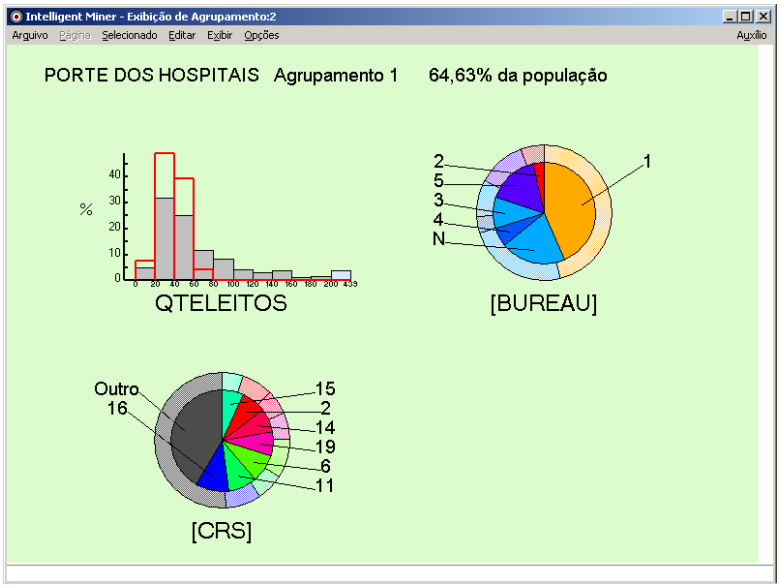


FIGURA 4.9 – Expansão do resultado do agrupamento demográfico gerado pelo IM.

O histograma representa valores numéricos que foram discretizados pelo algoritmo em intervalos de 20 unidades cada, com exceção do último intervalo à direita que é formado por *outliers*. As barras cheias representam a distribuição de leitos em relação ao todo e as barras vermelhas transparentes representam a distribuição de leitos no agrupamento 1.

Cada gráfico setorial mostra duas distribuições: o anel de fora mostra a distribuição relativa à amostra toda, o aro de dentro mostra a distribuição relativa ao grupo a que ele está associado. Por exemplo, no gráfico setorial de BUREAU, o anel de fora representa a distribuição de *bureaux* em todos os dados, e o valor 1 é o de maior tamanho. O aro de dentro representa que esse grupo é também formado em sua maioria por registros cujo valor é 1.

Nos gráficos do IM, é possível julgar a significância de cada variável por meio da comparação dos tamanhos das distribuições no agrupamento e em relação ao todo. Quanto maior a diferença entre as duas distribuições, mais significantes são as variáveis.

Juntamente com os resultados gráficos, são gerados os seguintes relatórios estatísticos: relatório geral de todas as partições (Figura 4.10), relatórios de cada partição ou agrupamento e relatórios de cada campo exibido na visualização. Nesses relatórios, o item *Características do Campo de Referência* apresenta o valor modal, o qual representa o centro dos agrupamentos criados.

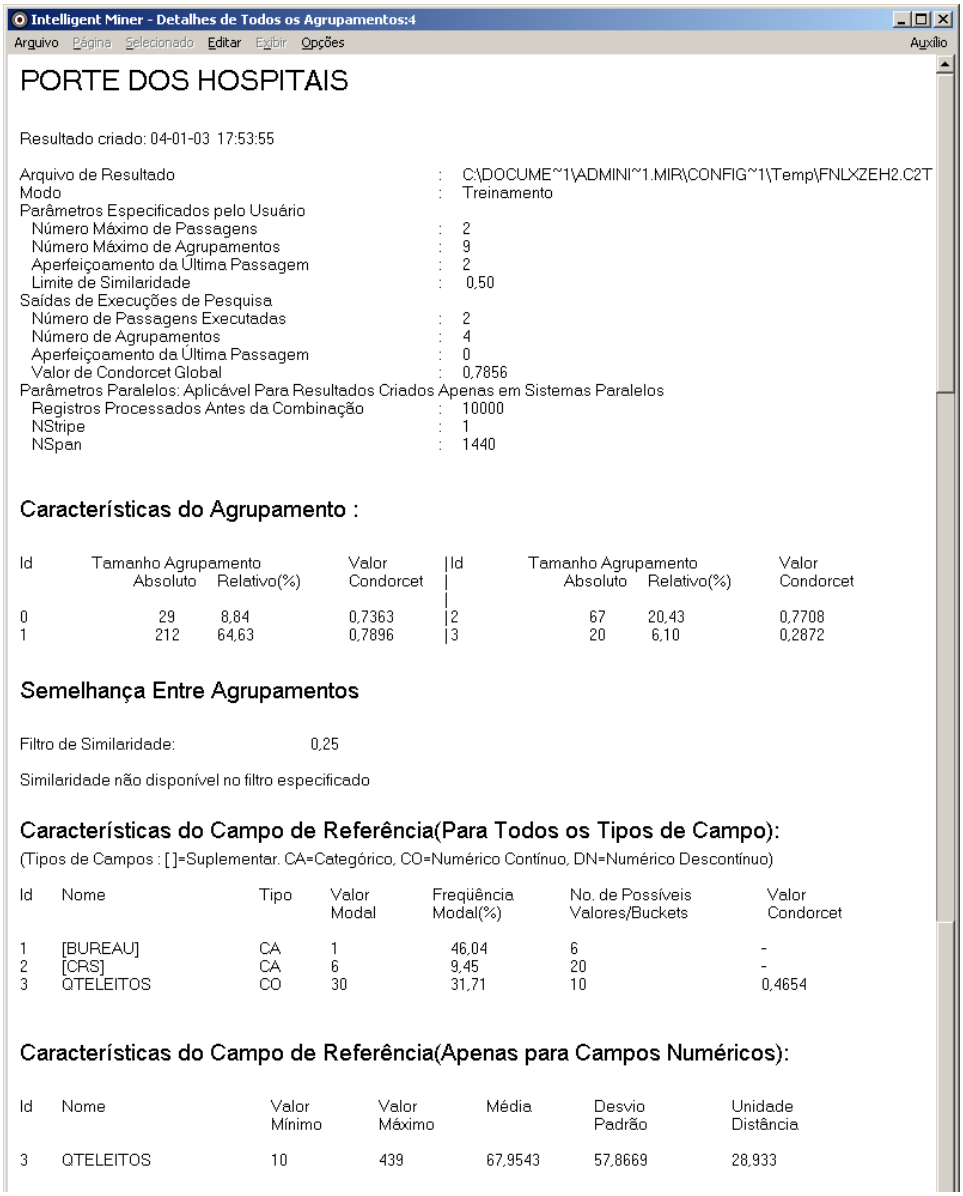


FIGURA 4.10 – Detalhes dos resultados contendo informações estatísticas de todas as partições.

O Assistente de Pesquisa de Agrupamento permite, também, a criação de um arquivo de saída de dados, o qual permite a inclusão de campos de entrada selecionados pelo usuário e solicita que seja especificado o campo obrigatório *ID do Agrupamento*, cujos valores serão o identificador do agrupamento de melhor ajuste para o registro de entrada correspondente.

Sugere também a criação de dois campos opcionais: *Contagem de Registro*, que conterá os valores do registro de entrada correspondente de melhor qualidade para o

agrupamento de melhor ajuste e *Confiança*, que conterá os valores de confiança calculados quando os registros de entrada foram atribuídos aos agrupamentos. O objeto de definição de dados de saída poderá estar no formato texto ou como tabelas ou *views* de banco de dados, servindo como entrada para outros *softwares*.

2) Agrupamento Neural

O Agrupamento Neural utiliza a rede neural de Mapas Auto-Organizáveis ou Mapa de Kohonen, cujo algoritmo foi citado no Capítulo 3, Seção 3.6.5, em que se utiliza um processo de auto-organização para agrupar registros de entradas semelhantes.

A tarefa principal do agrupamento neural é encontrar um centro para cada agrupamento, o qual também é chamado de protótipo do agrupamento. Para cada registro nos dados de entrada, a função de pesquisa Agrupamento Neural calcula o protótipo do agrupamento que se encontra mais próximo do registro [IBM 99].

A contagem de cada registro de dados é representada pela distância euclidiana do protótipo do agrupamento. As contagens mais próximas a zero possuem um grau mais alto de semelhança com o protótipo do agrupamento, ao passo que uma contagem mais alta representa maior dissemelhança entre o protótipo e o registro [IBM 99].

Os passos de utilização do Agrupamento Neural são semelhantes aos do Agrupamento Demográfico, com exceção da definição para alguns parâmetros tais como [IBM 99]:

Parâmetros de modo:

- *Número máximo de passagens* sobre os dados de entrada: um número limitado de passagens reduz o tempo de processamento, mas também reduz a acurácia. Um número mais alto aumenta a qualidade dos agrupamentos, mas aumenta também o custo de desempenho. Cinco a dez passagens normalmente são suficientes. O valor padrão do IM é de 5 passagens.

- *Número máximo de agrupamentos* a ser gerado pelo algoritmo: no modo de agrupamento, a função de pesquisa cria o número de agrupamentos especificado na forma de uma grade retangular, em que o *número máximo de linhas* é uma dimensão da grade e o *número máximo de colunas* é a outra dimensão da grade. A grade é igual ao número máximo de linhas e menor ou igual ao número máximo de colunas fornecidas. Os valores padrões do IM são de 3 linhas e 3 colunas no máximo ou 9 agrupamentos máximos.

O agrupamento neural do IM não apresenta os parâmetros de campo, os parâmetros de campo adicionais, matriz de semelhança e parâmetros paralelos, que o agrupamento demográfico apresenta.

No agrupamento neural, os dados de entrada devem ser normalizados ou representados em uma escala de 0,0 a 1,0. É mais indicado para valores numéricos. No caso de valores categóricos, estes devem ser convertidos em um código numérico para apresentação à rede neural [IBM 99].

O IM oferece a opção de normalizar os dados de entrada, representando em escalas campos numéricos contínuos e descontínuos em uma faixa de 0,0 a 1,0 e converte dados categóricos em 1-de-*N* vetores. Em um campo categórico com *N* valores diferentes, um valor de entrada é convertido em um índice exclusivo *i*. No vetor de entrada interno de tamanho *N*, um único valor 1 será colocado na posição do valor de índice do campo, e os valores 0 serão colocados em todas as outras posições do vetor. Significa que um valor categórico com um número grande de valores descontínuos pode causar uma expansão no número de unidades de entrada na rede neural [IBM 99].

A Tabela 4.3 exibe as definições que resumem os parâmetros da função de pesquisa agrupamento neural para o mesmo exemplo mostrado anteriormente:

TABELA 4.3 – Definições de parâmetros para a função de pesquisa Agrupamento Neural.

Página Assistente	**Parâmetro**	**Valor**
Função de pesquisa	Nome Comentário Função de pesquisa	PORTE DOS HOSPITAIS Agrupamento neural
Dados de entrada	Dados de entrada	HOSPITAIS
Parâmetros de modo (valores padrões)	Modo de uso No. de passagens máx No. máximo de linhas No. máximo de colunas	Modo de agrupamento 5 3 3
Campos de entrada	Ativos Suplementares	QTELEITOS (quant. de leitos) BUREAU (cód. do *bureau*) CRS (cód. da Coord. Reg. de Saúde)
Tratamento de *outlier*		Substituir *outlier* por mín ou máx
Campos de saída		Não criar saída
Resultados	Nome dos resultados Comentário	PORTE DOS HOSPITAIS

4.1.5 Avaliação

Nesse estágio, foram construídos modelos que aparentam ter alta qualidade, sob a perspectiva da análise de dados. Antes de partir para a aplicação final do modelo, é importante avaliar e rever todo o processo para se ter certeza de que foram atingidos os objetivos da aplicação. Se todas as questões importantes foram suficientemente consideradas. Ao final desta fase, deve ser tomada uma decisão sobre os resultados alcançados [CHA 99].

A avaliação, na análise de agrupamentos, consiste na identificação de padrões realmente interessantes que representem conhecimento baseado em algumas medidas de interesses. Esta fase, na verdade, ocorre alternadamente com a fase de modelagem ou mineração, visto que é comum aplicar o método de mineração, avaliar os resultados, validar esses resultados com os especialistas e, normalmente, repetir o processo.

A apresentação do conhecimento descoberto é feita com técnicas de visualização e representação de conhecimento que são usadas para apresentar o conhecimento modelado ao usuário.

Muitas vezes, a descoberta de agrupamentos revela padrões interessantes, mas que requerem um refinamento com a utilização de outras técnicas de MD, como por exemplo, classificação ou regras associativas.

4.1.6 Aplicação

A fase final é a *aplicação*, em que o conhecimento obtido deve ser organizado e apresentado ao usuário de forma que este possa usá-lo com facilidade. Conforme os requisitos do projeto, esta fase pode compreender desde uma simples geração de relatórios até a implementação de um processo de MD reusável em novas aplicações. O usuário final tem papel importante em cumprir as ações necessárias para a utilização dos modelos criados.

4.2 Considerações

A Figura 4.11 resume as fases descritas anteriormente, com suas tarefas genéricas (em negrito) e seus produtos (em itálico). São mostradas todas as etapas do modelo, no entanto, nem todas precisam ser realizadas em uma aplicação. Devem ser realizadas aquelas que se aplicam a um determinado projeto de mineração de dados [CHA 99]:

Compreensão do domínio	Compreensão dos dados	Preparação de dados	Modelagem	Avaliação	Aplicação
Determinação dos objetivos da aplicação *Cenário* *Objetivos da aplicação* *Critérios de sucesso da aplicação* **Situação a ser avaliada** *Inventário de recursos* *Requisitos, suposições e limitações* *Terminologia* *Custos e benefícios* **Determinação das metas de MD** *Metas* *Critérios de sucesso de MD* **Produção do projeto** *Projeto* *Avaliação inicial de ferramentas e técnicas*	**Coleta de dados inicial** *Relatório da coleta de dados inicial* **Descrição dos dados** *Relatório da descrição dos dados* **Exploração dos dados** *Relatório da exploração dos dados* **Verificação da qualidade dos dados** *Relatório de qualidade dos dados*	*Conjunto de dados* *Descrição do conjunto de dados* **Seleção de dados** *Racionalizar para inclusão/ exclusão* **Limpeza de dados** *Relatório da limpeza de dados* **Construção de dados** *Atributos derivados* *Registros gerados* **Integração de dados** *Dados mesclados* **Formatação de dados** *Dados reformatados*	**Seleção da técnica de modelagem** *Técnica de modelagem* *Suposições de modelagem* **Geração do projeto de teste** *Projeto de teste* **Construção do modelo** *Configurações de parâmetros* *Modelos* *Descrição dos modelos* **Modelo a ser avaliado** *Avaliação do modelo* *Configurações de parâmetros revisadas*	**Avaliação dos resultados** *Avaliação dos resultados de MD em função dos critérios de sucesso da aplicação* *Modelos aprovados* **Revisão do processo** *Revisão do processo* **Determinação dos próximos passos** *Lista de possíveis ações* *Decisões*	**Aplicação do projeto** *Plano de aplicação* **Plano de monitoramento e manutenção** *Monitoramento e manutenção do plano* **Produção do relatório final** *Relatório final* *Apresentação final* **Revisão do projeto** *Documentação da experiência*

FIGURA 4.11 – Fases, tarefas genéricas e produtos do Modelo de Referência do CRISP-DM.
Fonte: CHA, 99. p. 8.

Uma observação importante sobre a fase de pré-processamento é que existem maneiras automatizadas que ajudam a verificar a qualidade dos dados, mas é importante entender a origem destes e o que contêm. Em alguns casos, somente o conhecimento do domínio poderá realmente esclarecer o significado de um atributo ou variável. De outra forma, por melhores que sejam as técnicas de modelagem, todo o trabalho ficará prejudicado.

O próximo capítulo descreve o estudo de caso realizado com a utilização da metodologia apresentada, as principais dificuldades encontradas durante o processo de MD e os achados obtidos pela utilização das técnicas de agrupamentos de dados.

5 Estudo de Caso

Este trabalho se propõe a realizar a mineração de dados com a utilização de métodos de agrupamento sobre uma base de dados real da área de saúde, a base de dados das internações hospitalares do RS controladas pela Secretaria Estadual de Saúde do Rio Grande do Sul (SES) no período de maio a dezembro de 2000.

A base de dados em estudo possui uma grande quantidade de dados, alta dimensionalidade e tipos diferentes de dados, constituindo um domínio rico em informações e de alta complexidade.

O período de maio a dezembro de 2000 foi estabelecido em virtude do controle automatizado dos registros bloqueados pelos auditores ter iniciado em maio/2000, registros estes que compõem a base de dados em estudo.

O processo de MD foi realizado segundo a metodologia apresentada no capítulo anterior. Na fase de modelagem, foram gerados modelos de mineração de dados com a utilização de agrupamento, que foram reunidos em dois experimentos, realizados de acordo com os objetivos discriminados abaixo:

1. Experimento 1: avaliar os algoritmos de agrupamento demográfico e neural do *Intelligent Miner*, quanto à sensibilidade dos parâmetros de tratamento de *outliers*, e como isto se aplica aos parâmetros reais dos dados da saúde.

2. Experimento 2: identificar as tendências das internações hospitalares com pagamento bloqueado pela auditoria médica da SES. A idéia é observar os agrupamentos e verificar se os padrões encontrados conduzem a resultados indicativos de pontos em que os critérios de bloqueio técnico utilizado possam estar falhando ou indicativos da criação de novos critérios.

Com relação à metodologia utilizada, foi verificado, na prática, que uma alteração na ordem de execução da tarefa genérica de seleção de dados produzirá melhores resultados para a base de dados estudada, conforme será visto adiante.

A seguir, serão descritas as principais etapas da aplicação de mineração de dados neste estudo de caso.

5.1 Compreensão do domínio da aplicação

5.1.1 Determinação dos objetivos da aplicação

A expectativa da aplicação de técnicas de MD para a solução de problemas reais, como os da área da saúde, é de que esta poderá resultar em economia de grande quantidade de recursos financeiros governamentais e melhorar as condições do sistema de saúde através da otimização do emprego dos recursos existentes, além de promover a democratização de tecnologias da informação avançadas [ENG 2000].

- **Cenário**

A Constituição Federal de 1988 instituiu o Sistema Único de Saúde (SUS), em que o custeio da Assistência à Saúde no país é descentralizado, permitindo aos Estados e Municípios mecanismos de organização e controle dos sistemas locais e regionais de saúde. Os Estados e Municípios que estão habilitados em *Gestão Plena do Sistema* recebem os recursos diretamente em seus Fundos de Saúde, ao passo que os não habilitados são custeados pelo *Teto Financeiro da Assistência* e o valor do custeio é determinado pelo Ministério da Saúde (MS) [RIO 2000].

A SES é responsável pela gestão Estadual dos municípios custeados pelo Teto. Esta atua no recebimento das faturas dos serviços, avalia e autoriza que o pagamento seja efetuado pelo MS. Em 2000, apenas 10 municípios possuíam gestão plena no RS, ou seja, não estavam sob a responsabilidade da SES [RIO 2000].

Para a gestão mencionada acima, a SES utiliza o Sistema de Informações Hospitalares do SUS (SIH/SUS), implantado em âmbito nacional, que tem como instrumento a Autorização de Internação Hospitalar (AIH), para o registro de todos os dados pertinentes às internações hospitalares. O fluxo geral do sistema pode ser observado na Figura 5.1.

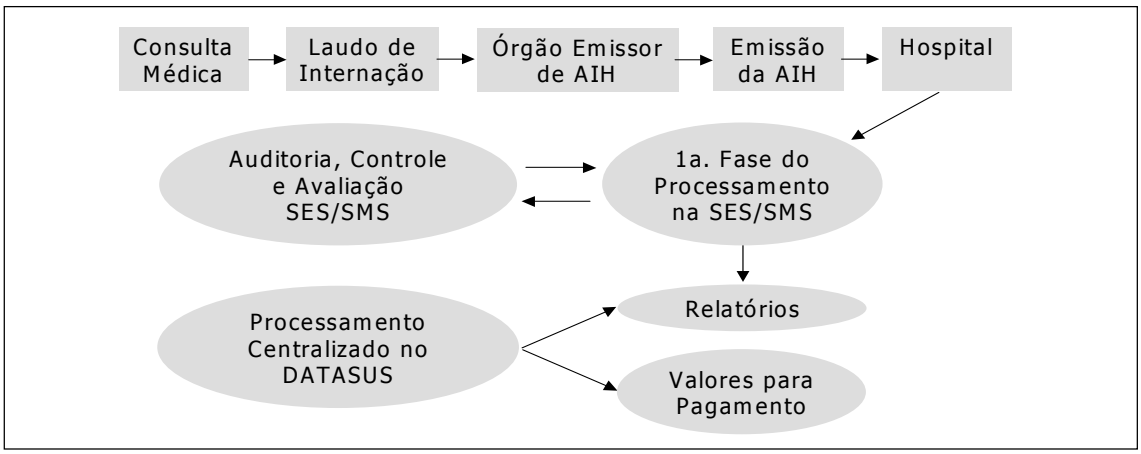


FIGURA 5.1 – Fluxo do SIH/SUS.
Fonte: SUS, 2001. p. 3.

A base de dados da SES referente ao SIH/SUS possui uma enorme quantidade de dados sobre a movimentação das internações dos municípios sem gestão plena do Estado. Mensalmente, a Auditoria Médica Estadual, após realizar uma análise técnica nas AIHs, bloqueia o pagamento de um certo número de internações por apresentarem alguma impropriedade. Essa estratégia é utilizada pela SES com a finalidade de manter o pagamento dos serviços dentro do teto financeiro definido pelo MS na cobrança das internações.

No ano de 2000, a equipe de controle, avaliação e auditoria da SES elegeu os seguintes critérios técnicos de bloqueios para as AIHs apresentadas: septicemia, cuidados prolongados, politraumatizados, cirurgias múltiplas, transplante, AVC agudo e homônimos. Significa, por exemplo, que se o motivo de uma internação for septicemia,

esta é automaticamente submetida a algumas regras. Se for bloqueada, é separada para análise pelos auditores. Além das auditorias pelos critérios técnicos, que representaram 97,17% do total no referido ano, outros tipos de auditoria foram os seguintes: laudos médicos excedentes, denúncias ou situações com suspeita de irregularidades, auditorias especiais solicitadas pelas Coordenadorias Regionais de Saúde e Secretarias Municipais, que juntos representaram 2,83% do total realizado [RIO 2000].

O trabalho de auditoria é intenso, uma vez que os bloqueios são analisados caso a caso. Conforme relatado pelos auditores, do total de AIHs apresentadas mensalmente, apenas cerca de 40% das mesmas são analisadas. Em 2000, 33% do total das AIHs bloqueadas não foram pagos. Nesse ano, os recursos do Teto Financeiro de Assistência destinados ao RS foram insuficientes para o custeio das despesas e seu déficit teve de ser coberto pelos recursos do Tesouro do Estado.

A Figura 5.2 mostra o fluxo do sistema da auditoria médica da SES para o bloqueio de internações hospitalares segundo os critérios técnicos e as normas do SUS.

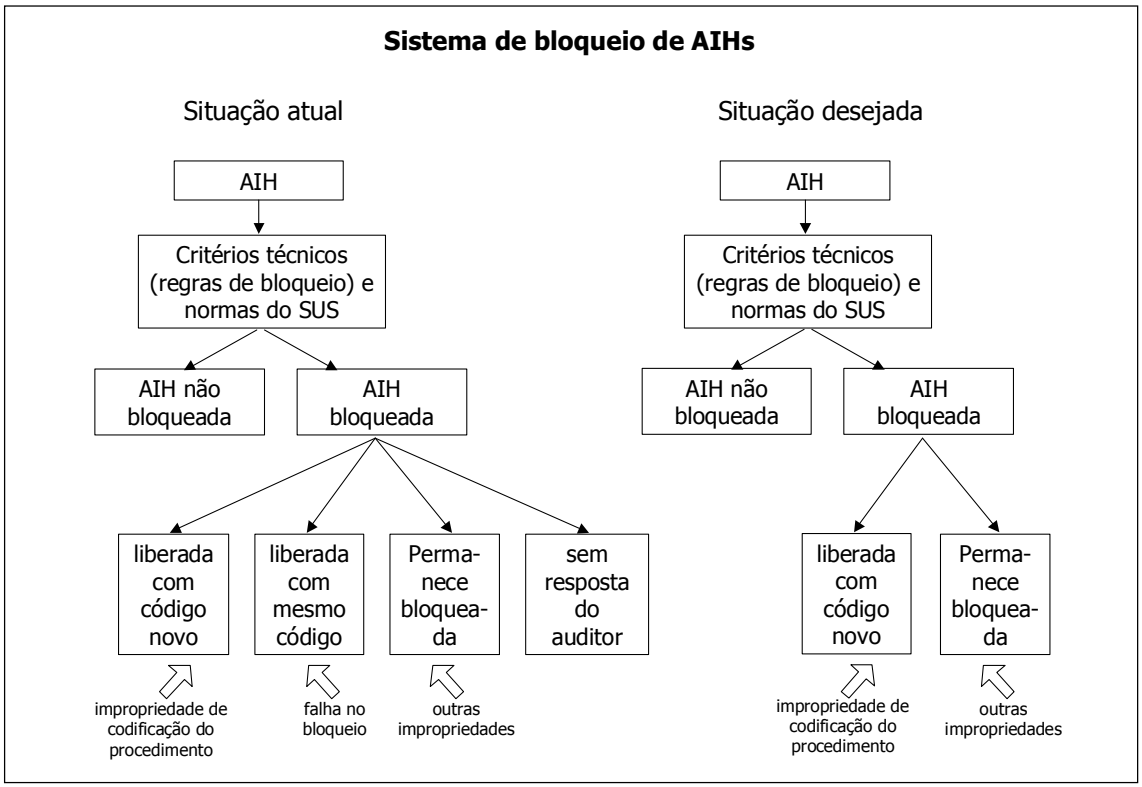


FIGURA 5.2 – Sistema de bloqueio de AIHs utilizado pela SES.

A situação que se apresenta até hoje, é mostrada no lado esquerdo da Figura 5.2. Cada internação cobrada é submetida aos critérios técnicos, que se apresentam sob a forma de regras para bloqueio de uma internação, e também às normas do SUS, que avaliam 81 itens para verificar se há impropriedades, tais como, preenchimento incompleto, rasuras, lançamento de procedimento não realizado e outros. Se a AIH não se enquadrar em nenhum desses critérios ou normas, é encaminhada para pagamento. Caso contrário, ela é bloqueada e ocorre um dos quatro procedimentos após a análise da auditoria:

a) *A AIH é liberada com código novo*: nessa opção estão enquadrados os casos de impropriedade de codificação de procedimento. É cobrado um procedimento com valor maior, enquanto que deveria ter sido cobrado um procedimento com valor menor. Então, a AIH é liberada para pagamento com o código do procedimento de menor valor.

b) *A AIH é liberada com o mesmo código*: neste caso a auditoria não conseguiu identificar nenhuma impropriedade e acaba liberando o pagamento da internação. Segundo a auditoria, o ideal é que só fossem bloqueadas as internações com alguma impropriedade. O grande número dessas internações significa que o sistema de bloqueios precisa ser aperfeiçoado.

c) *A AIH permanece bloqueada*: neste caso, são avaliadas e se apresentarem impropriedades, segundo as normas do SUS, tais como dados incorretos, com rasuras ou procedimentos inexistentes, permanecem bloqueadas e não são pagas.

d) *Sem resposta do auditor*: o auditor não consegue dar um parecer sobre o caso.

Ao lado direito da Figura 5.2, aparece o fluxo da situação idealizada pela auditoria: a de bloquear somente os casos de impropriedades.

Há, portanto, a necessidade de se descobrir meios para que a SES possa melhor auditar as AIHs para que haja melhor aplicação dos recursos públicos.

- **Objetivos da aplicação**

Diante deste cenário, esta aplicação é voltada para a busca de meios que possam otimizar o trabalho dos auditores, com o emprego de MD baseado na aplicação da análise de agrupamentos.

Seus objetivos foram estabelecidos e nortearam um ciclo completo do processo de MD. Esses objetivos são os seguintes:

- Identificar as tendências das internações hospitalares com pagamento bloqueado pela auditoria médica da SES. Observar os agrupamentos e verificar se os padrões encontrados conduzem a resultados indicativos de pontos em que os critérios de bloqueio técnico utilizado possam estar falhando ou indicativos da criação de novos critérios.

- **Critérios de sucesso**

Para o alcance dos objetivos da aplicação, foram estabelecidos como critérios importantes a aquisição do conhecimento dos especialistas, a coleta de dados relevantes e consistentes com a minimização da perda ou incorreção destes, bem como a utilização de ferramentas que permitissem a MD com análise satisfatória dos resultados, de forma a detectar descobertas importantes sobre o assunto.

Na prática, foi verificada a dificuldade na aquisição de conhecimento dos especialistas. As reuniões, inicialmente realizadas somente com os médicos auditores, mostraram a necessidade do envolvimento de técnicos de informática da SES, uma vez que foi preciso esclarecer dúvidas sobre os bancos de dados estudados. Essa equipe foi se modificando com o passar do tempo e as opiniões sobre quais fatores influenciam na descoberta de padrões interessantes também. Desta forma, a subjetividade das atividades da auditoria médica conduziu a uma análise do problema que, por diversas vezes teve seus objetivos redirecionados, até que se identificasse o que realmente se queria abordar.

Houve também dificuldades com a qualidade os dados, conforme será relatado adiante.

Quanto à utilização da ferramenta escolhida, esta foi bem sucedida na modelagem e na apresentação dos resultados, ficando apenas prejudicada a utilização de suas tarefas de pré-processamento, o que também será explicado no item oportuno.

5.1.2 Situação a ser avaliada

Nesta etapa, os recursos necessários para a aplicação foram relacionados em um inventário e sua adequação para atingir os objetivos e metas almejados foi verificada. Foram também identificados requisitos, suposições e limitações sobre a situação a ser avaliada.

- **Inventário dos recursos**

Os recursos disponibilizados para a análise que será aqui apresentada são os seguintes:

TABELA 5.1 – Inventário dos recursos.

Recursos	Descrição
Humanos	Equipe de especialistas em MD, constituída por professores e alunos de pós-graduação da UFRGS, envolvidos no projeto "*Desenvolvimento de Metodologia para Extração de Conhecimento de Bases de Dados da Saúde do Estado para Avaliação e Planejamento*", convênio UFRGS/SES. Equipe de especialistas do domínio da aplicação, constituída por auditores médicos e técnicos em informática do Setor de Auditoria Médica da SES.
Dados	Bases de dados obtida junto à SES.
Documentos	Relatório de dados, normas, legislações, manual do SIH/SUS, relatório de atividades da SES/2000, outros relatórios.
Hardware	Rede de computadores do Instituto de Informática da UFRGS e os equipamentos adquiridos pelo projeto, que se encontram no referido Instituto, com as seguintes configurações: Clientes: • Microcomputadores Pentium 3 com 191 MB de RAM, HD de 20 GB, drive de 3 ½ ", drive de CD. • Microcomputadores Pentium 3 com 522 MB de RAM, HD de 20 GB, drive de 3 ½ ", drive de CD, gravador de CD.

Recursos	Descrição
Hardware	Servidor: • Microcomputador Pentium 4 com 522 MB de RAM, HD de 40 GB, drive de 3 ½ “, drive de CD e o sistema operacional Windows 2000 Server.
Sistemas Operacionais	Microsoft Windows 98 Microsoft Windows 2000 Professional Microsoft Windows 2000 Server
Software	Ferramenta de para a mineração de dados: IBM DB2 Intelligent Miner for Data. Outros *softwares*: Microsoft Office 2000.

Foi possível constatar durante o desenvolvimento do processo de MD que a disponibilidade dos recursos adequados também é um fator decisivo para o sucesso da aplicação. Todos os itens citados acima são de importância fundamental para essa finalidade.

- **Requisitos, suposições e limitações**

O trabalho requisitou não apenas recursos pessoais e materiais, mas também outros, como por exemplo, permissão para utilizar as bases de dados, obtidas por intermédio do projeto “*Desenvolvimento de Metodologia para Extração de Conhecimento de Bases de Dados da Saúde do Estado para Avaliação e Planejamento*”, realizado como colaboração entre pesquisadores da UFRGS e a SES. Foi ainda necessária a definição de situações de interesse mais específicas.

Para este estudo, uma situação de interesse foi a análise das tendências de hospitais no que se refere às internações hospitalares realizadas nos meses de maio a dezembro de 2000, bem como a análise do perfil de internações bloqueadas, nas quais se concentram as situações de impropriedades ocorridas no sistema.

Apesar do grande número de regras de integridade implementadas no banco de dados, como por exemplo, a filtragem de procedimentos que se aplicam somente ao sexo feminino ou a uma determinada faixa de idade, a maioria dos casos de impropriedades só é detectada com a observação caso a caso, e depende muito da experiência médica do auditor.

No entanto, a mineração de dados com a utilização das técnicas de agrupamento, pode evidenciar padrões interessantes que podem alertar os auditores para situações que ainda não foram percebidas e que podem constituir novos critérios de bloqueio ou tornar os critérios existentes mais eficazes, conforme será mostrado nos experimentos deste estudo de caso.

As limitações identificadas neste estudo se referem a informações julgadas importantes para a MD, que ainda não foram disponibilizadas nas bases de dados, dificuldades para integrar bases de dados, pela falta de um identificador único que facilitaria essa operação e a falta de documentação detalhada do sistema de controle das AIHs bloqueadas.

5.1.3 Metas de mineração de dados

A mineração deverá resultar em agrupamentos ótimos de subconjuntos da base de dados, que apresentem acurácia e que apresentem uma avaliação de resultados satisfatória, tanto quantitativa, como qualitativamente. Os resultados devem ser de fácil visualização para os usuários do sistema, devem sugerir ações úteis e revelar padrões desconhecidos ou validar hipóteses que os auditores desejem confirmar.

5.1.4 Avaliação inicial de ferramentas e técnicas

Conforme foi visto no Capítulo 3, existem inúmeras técnicas e algoritmos para a análise de agrupamentos em MD. Assim, uma das etapas deste trabalho foi fazer a busca de ferramentas que implementassem essas técnicas. Procurou-se na *internet* aquelas de domínio público, isto é, *freeware* ou *shareware*. Muitas ferramentas estão disponíveis na *Web*, no entanto, as limitações impostas, como por exemplo, um número de registros bastante reduzido ou um tempo de uso restrito a poucos meses, inviabilizaram o uso dessas ferramentas nesta pesquisa.

Algumas das ferramentas analisadas para uso, no início do trabalho, apresentavam os resultados em relatórios e gráficos de difícil compreensão pelo usuário, outras necessitavam de um trabalho extra para codificar os dados de entrada, e assim foram abandonadas.

Em geral, as ferramentas são classificadas em três categorias: comercial, de domínio público e protótipos de pesquisas.

A ferramenta selecionada para este trabalho foi o *software* comercial IBM DB2 Intelligent Miner for Data © (IM), que pode ser utilizado de forma completa por membros e pesquisadores de instituições de ensino cadastrados no "*IBM Scholars Program*". A versão utilizada foi a Versão 6 Release 1. O motivo fundamental da escolha, no entanto, foi o de que o IM contempla os requisitos de mineração do problema estudado, conforme será visto a seguir.

- **IBM DB2 Intelligent Miner for Data (IM)**

O IM é uma suíte de ferramentas de mineração da IBM que se caracteriza por apresentar, dentre outras coisas: uma ampla seleção de algoritmos de mineração de valor comprovado; a interoperabilidade entre algoritmos, de tal forma que os resultados de um algoritmo podem ser passados como entrada para outro algoritmo; escalabilidade, suportando grandes volumes de dados. Também oferece ferramentas de visualização de dados para a exibição e interpretação de resultados [CAB 97].

Em geral, utiliza uma arquitetura cliente/servidor, em que a mineração é realizada no servidor e a definição dos dados e interpretação dos resultados são realizadas no cliente. A API (*interface* do programa de aplicação) fornece uma *interface* às funções que são acessíveis pelo cliente e assim acionam a execução das funções no servidor. O *software* do servidor é executado nos sistemas operacionais AIX, AS/400,

OS/390, Sun Solaris e Windows NT. Os clientes podem utilizar AIX, OS/2 e Windows [IBM 99].

O IM oferece uma série de funções estatísticas, de pré-processamento e de pesquisa (mineração) de dados, as quais podem ser empregadas independentemente, iterativamente ou com uma combinação dessas duas formas. O processo completo para a mineração de dados utilizado pelo IM suporta [IBM 99]:

– Preparação de dados

Os dados de entrada podem estar em formato de arquivos planos (arquivos em formato texto), tabelas ou *views* importados dos bancos de dados IBM DB2, Oracle e Sybase e outras fontes de dados relacionais. Disponibiliza os tipos de dados relacionados a seguir, os quais define como:

- *binário*: dados medidos pela escala de medida nominal e que admitem somente dois valores possíveis, 0 e 1;
- *categórico*: dados que são medidos pela escala de medida nominal;
- *contínuo*: dados que são medidos pela escala de proporção ou intervalo. Nesse caso, todos os valores da extensão do campo são divididos em *buckets*, aos quais são atribuídos valores separados. Por definição, a extensão de aproximadamente ±2 desvios padrão da média é dividida em 10 *buckets*. O número exato de *buckets* varia de acordo com a função utilizada;
- *numérico-discreto*: dados que são medidos pela escala de proporção ou intervalo. Todo valor de campo é tratado como está, sem nenhum processamento adicional. É útil quando se quer saber quantos valores diferentes existem em um determinado campo e com que freqüência esses valores ocorrem. Tem a desvantagem de que as menores diferenças entre os valores são tratadas como observações distintas, conduzindo a resultados muito grandes e detalhados, o que acaba ofuscando as informações que se encontram nos dados;
- *numérico*: dados que são medidos pela escala de proporção ou intervalo. É uma combinação do tipo de dados contínuo e do numérico-discreto. Se o campo contiver até 50 valores diferentes, este será tratado como numérico-discreto, caso contrário, será tratado como um campo contínuo.

A qualquer momento, podem ser usadas funções estatísticas para explorar e analisar os dados. Essas funções servem ainda para transformar os dados e criar campos de entrada para a pesquisa, além de terem utilidade na avaliação dos dados de saída gerados pela função de pesquisa. As funções que o IM oferece são: Regressão Linear, Ajuste da Curva Univariada, Análise de Componentes Principais, Análise de Fator e Estatística Bivariada.

Os dados de entrada podem ser transformados por meio de funções de pré-processamento, tais como agregação de valores e cálculo de novos valores com o uso de SQL, conversão de minúscula ou maiúscula, cópia e indexação de registros para arquivos planos, remoção ou codificação de valores nulos ou ausentes, discretização em quantis e intervalos, filtragem de campos e registros, obtenção de amostras aleatórias,

agrupamento de registros, ligação de fontes de dados relacionais, mapeamento de valores de entrada, campos calculados por funções matemáticas, divisão de registros de dados de entrada para múltiplos registros.

– Mineração de dados

As técnicas de modelagem utilizadas pelo IM para a mineração de dados são apresentadas como funções de pesquisa e são as seguintes: associação, classificação neural, classificação em árvore, agrupamento demográfico, agrupamento neural, padrões seqüenciais, seqüências similares, previsão neural e previsão RBF (Função de Base Radial).

– Análise dos resultados e assimilação do conhecimento

Os resultados são avaliados em relação aos objetivos da aplicação. As ferramentas de visualização permitem a exibição dos resultados e a identificação das informações importantes reveladas pelo processo. Os resultados podem ser exportados para exibição em uma estação de trabalho remota, no formato de arquivo de texto ou de tabelas ou *views* de banco de dados, ou copiados para a área de transferência para que fiquem disponíveis a outras ferramentas e podem também ser impressos.

5.2 Compreensão dos dados

Ao primeiro contato com os dados do SIH, verificou-se que os mesmos são de grande complexidade para o entendimento de pessoas estranhas ao sistema. No decorrer do trabalho, diversas reuniões com os especialistas do domínio da aplicação foram necessárias para a compreensão destes, a fim de que fosse possível realizar o pré-processamento dos mesmos.

- **Dados da coleta inicial**

Os conjuntos de dados para esta pesquisa foram obtidos junto à SES e coletados do Sistema de Internações Hospitalares do Estado do Rio Grande do Sul. Esses dados se referem apenas aos meses de maio a dezembro de 2000. Estão armazenados em diversos arquivos, discriminados na Tabela 5.2.

A obtenção de documentos que pudessem ajudar a compreender os dados e o domínio estudado também foi necessária. Alguns desses documentos foram: o manual do SIH/SUS, relatório de atividades da SES/2000, modelos de AIHs, descrição dos dados, critérios de bloqueios de AIH, relatórios de procedimentos de maior freqüência, procedimentos de maior custo médio e procedimentos de maior valor pago.

TABELA 5.2 – Arquivos de dados fornecidos pela SES no formato DBF.

Arquivo	Descrição	N° de objetos	N° de atributos
DSMS010	Movimento das AIHs.	375.408	75
DSMS020	Procedimentos especiais autorizados de AIH nos municípios do Estado no período.	86.905	9
DSMS030	Atos profissionais autorizados de AIH nos municípios do Estado no período.	2.250.372	14
DSMS040	Movimento dos hospitais com informações de lançamentos (pagamentos e descontos).	392	9
DSMS160	Valores da AIH (faturamentos cobrados).	375.408	24
DAIH050	Tabela de Atos.	4.899	15
DAIH150	Diagnósticos de acordo com a tabela CID.	14.196	6
CONTROLE	Controle anual dos registros bloqueados.	14.282	12
BUREAU	Hospitais distribuídos por Bureau	256	2
LEITOS	Hospitais distribuídos pelo número de leitos.	379	5

- **Descrição dos dados**

Os diversos arquivos de dados foram examinados e foi produzido, pela equipe do Projeto, um novo Relatório de Dados que contém a descrição geral de cada arquivo, bem como o nome e a descrição do significado de cada atributo.

- **Exploração dos dados**

Para a compreensão dos atributos das internações hospitalares realizadas, foram elaborados alguns gráficos com a estatística descritiva das internações no período estudado, bem como das internações bloqueadas, e também dos hospitais.

Com relação aos atributos que poderiam nortear a pesquisa, levou-se muito tempo para identificá-los e foram necessárias inúmeras entrevistas com os especialistas do domínio até que se pudesse chegar aos atributos mais interessantes para a MD. Os atributos mais utilizados estão relacionados na Tabela 5.3:

TABELA 5.3 – Principais atributos dos conjuntos de dados das internações hospitalares.

Atributo	Tipo	Descrição	Valores
APRES	Categórico	Mês de apresentação da AIH para pagamento	62000; 72000; 82000; 92000; 102000; 112000; 122000; 12001.
BUREAU	Categórico	Nome do Bureau (empresa que administra as internações de um determinado hospital)	1; 2; 3; 4; 5 e N: NENHUM.
COD_MOTIVO	Categórico	Código do motivo de bloqueio	00: Não bloqueada. 01 a 21: Motivos.
COD_TIPO	Categórico	Código do tipo de bloqueio	0: Não bloqueada; 1: Sustada; 2: Libera AIH com código novo; 3: Libera AIH com mesmo código; 4: Sem resposta do auditor.
CRS	Categórico	Código da Coordenadoria Regional de Saúde	1 a 19

Atributo	Tipo	Descrição	Valores
CUSTO_AIH	Categórico	Custo total da AIH	MB (muito baixo): <=200 reais; B (baixo): >200 e <=500 reais; M (médio): >500 e <=1000 reais; A (alto): >1000 e <=2000 reais; MA (muito alto): >2000 reais.
DIAG_PRI	Categórico	Código do diagnóstico principal, segundo a CID	14.196 diagnósticos.
ESPEC	Categórico	Código da especialidade da AIH	1 a 9: Especialidades.
FAIXADIAS	Categórico	Faixa com o número de dias que o paciente ficou internado.	0 (não completou 1 dia); 1 a 3 dias; 4 a 6 dias; 7 a 9 dias; 10 a 12 dias; 13 a 15 dias; 16 a 18 dias; >18dias.
GRUPOCID	Categórico	Código do grupo de doenças, conforme o diagnóstico segundo a CID	GP01 a GP21.
HOSP	Categórico	Código do hospital	HOSP1 a HOSP328.
PORTE	Categórico	Código do porte do hospital	1: 10 a 64 leitos; 2: 65-111 leitos; 3: 112-164 leitos; 4: 172 a 439 leitos.
PROC_REA	Categórico	Código do procedimento médico realizado	
QTELEITOS	Numérico	Quantidade de leitos de um hospital	10 a 439 leitos.
VALTOTAL	Numérico	Valor total da AIH	R$ 0,00 a 112.926,00.

- **Verificação da qualidade dos dados**

Em relação à qualidade dos dados, inúmeros problemas foram detectados, incluindo dados inconsistentes e com ruídos, dados ausentes, dados com valores de atributos diferentes, mas com o mesmo significado e a falta de identificadores únicos nas diversas tabelas, o que dificultou bastante o pré-processamento dos mesmos. Esses problemas serão abordados nas diversas etapas da preparação de dados.

5.3 Preparação de dados

Ao início da fase de preparação de dados, ainda não havia sido feita a escolha da ferramenta de MD. Assim, as diversas etapas de preparação de dados foram realizadas com a utilização do *software* Microsft Access. A idéia foi formar um conjunto único de dados que contivesse todos os atributos relevantes para a aplicação, o qual seria levado para a ferramenta de mineração. Ao final, optou-se por gerar dois conjuntos de dados, conforme se explica mais adiante. Houve, no entanto, grande dificuldade na formação destes. Alguns dos obstáculos encontrados foram:

- Na ligação de tabelas, descobrir quais atributos identificavam uma AIH, uma vez que essas ligações geravam registros duplicados, os quais não se sabia se eram falsas duplicatas ou duplicatas possíveis.

- As limitações do Microsoft Access para processar consultas com grandes quantidades de dados, o que ocorria com muita lentidão e, freqüentemente, era alcançado o tamanho máximo do banco de dados, sendo necessária a criação de um novo arquivo de banco de dados. As diversas operações com as tabelas de dados foram realizadas mês a mês.

O IM permite a realização de uma série de tarefas de pré-processamento. Para isso, os dados precisam estar em tabelas de bancos de dados do DB2. Porém, no início da utilização da ferramenta, o servidor do sistema não possuía capacidade suficiente para a manipulação das *tablespaces* do DB2. Posteriormente, com a aquisição de um servidor mais potente, não foi mais possível investir na utilização do *software* gerenciador de bancos de dados DB2. No entanto, isto é recomendável, pois permitirá o uso pleno de todas as funcionalidades oferecidas pelo IM.

- **Seleção de dados**

No início do trabalho, antes da limpeza de dados, foram examinados todos os arquivos de dados fornecidos e foram selecionados, no Relatório de Dados, os atributos julgados de maior interesse. Porém, mais tarde, após novas entrevistas com os especialistas, foi constatado que alguns atributos importantes haviam sido eliminados dos conjuntos de dados produzidos para a MD.

Assim, foi preciso reiniciar o trabalho. Porém, dessa vez, foi realizada a integração de dados de tal forma que o arquivo resultante contivesse todos os atributos dos diversos arquivos de dados, exceto aqueles que não apresentavam nenhum valor em todos os registros ou os que apresentavam o mesmo valor em todos os registros.

Como o SIH é bastante complexo, e até mesmo os especialistas divergem sobre a escolha de alguns atributos, julga-se que é importante levar todos os atributos para a ferramenta de mineração para que se possa explorar as diversas visões sobre os dados.

Essa decisão de alteração será particularmente importante se houver a realização de novos ciclos do processo, com novos objetivos, quando outros atributos forem necessários para a mineração.

Com base nesta experiência e na observação de que a metodologia CRISP-DM organiza o processo de MD de forma que a tarefa genérica de seleção de dados é realizada antes da limpeza, construção, integração e formatação de dados, é feita uma sugestão de que esta tarefa seja realizada depois dessas outras tarefas, caso em que se aproxima da descrição das etapas de DCBD apresentadas na Seção 2.1.1 e Figura 2.1 do Capítulo 2 deste trabalho, que representa um processo semelhante ao utilizado em *data warehouses*.

- **Limpeza de dados**

Ao exame minucioso dos diversos arquivos de dados, foram identificados diversos problemas, alguns dos quais são exemplificados na Tabela 5.4.

TABELA 5.4 – Problemas relacionados à limpeza de dados.

Descrição	**Ação**
Atributos que não possuíam valor em nenhum dos registros.	Foram eliminados dos conjuntos de dados preparados para a MD

Descrição	Ação
Atributos que possuíam o mesmo valor em todos os registros.	Foram eliminados dos conjuntos de dados preparados para a MD
O arquivo BUREAU apresentava o atributo CGC com barra (/) antes dos dois dígitos finais em alguns registros e em outros não. Esse fato gerou erro na ligação entre essa tabela e a tabela de internações, pois esse campo determinava o relacionamento entre as duas tabelas.	Remoção de todas as barras, tornando os valores uniformes quanto ao formato.
Registros que em um arquivo pertenciam a um determinado município, em outro, pertenciam a outro município.	Verificado o valor desse atributo em diversas tabelas, optando-se pelo valor que ocorria mais vezes.
Datas com registros contendo ora 4 dígitos, ora 2 dígitos para indicar o ano.	Mudado o formato para data abreviada (ano com dois dígitos)
Os atributos motivo de bloqueio e o tipo de bloqueio apresentavam diversos valores para o mesmo significado.	Foram normalizados.
Atributos importantes com registros sem valores, como por exemplo, a quantidade de leitos.	Estes valores foram solicitados à equipe da SES.

- **Construção ou transformação de dados**

No início e durante o processo de MD foram criados atributos derivados ou que sofreram transformações, como por exemplo, os mostrados na Tabela 5.5.

TABELA 5.5 – Atributos derivados ou que sofreram transformações.

Tipo de derivação ou transformação	Descrição
Discretização – variáveis numéricas ou categóricas foram discretizadas em medidas de escala categórica.	CUSTO_AIH (custo total de uma AIH) – intervalos contendo grupos com base no valor total da internação. FAIXADIAS (período de internação) – intervalos contendo grupos com base no número de dias de internação. GRUPOCID (grupo do diagnóstico principal segundo a CID) - intervalos contendo grupos com base nos diagnósticos principais segundo a CID.
Mapeamento – uma variável nominal recebeu um outro valor, por questões éticas.	CGC – recebeu o nome HOSP1, por exemplo. *BUREAU* - recebeu os valores categóricos de 1 a 5 para o nome dos *BUREAUX* e N para o valor NENHUM.
Operações Matemáticas – diversas variáveis numéricas foram representadas em um só atributo.	VALTOTAL (valor total da AIH) – soma dos valores dos diversos serviços.
Atribuição de valores – atribuição de medidas de escala categórica para um novo valor de atributo, após a integração de tabelas.	CODTIPO (código do tipo de bloqueio) – o valor 0 foi atribuído para os não bloqueados. CODMOTIVO (código do motivo de bloqueio) – o valor 00 foi atribuído para os não bloqueados.

- **Integração de dados**

Para que fosse possível obter determinadas informações de uma internação, foi necessário a mesclagem de vários arquivos de dados em um só arquivo. Deste processo, resultaram não apenas um, mas dois conjuntos de dados para a mineração, os quais estão descritos na Tabela 5.6.

TABELA 5.6 – Conjuntos de dados preparados para a mineração.

Conjunto de Dados	Bases de dados que foram mescladas	No. de Registros	No. de Bloqueios	No. de atributos	Descrição
C1	DSMS04, BUREAU e LEITOS	328	–	11	Hospitais que realizaram internações no período de maio a dezembro/2000
C2	DSMS010, DSMS160, DSMS040, LEITOS, BUREAU, CONTROLE e DAIH050	375.408	14.314	70	Corresponde ao número total de internações realizadas de maio a dezembro/2000.

Esses conjuntos resultaram da união das bases de dados mostradas na Tabela 5.6, cuja descrição se encontra na Tabela 5.2.

O conjunto C1 totalizou 328 hospitais que realizaram internações no período de maio a dezembro/2000.

O conjunto C2 totalizou 375.408 registros, que correspondem ao número de internações realizadas de maio a dezembro de 2000. Desse total, 14.314 foram internações bloqueadas.

- **Formatação de dados**

Sobre esta etapa, observa-se que não há a necessidade de ordenação dos atributos para a realização do agrupamento.

Os dados pré-processados no Microsoft Access devem ser convertidos em arquivos de textos com largura fixa, para que possam servir de entrada na criação de objetos de dados no IM, no qual serão chamados de arquivos planos.

5.4 Modelagem

Esta fase se refere à utilização do IM para a mineração de dados. Foram realizados experimentos com a *Pesquisa de Agrupamento Demográfico*, que utiliza o algoritmo demográfico para a formação de agrupamentos e a *Pesquisa de Agrupamento Neural*, que utiliza mapas auto-organizáveis ou mapas de Kohonen.

Conforme foi citado no início deste capítulo, os experimentos foram agrupados com diferentes objetivos. O Experimento 1 busca a compreensão da técnica, com relação à configuração de parâmetros para tratamento de *outliers*. O Experimento 2 consiste na aplicação da técnica com a finalidade de atender aos objetivos da aplicação sobre os dados da saúde.

5.4.1 Experimentos

Na etapa de mineração dos dados do SIH/SUS, foram realizados inúmeros experimentos, em busca de resultados que atendessem aos objetivos desta pesquisa.

A cada realização de um conjunto de experimentos, estes eram levados aos especialistas da aplicação, que os analisavam e, por diversas vezes, sugeriram mudanças, ora apontando novos atributos para a MD, ora observando o que era e o que não era interessante, validando ou não o conhecimento encontrado.

Ao final do primeiro ciclo de MD, foram selecionados dois experimentos que demonstram a utilização das técnicas de agrupamento neste estudo de caso.

5.4.2 Experimento 1 – Compreensão das técnicas de agrupamento, com relação à configuração de parâmetros para tratamento de *outliers*.

Antes de se analisar os problemas específicos da área da saúde, houve a necessidade de explorar as opções que a ferramenta oferece para o tratamento de *outliers*, uma vez que a análise desses objetos é útil, tanto para a investigação do comportamento de impropriedades, como para o pré-processamento dos dados. Com este experimento foi possível a compreensão desse tipo de parametrização, utilizada para a detecção de *outliers* em atributos numéricos.

O experimento se baseia na análise do atributo porte[13] dos hospitais em relação à quantidade de leitos que estes possuem. Esta informação é importante para que se possa encontrar um perfil dos hospitais e analisar se hospitais do mesmo porte apresentam comportamentos semelhantes ou não, com relação às internações hospitalares. A SES não utiliza, atualmente, a informação de porte de um hospital. Apenas os hospitais com 50 leitos ou menos são considerados hospitais pequenos.

Portanto, este primeiro experimento tem como objetivos:

1) A avaliação da parametrização dos algoritmos demográfico e neural com relação ao tratamento de *outliers*.

2) A distribuição dos hospitais em faixas conforme o número de leitos, ou discretização, com a utilização da pesquisa de agrupamento do IM, definindo assim um porte para os hospitais que realizaram internações no período de maio a dezembro de 2000.

Neste experimento, foi utilizado o conjunto de dados C1, que contém 328 hospitais, com quantidades de leitos que vão de um mínimo de 10 a um máximo de 439 leitos. O campo ativo para agrupamento foi QTELEITOS (quantidade de leitos) e os campos suplementares (apenas ilustrativos) foram HOSP (identificação do hospital) e

[13] O porte dos hospitais é considerado, neste trabalho, apenas com relação à sua quantidade de leitos. Em entrevistas recentes com os especialistas da saúde, foi informado, no entanto, que existem resoluções do MS que estabelecem o porte em função não só da quantidade de leitos, mas de uma série de outros atributos. Porém, essa informação ainda não foi inserida nas bases de dados da SES.

BUREAU (empresa que administra as internações de um determinado hospital). Todos os parâmetros permanecem com valores padrões do IM, com exceção do TRATAMENTO DE *OUTLIERS* que será avaliado em suas quatro opções, com o agrupamento demográfico e em suas três opções, com o agrupamento neural.

Mineração 1.1 – Agrupamento Demográfico com tratamento de *outliers* como valores ausentes.

Foi executada a Pesquisa de Agrupamento Demográfico do IM com o parâmetro TRATAMENTO DE *OUTLIERS* permanecendo com o valor padrão *Tratar outliers como valores ausentes*, opção em que os *outliers* são desprezados durante a execução de um agrupamento.

- **Resultado**

A mineração gerou 5 agrupamentos, conforme se observa na Figura 5.3. O valor do Condorcet global foi de 0,788 e o tempo de mineração foi de 0min02. A utilização do comando Exibir *outlier*, que a ferramenta oferece, permite visualizar os *outliers* no histograma correspondente, em uma barra vertical de cor azul clara. Estes se encontram no agrupamento 4, conforme indica a barra vermelha transparente situada no intervalo em que os *outliers* estão posicionados.

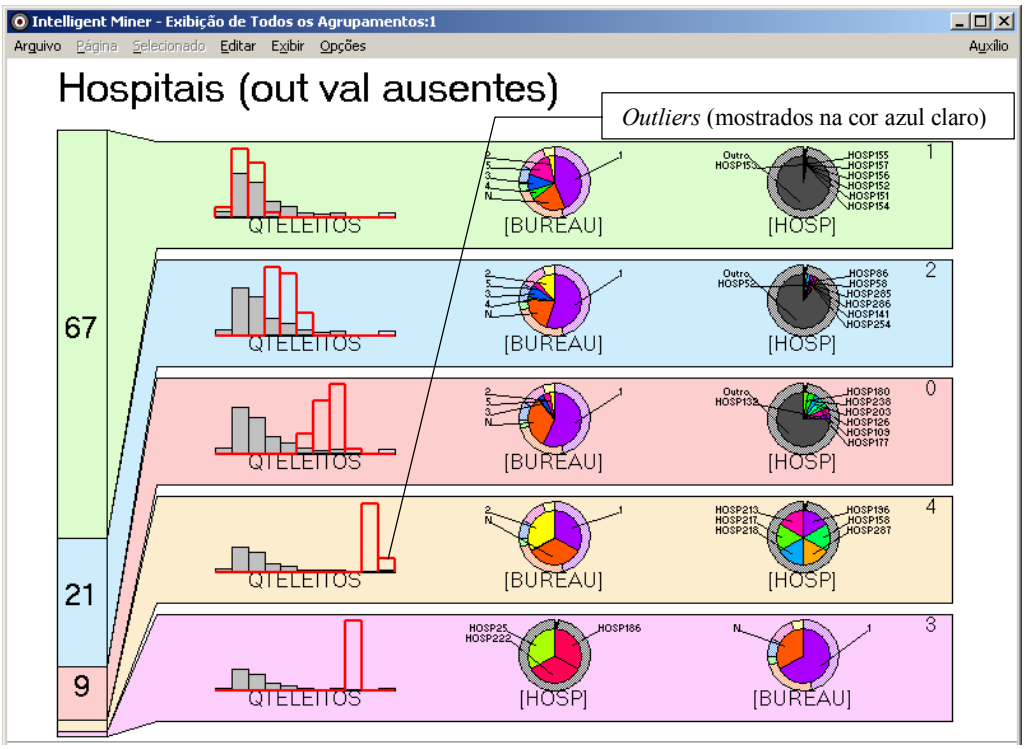


FIGURA 5.3 – Visualização do resultado da Mineração 1.1.

Mediante a análise do relatório dos agrupamentos gerados pela mineração, foi observado que 12 registros foram ignorados e, apesar de aparecerem na visualização, não constam nas estatísticas dos agrupamentos (Figura 5.4).

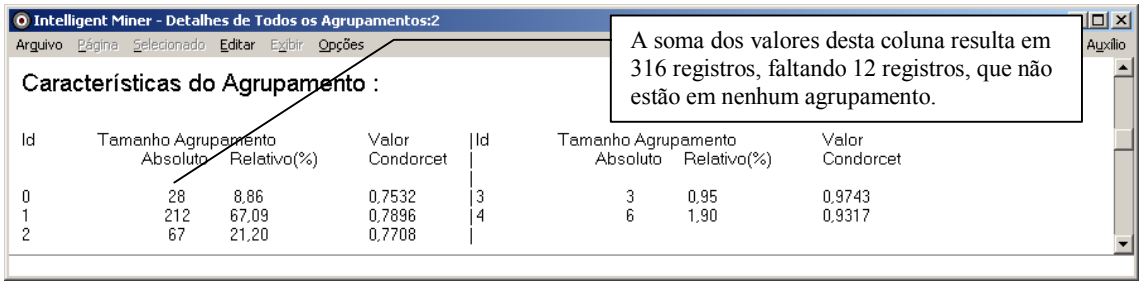


FIGURA 5.4 – Relatório gerado após a Mineração 1.1.

Para descobrir quais os registros ignorados, foi criado um objeto de saída de dados da mineração, que mostrou os 12 registros os quais não receberam valor para o campo ID (identificador do agrupamento), conforme se observa na Figura 5.5.

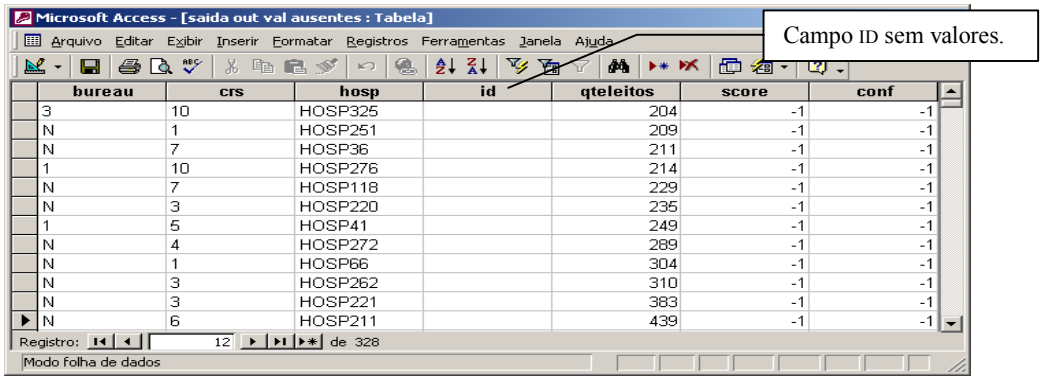


FIGURA 5.5 – Objeto de saída de dados que mostra os 12 registros que não receberam o ID do agrupamento.

Ao se analisar o atributo QTELEITOS, foi verificado que os hospitais que não receberam rótulo de nenhum agrupamento apresentavam um número de leitos alto em relação aos demais. Dessa forma, o algoritmo classificou como *outliers* os valores de 204 a 439 leitos, motivado pela parametrização do item TRATAMENTO DE *OUTLIERS*, cujo valor é *Tratar outliers como valores ausentes*, opção em que esses objetos são desprezados durante a execução de um agrupamento.

A Tabela 5.7 mostra a distribuição dos hospitais nos agrupamentos criados, conforme a quantidade de leitos que possuem. Mostra também o número de hospitais e o valor modal[14] em cada agrupamento. Segundo essa mineração, os hospitais seriam classificados em 5 faixas de portes, porém aqueles com mais de 173 leitos não se encontrariam em nenhuma delas.

TABELA 5.7 – Faixas de portes dos hospitais geradas pela Mineração 1.1.

Agrupamento	Faixa de leitos	No. de hospitais	Valor modal
1	10-64	212	30
2	65-111	67	70
0	112-161	28	150
4	195-200	6	190
3	164-173	3	170

[14] Valor mais freqüente em um intervalo de dados.

Mineração 1.2 – Agrupamento Demográfico com a criação de intervalos inferiores e superiores para acomodar os *outliers*.

Foi executada a Pesquisa de Agrupamento Demográfico do IM, com a opção *Criar buckets inf e sup até que outlier esteja acomodado* para o TRATAMENTO DE *OUTLIERS*, na qual os *buckets*[15] são incluídos em ambas as extremidades do intervalo de valores até que todos os *outliers* estejam contidos em um *bucket*. O tamanho do *bucket* duplica cada vez que um outro *bucket* é incluído.

- **Resultado**

Foram gerados 8 agrupamentos. O valor do Condorcet global foi de 0,792 e o tempo de mineração foi de 0min02. Os registros que foram considerados *outliers* na mineração anterior não mais aparecem como tal nesta mineração. Todos os registros receberam o ID do agrupamento para o qual foram atribuídos. Esta mineração encontrou as faixas de classificação dos hospitais quanto ao número de leitos mostradas na Tabela 5.8.

TABELA 5.8 – Faixas de portes dos hospitais geradas pela Mineração 1.2.

Agrupamento	Faixa de leitos	No. de hospitais	Valor modal
1	10-64	212	30
2	65-111	67	70
0	112-161	28	150
6	195-235	12	220
3	164-173	3	170
5	289-310	3	280
7	383-439	2	379,5
4	249-249	1	280

O agrupamento 4 contém um único hospital, o HOSP41, que possui 249 leitos e valor modal de 280, mesmo valor modal do agrupamento 5. O valor modal (centro do agrupamento) não é considerado em função dos valores mínimos e máximos do agrupamento gerado, e sim, em função dos valores mínimos e máximos do *bucket* em que estão incluídos. Foi possível verificar no relatório estatístico que estes pertencem ao *bucket* que vai de 240 a 320.

Observa-se na Figura 5.6 que o tamanho dos *buckets* duplicou cada vez que um outro *bucket*, contendo os registros que foram considerados *outliers* pelo experimento anterior, foi incluído (até 200, o intervalo era de 20, depois passou para 40, 80 e alcançou o valor máximo 439).

[15] Os valores dos campos numéricos contínuos são atribuídos aos *buckets*. Cada *bucket* representa um intervalo de valores. O número mínimo desse intervalo de valores é o limite inferior do *bucket* que cobre os valores mais baixos, enquanto que o número máximo é o limite superior do *bucket* que cobre os valores mais altos. Os *outliers* são valores menores que o mínimo e maiores que o máximo desse intervalo de valores [IBM 99].

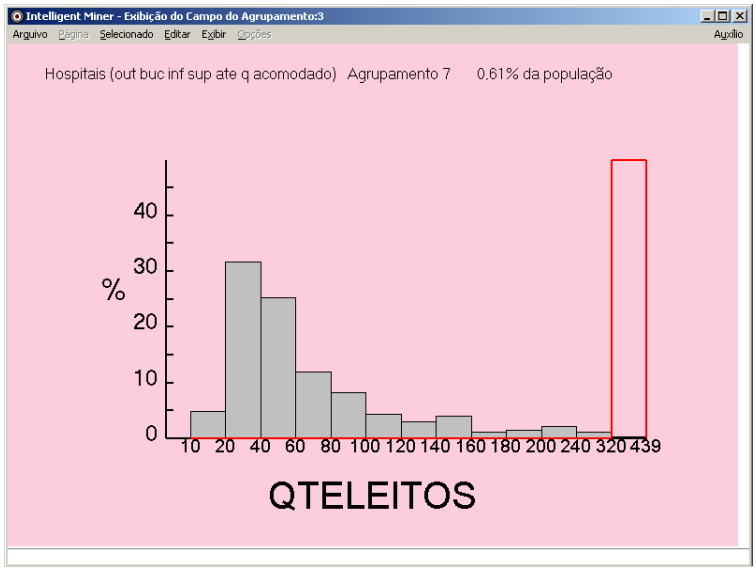


FIGURA 5.6 – Detalhe do atributo QTELEITOS no agrupamento 7 da Mineração 1.2.

Mineração 1.3 – Agrupamento Demográfico com *outliers* colocados em um intervalo inferior ou superior.

Foi realizada a Pesquisa de Agrupamento Demográfico do IM, com a opção *colocar outliers em um bucket inferior e superior* para o TRATAMENTO DE *OUTLIERS*, na qual um único *bucket* pode ser incluído em ambas as extremidades do intervalo de valores a fim de envolver os *outliers*.

- **Resultado**

Foram gerados 7 agrupamentos. O valor do Condorcet global foi de 0,787 e o tempo de mineração foi de 0min02.Os *outliers* são rotulados com o ID do agrupamento nos dados de saída e também são mostrados na visualização. As faixas de classificação quanto ao número de leitos são mostradas na Tabela 5.9.

TABELA 5.9 – Faixas de portes dos hospitais geradas pela Mineração 1.3.

Agrupamento	Faixa de leitos	No. de hospitais	Valor modal
1	10-64	212	30
2	65-111	67	70
0	112-161	28	150
8	211-439	9	319,5
6	195-209	8	190
3	164-173	3	170
5	383-383	1	319,5

Os registros considerados *outliers* pelo algoritmo ficaram nos agrupamentos 8, 6 e 5. Um único *bucket* foi criado com intervalo de 200 a 439, para incluir os *outliers*.

Mineração 1.4 – Agrupamento Demográfico com *outliers* substituídos pelo valor mínimo ou pelo valor máximo.

Foi realizada a Pesquisa de Agrupamento Demográfico do IM, com a opção *Substituir outliers por mín ou máx* para o TRATAMENTO DE *OUTLIERS*, na qual qualquer *outlier* menor que o mínimo será substituído pelo valor mínimo, e qualquer *outlier* maior que o máximo será substituído pelo valor máximo.

- **Resultado**

Foram gerados 4 agrupamentos. O valor do Condorcet global foi de 0,787 e o tempo de mineração foi de 0min02.Os *outliers* são rotulados com o ID do agrupamento nos dados de saída e também são mostrados na visualização. As faixas de classificação quanto ao número de leitos são mostradas na Tabela 5.10.

TABELA 5.10 – Faixas de portes dos hospitais geradas pela Mineração 1.4.

Agrupamento	Faixa de leitos	No. de hospitais	Valor modal
1	10-64	212	30
2	65-111	67	70
0	112-164	29	150
3	172-439	20	319,5

Também neste caso, um único *bucket* foi criado com intervalo de 200 a 439, para incluir os *outliers*.

Mineração 1.5 – Agrupamento Neural com *outliers* posicionados em um intervalo intermediário.

Foi realizada a Pesquisa de Agrupamento Neural do IM, com a opção *Tratar outliers como valores ausentes* para o TRATAMENTO DE *OUTLIERS*. Neste caso, os *outliers* não são desprezados, mas sim mapeados para um valor representado em escala por 0,5 da unidade de entrada neural correspondente.

- **Resultado**

Foram gerados 9 agrupamentos. O desvio de agrupamento final foi de 0,0005606 e o tempo de mineração foi de 0min02. Os *outliers* são rotulados com o ID do agrupamento nos dados de saída e também são mostrados na visualização. As faixas de classificação quanto ao número de leitos são mostradas na Tabela 5.11.

Observa-se que a faixa de leitos do agrupamentos 0 está dentro do intervalo da faixa de leitos do agrupamento 1. Ao examinar o objeto de saída de dados, verificou-se que o agrupamento 1 apresenta valores que vão de 67 a 82 e 204 a 439. Portanto, os *outliers* se encontram nesse agrupamento, mas a distribuição parece confusa.

TABELA 5.11 – Faixas de portes dos hospitais geradas pela Mineração 1.5.

Agrupamento	Faixa de leitos	No. de hospitais	Valor modal
7	29-37	57	30
8	10-28	53	30
1	67-439	44	70
6	45-51	43	50
3	84-115	34	90
0	120-200	33	150
5	38-43	31	50
4	52-57	17	50
2	59-65	16	70

Mineração 1.6 – Agrupamento Neural com *outliers* substituídos pelo valor mínimo ou pelo valor máximo.

Foi realizada a Pesquisa de Agrupamento Neural do IM, com a opção *Substituir outliers por Min ou Max* para o TRATAMENTO DE *OUTLIERS*, na qual todo *outlier* se menor que o mínimo será substituído pelo valor mínimo e se maior que o máximo será substituído pelo valor máximo.

- **Resultado**

Foram gerados 9 agrupamentos. O desvio de agrupamento final foi de 0,0006369 e o tempo de mineração foi de 0min02. Os *outliers* são rotulados com o ID do agrupamento nos dados de saída e também são mostrados na visualização. As faixas de classificação quanto ao número de leitos são mostradas na Tabela 5.12.

TABELA 5.12 – Faixas de portes dos hospitais geradas pela Mineração 1.6.

Agrupamento	Faixa de leitos	No. de hospitais	Valor modal
8	10-30	68	30
5	31-39	52	30
0	138-439	36	150
1	73-95	36	90
7	40-46	36	50
2	47-53	31	50
3	97-136	30	110
6	61-70	21	70
4	54-60	18	50

Aqui, os *outliers* ficaram no agrupamento 0 e não houve intervalos sobrepostos como no caso anterior.

Mineração 1.7 – Agrupamento Neural com tratamento de *outliers* como valores válidos.

Foi realizada a Pesquisa de Agrupamento Neural do IM, com a opção *Tratar outliers como valores válidos* para o TRATAMENTO DE *OUTLIERS*, na qual os *outliers* são tratados como valores normais.

- **Resultado**

Foram gerados 9 agrupamentos. O desvio de agrupamento final foi de 0,002049 e o tempo de mineração foi de 0min01. Os *outliers* são rotulados com o ID do agrupamento nos dados de saída e também são mostrados na visualização. As faixas de classificação quanto ao número de leitos são mostradas na Tabela 5.13.

TABELA 5.13 – Faixas de portes dos hospitais geradas pela Mineração 1.7.

Agrupamento	Faixa de leitos	No. de hospitais	Valor modal
3	69-125	68	90
5	29-36	56	30
8	10-28	53	30
0	129-439	40	150
2	43-48	33	50
7	37-42	27	50
4	49-53	18	50
6	54-60	18	50
1	61-68	15	70

Também neste caso os *outliers* ficaram no agrupamento 0 e não houve intervalos sobrepostos.

5.4.3 Conclusões e validação do Experimento 1

Apesar de o atributo QTELEITOS ser um atributo numérico, dentre os modelos de mineração gerados, o modelo da Mineração 1.4, que utiliza a Pesquisa de Agrupamento Demográfico, e resultou em hospitais distribuídos em 4 faixas conforme a quantidade de leitos, atende melhor o objetivo deste experimento, no que se refere a simplesmente agrupar os hospitais em faixas de leitos que representem o seu porte. Seus resultados para hospitais com aproximadamente 50 leitos se aproximam da classificação de hospitais pequenos que a SES utiliza.

O agrupamento demográfico apresentou, mediante as diversas opções de tratamento de *outliers*, intervalos mais regulares, com quantidades de registros bem variadas em cada faixa de leitos. O valor do Condorcet global, em média, foi de 0,788. Em agrupamentos perfeitos esse valor seria igual a 1.

O agrupamento neural apresentou intervalos mais variáveis e quantidades de registros mais regulares em cada faixa de leitos. Foi feita uma última mineração com o neural com 4 agrupamentos máximos e as faixas encontradas foram de 10 a 37, 38 a 51, 52 a 65 e 67 a 439, intervalos bem diferentes dos encontrados com o algoritmo demográfico. O desvio de agrupamento final de menor valor foi 0,0005606. Em agrupamentos perfeitos esse valor seria igual a 0.

Os modelos gerados no Experimento 1 demonstram como é possível observar desvios em atributos numéricos com as diversas opções de tratamento de *outliers*. A escolha de qual opção deve ser utilizada depende muito do objetivo e do conhecimento que se tem sobre os dados.

No caso deste experimento, por exemplo, o objetivo não é descartar hospitais cuja quantidade de leitos é considerada elevada em comparação com os demais. No entanto, em um experimento semelhante com o atributo VALTOTAL (valor total de uma AIH), foram geradas faixas de valores destoantes da realidade. Isso ocorreu porque o valor máximo desse atributo (R$ 112.926,33) não era um valor válido, decorreu de um erro ao ser lançado no sistema, e deveria ter sido eliminado ou corrigido na etapa de limpeza dos dados. Porém, somente com o conhecimento dos especialistas da saúde é que se chegou a essa conclusão. Nessa situação, tratar *outliers* como valores ausentes seria a opção mais adequada.

O Experimento 1, ao ser levado para a validação pelos especialistas da SES, foi facilmente entendido e seus resultados foram considerados válidos. Apesar de o atributo PORTE não ser ainda utilizado nas análises de auditoria, cujo sistema de implantação automatizado ainda se encontra em andamento, e apesar da quantidade de leitos não ser o único atributo usado para definir o porte de um hospital, o resultado da Mineração 1.4 se aproxima da realidade e pode ser utilizado provisoriamente para a obtenção de perfis dos hospitais.

Os diversos resultados obtidos, no entanto, podem ser úteis para outros tipos de análises. Como já foi citado neste trabalho, a validade do conhecimento descoberto é avaliada conforme os objetivos da aplicação.

5.4.4 Experimento 2 – Construção de modelos de mineração de dados com a utilização de agrupamento sobre os dados da saúde.

Na fase de compreensão do domínio da aplicação, colocou-se como situação de interesse para esta aplicação, a análise das tendências de hospitais no que se refere às internações hospitalares realizadas nos meses de maio a dezembro de 2000, bem como a análise do perfil de internações bloqueadas, nas quais se concentram as situações de impropriedades ocorridas no sistema.

No decorrer deste trabalho, muitos experimentos foram realizados nesse sentido. Diversos modelos foram criados e foram identificados desvios, como por exemplo, um agrupamento com 5 internações cujos valores iam de R$105.916,90 a R$112.926,33, para tratamento oncológico e tratamento psiquiátrico. Neste caso, foi constatado que houve um engano no lançamento de valores. Segundo os auditores, no ano 2000, alguns casos, ainda não detectados pela SES, eram bloqueados mais adiante, na auditoria do DATASUS. Assim, a sugestão dos especialistas foi a de que se descartasse esses registros. Apesar de a ferramenta identificar esses valores como *outliers*, somente com o conhecimento do domínio foi possível afirmar que se tratava de valores incorretos.

Após diversas experiências infrutíferas com a finalidade de que os desvios assim analisados resultariam em padrões interessantes para o objetivo proposto, passou-se a utilizar uma outra metodologia, a de analisar as tendências das internações hospitalares realizadas e das internações bloqueadas no decorrer do tempo.

Estudos sobre detecção de fraudes mostram que uma solução muito usada é a análise do perfil dos objetos, e, no decorrer do tempo, cada nova transação é comparada com esse perfil para verificar se ocorrem desvios em relação ao comportamento médio

desses objetos [MOR 96, MOR 97, YAM 2000]. Partiu-se, então, do princípio de que uma análise semelhante poderia ser válida para as internações hospitalares, com a finalidade de avaliar mês a mês se ocorrem alterações significativas no sistema.

Nesse sistema específico, as ações voltadas para a captura de situações de desvios, aqui entendidos como as situações de impropriedades nas internações hospitalares, se dão por intermédio da utilização de critérios de bloqueio técnico estabelecidos com base na experiência da auditoria médica, bem como a utilização de normas do SUS. A situação a ser avaliada é a de que esses critérios precisam ser revistos, para que sejam aperfeiçoados ou para que novos critérios sejam criados.

Este trabalho pretende mostrar como a análise de agrupamentos pode ser bastante útil para melhorar e apontar novos critérios de bloqueios técnicos de AIHs.

Portanto, os objetivos deste experimento são:

1) Utilizar a Pesquisa de Agrupamento Demográfico, a qual é mais indicada para dados categóricos, para a análise dos padrões mais freqüentes de comportamento nos dados da saúde.

2) Examinar os padrões dos registros bloqueados, nos quais estão inseridos os casos de irregularidades encontrados pelos especialistas nos conjuntos de dados estudados e verificar qual o porte dos hospitais mais problemáticos.

3) Verificar os padrões mais freqüentes das internações realizadas de hospitais com o mesmo porte dos hospitais mais problemáticos.

4) Verificar os padrões mais freqüentes das internações bloqueadas por tipo de problema apresentado.

5) Extrair e avaliar situações de possíveis desvios com base nos padrões encontrados nas internações hospitalares bloqueadas.

6) Examinar os padrões mais freqüentes dos hospitais mais problemáticos, por mês, para avaliar as alterações que ocorrem nesses dados.

Os parâmetros de modo do algoritmo demográfico do IM, para os modelos de mineração que serão gerados, permanecerão com os valores padrões do IM, ou seja, 2 passagens máximas sobre o BD, 9 agrupamentos máximos e melhora na precisão entre duas passagens de 2%.

Para o tratamento de *outlier*, nos casos em que forem utilizados atributos numéricos, será escolhida a opção *Substituir outliers por mín ou máx*, com base nos resultados do experimento anterior. Quanto aos *outliers* em atributos categóricos, estes serão observados pela visualização das características de um determinado agrupamento.

Deve ser observado que cada linha, nas tabelas que serão apresentadas nos experimentos, contém os valores dos atributos do protótipo de um agrupamento, que são os valores mais freqüentes para os atributos nesse agrupamento, não significando que sejam os valores de um determinado registro do agrupamento. Por exemplo, se em

um agrupamento, o valor mais freqüente para o atributo procedimento realizado (PROC_REA) for 31000002 (cirurgia múltipla), e se o valor mais freqüente para o atributo novo código de procedimento (NOVO_COD) for 77500130 (hipertensão maligna), não significa que cirurgia múltipla recebeu o código 77500130 como o novo valor de procedimento, uma vez que são procedimentos de grupos distintos, mas apenas que esse valor foi o mais freqüente de todos os valores para esse atributo nesse agrupamento.

Mineração 2.1 – Análise das AIHs bloqueadas no período de maio a dezembro/2000.

O ponto de partida para este experimento foi a análise dos registros bloqueados. Este modelo deve revelar os hospitais com maior número de internações problemáticas.

Para a criação deste modelo de mineração foi utilizado o subconjunto de 14.314 internações bloqueadas no período estudado, extraído do conjunto de dados C2. As internações bloqueadas estão subdivididas nas seguintes categoriais por tipo de problema:

–	Liberadas com código novo:	2.257 internações	15,77%
–	Liberadas com mesmo código:	9.517 internações	66,49%
–	Pemanecem bloqueadas:	1.666 internações	11,64%
–	Sem resposta do auditor:	874 internações	6,11%

Conforme foi constatado, a análise mês a mês é bastante significativa. Assim, o mês de apresentação da AIH (APRES) é um atributo bastante considerado nesta pesquisa.

Os campos ativos foram APRES e HOSP e os campos suplementares foram ESPEC, FAIXADIAS, CRS, VALTOTAL, CUSTO_AIH, DIAG_PRI, GRUPOCID, PROC_REA e PORTE. Foram tentadas outras configurações com o uso de mais campos ativos, entretanto, o valor do Condorcet fica muito baixo, o que significa agrupamentos de baixa qualidade.

- **Resultado**

Os padrões mais freqüentes encontrados são mostrados na Tabela 5.14. Os valores dos atributos em cada linha da tabela são os centros dos agrupamentos criados. Esses valores foram extraídos do gráfico de visualização dos resultados (Figura A do Anexo) e dos relatórios gerados pela mineração e são dados pelos valores modais (valores mais freqüentes) dos atributos em cada agrupamento. Os agrupamentos foram ordenados pelo campo APRES para se avaliar os padrões no decorrer do tempo.

Observa-se que os padrões encontrados apresentam um comportamento variável de um mês para outro, conforme demonstram os valores da maior parte dos atributos. Porém, ocorre a predominância de hospitais que se classificam como PORTE 4 (172 a 439 leitos), de acordo com o resultado da mineração 1.4 do Experimento 1. Portanto, as próximas minerações serão concentradas na análise específica dos hospitais PORTE 4, a fim de evidenciar um padrão médio de comportamento.

Observa-se ainda o agrupamento 2. Em relação aos outros, ele pode ser visto como um *outlier*, uma vez que foi formado somente por internações do HOSP326. Essas internações possuem os mesmos padrões mais freqüentes em todos os meses e os valores de seus atributos CRS, PORTE e HOSP tiveram maior influência para a formação do agrupamento do que os valores do campo ativo APRES (vide Figura A do Anexo).

TABELA 5.14 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas.

AGRUP	APRES	HOSP	ESPEC	FAIXA DIAS	CRS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPO CID	PROC_ REA	PORTE
8 (8,25%)	62000	HOSP66	03	4 a 6	1	300,00	B	J449	GP10	76500225	4
6 (9,21%)	72000	HOSP66	03	1 a 3	1	300,00	B	J449	GP10	76500225	4
5 (21,04%)	82000	HOSP66	03	4 a 6	1	300,00	B	G458	GP06	81500106	4
7 (13,13%)	92000	HOSP211	03	1 a 3	1	300,00	B	G458	GP06	81500106	4
0 (10,68%)	102000	HOSP66	03	4 a 6	1	300,00	B	G458	GP01	81500106	4
2 (1,18%)	102000	HOSP326	03	1 a 3	13	300,00	B	A419	GP01	74500244	3
4 (11,67%)	112000	HOSP211	03	4 a 6	1	300,00	B	G458	GP06	81500106	4
3 (14,13%)	122000	HOSP272	03	1 a 3	1	300,00	B	O829	GP15	35009012	4
1 (10,70%)	12001	HOSP211	03	1 a 3	1	300,00	B	G458	GP06	81500106	4

Legenda:
ESPEC: 03 – especialidade médica em clínica médica.
DIAG_PRI: J449 – Doença pulmonar obstrutiva crônica neonatal; G458 – Outros acidentes isquêmicos cerebrais trans sindr corr; A419 – Septicemia neonatal; O829 – Parto p/cesariana NE.
GRUPOCID: GP10 - Doenças do aparelho respiratório; GP06 – Doenças do sistema nervoso; GP01 – Algumas doenças infecciosas e parasitárias; GP15 – Gravidez , parto e puerpério.
PROC_REA: 76500225 – Doenca pulmonar obstrutiva crônica; 81500106 – AVC agudo; 74500244 – Septicemia (clínica médica); 35009012 – Cesariana.

O valor do Condorcet global para os agrupamentos gerados foi de 0,5101. Apesar deste valor não ter sido muito elevado, o algoritmo agrupou os registros corretamente, a maior parte por mês de apresentação e os registros do agrupamento 2 por CRS. O tempo de execução da mineração foi de 1min27. Este tempo sofre pequenas variações conforme o hardware em que foi executado o algoritmo ou conforme os atributos selecionados, segundo foi observado na geração dos modelos.

Mineração 2.2 – Análise das AIHs de hospitais PORTE 4 no período de maio a dezembro/2000.

Para a criação deste modelo de mineração, foi utilizado o subconjunto de 95.991 internações realizadas por hospitais PORTE 4 no período estudado. Este subconjunto foi extraído do conjunto de dados C2.

Os campos ativos foram APRES e HOSP e os campos suplementares foram ESPEC, FAIXADIAS, CRS, VALTOTAL, CUSTO_AIH, DIAG_PRI, GRUPOCID, PROC_REA.

- **Resultado**

Os padrões encontrados são bem regulares no decorrer do período estudado, conforme mostra a Tabela 5.15. As internações de hospitais PORTE 4 apresentam as seguintes características: o hospital HOSP211 é o mais freqüente em todos os meses. Predominam os valores de especialidade 03 (clínica médica), o período de internação de 1 a 3 dias, valor total de internação de R$300,00, o procedimento realizado 35001011

(parto normal), e o diagnóstico principal O809 (parto normal) do grupo GP15 (gravidez, parto e puerpério).

TABELA 5.15 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs de hospitais PORTE 4.

AGRUP	APRES	HOSP	ESPEC	FAIXA DIAS	CRS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPO CID	PROC_ REA
2 (12,68%)	62000	HOSP211	03	1 a 3	3	300,00	B	O809	GP15	35021012
8 (11,87%)	72000	HOSP211	03	1 a 3	3	300,00	B	O809	GP15	35021012
3 (12,48%)	82000	HOSP66	01	1 a 3	1	300,00	B	O809	GP15	35001011
4 (12,17%)	82000	HOSP211	03	1 a 3	3	300,00	B	O809	GP15	35001011
0 (9,80%)	92000	HOSP211	03	1 a 3	6	300,00	B	O809	GP15	35001011
5 (10,03%)	102000	HOSP211	03	1 a 3	10	300,00	B	O809	GP15	35001011
7 (10,33%)	112000	HOSP211	03	1 a 3	6	300,00	B	O809	GP15	35001011
6 (10,52%)	122000	HOSP211	03	1 a 3	6	300,00	B	O809	GP15	35001011
1 (10,14%)	12001	HOSP211	03	1 a 3	6	300,00	B	O809	GP15	35001011

Legenda:
ESPEC: 03 – especialidade médica em clínica médica; 01 – especialidade médica em cirurgia geral.
DIAG_PRI: O809 – Parto único espontâneo NE.
GRUPOCID: GP15 – Gravidez , parto e puerpério.
PROC_REA: 35021012 – Parto normal c/ atendimento RN sala de parto; 35001011– Parto normal.

O agrupamento 3 possui um tamanho significativo. É o segundo maior agrupamento e é formado somente por internações do HOSP66 realizadas nos diversos meses do ano, conforme se visualiza na Figura B do Anexo. Diferencia-se dos outros padrões apenas pelo valor 01 do atributo especialidade médica (ESPEC), que se refere a cirurgia geral. Significa que as internações realizadas por esse hospital possuem os mesmos padrões mais freqüentes em todos os meses. Este agrupamento pode ser visto como um *outlier*, tendo em vista que esse hospital se sobressai em relação aos outros.

Este modelo, ao ser validado pelos auditores, chamou a atenção, uma vez que no ano 2001, foi estabelecido um novo critério técnico que bloqueia internações com baixa permanência, ou seja, com 48 horas ou menos de duração, ante a observação da manipulação de códigos de doenças simples para doenças mais graves. Os auditores visualizaram este fato rapidamente, com a análise dos valores do atributo FAIXADIAS no gráfico de visualização dos agrupamentos gerados (Figura B do Anexo), mas comentaram que esse critério foi resultado de observações de levantamentos estatísticos trabalhosos.

O valor do Condorcet para os agrupamentos gerados foi de 0,5408 e o tempo de execução foi de 2min40.

Apesar desta análise se concentrar em internações de hospitais PORTE 4, foram criados modelos para os outros portes de hospital e também um modelo de todas as internações realizadas, e verificou-se que os padrões mais freqüentes são bem regulares em todos os modelos e que existe uma pequena diferença entre os padrões mais freqüentes para os diversos portes, o que valida a classificação destes hospitais segundo os resultados do Experimento 1.

O próximo modelo de mineração analisa as internações bloqueadas apenas de hospitais PORTE 4.

Mineração 2.3 – Análise das AIHs bloqueadas de hospitais PORTE 4 no período de maio a dezembro/2000.

Para a criação deste modelo de mineração, foi utilizado o subconjunto de 5.209 internações bloqueadas no período estudado. Este subconjunto foi extraído do conjunto de dados C2. As internações bloqueadas de hospitais PORTE 4 estão subdivididas em:

- Liberadas com código novo: 429 internações 8,24%
- Liberadas com mesmo código: 3.976 internações 76,33%
- Pemanecem bloqueadas: 391 internações 7,51%
- Sem resposta do auditor: 413 internações 7,93%

Os campos ativos foram APRES e HOSP e os campos suplementares foram ESPEC, FAIXADIAS, CRS, VALTOTAL, CUSTO_AIH, DIAG_PRI, GRUPOCID, PROC_REA.

- **Resultado**

Esta mineração mostra os padrões mensais das internações bloqueadas dos hospitais PORTE 4. Os centros dos agrupamentos gerados revelam um comportamento irregular no período estudado, conforme mostra a Tabela 5.16.

TABELA 5.16 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas de hospitais PORTE 4.

AGRUP	APRES	HOSP	ESPEC	FAIXA DIAS	CRS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPO CID	PROC_ REA
4 (8,16%)	72000	HOSP211	01	1 a 3	6	375,00	B	I509	GP09	31000002
0(16,78%)	82000	HOSP211	03	4 a 6	3	375,00	B	G458	GP06	81500106
5 (8,50%)	92000	HOSP211	01	1 a 3	6	375,00	M	A419	GP01	81500106
6(19,70%)	102000	HOSP66	01	1 a 3	1	375,00	B	G458	GP06	81500106
3(10,56%)	102000	HOSP251	03	1 a 3	1	375,00	B	A419	GP11	31000002
7 (7,39%)	112000	HOSP211	03	4 a 6	6	375,00	B	A419	GP01	81500106
1 (10,04%)	122000	HOSP272	03	1 a 3	4	375,00	B	O829	GP15	35009012
2 (9,23%)	122000	HOSP211	03	1 a 3	6	375,00	B	O829	GP15	35009012
8 (9,64%)	12001	HOSP211	01	1 a 3	6	375,00	B	O829	GP15	35009012

Legenda:
ESPEC: 03 – especialidade médica em clínica médica; 01 – especialidade médica em cirurgia geral.
DIAG_PRI: I509 – Insuficiência cardíaca NE; G458 – Outros acidentes isquêmicos cerebrais trans sindr corr; A419 – Septicemia neonatal; O829 – Parto p/cesariana NE.
GRUPOCID: GP09 – Doenças do aparelho circulatório; GP06 – Doenças do sistema nervoso; GP01 – Algumas doenças infecciosas e parasitárias; GP11 – Doenças do aparelho digestivo; GP15 – Gravidez , parto e puerpério.
PROC_REA: 31000002– Cirurgia múltipla; 81500106 – AVC agudo; 35009012 – Cesariana.

Nos últimos meses, houve uma tendência para os diagnósticos do grupo CID GP15 (gravidez, parto e puerpério), com procedimento realizado mais comum de 35009012 (cesariana).

A validação com os auditores esclareceu que este resultado está de acordo com uma determinação federal que determinou os bloqueios pela SES, a partir de novembro/2000, das internações por cesariana que excedessem a taxa de 30% de cesáreas por hospital, estabelecida pelo SUS. Então, para se ter uma idéia mais clara dos padrões mais freqüentes desses agrupamentos, os registros bloqueados pelo motivo de

ajuste de 30% da taxa de cesárea foram retirados do conjunto de dados que está sendo avaliado. Um novo modelo foi criado sem esses registros e o atributo código do motivo de bloqueio foi acrescentado. O número de internações passou para 4.820. O resultado é mostrado na Tabela 5.17 e pode ser visualizado na Figura C do Anexo.

TABELA 5.17 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas de hospitais PORTE 4, sem os bloqueios por cesariana.

AGRUP	APRES	HOSP	ESPEC	FAIXA DIAS	CRS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPO CID	PROC_ REA
8 (12,59%)	62000	HOSP211	01	4 a 6	6	375,00	B	A419	GP09	31000002
4(8,82%)	72000	HOSP211	01	1 a 3	6	375,00	B	I509	GP09	31000002
0(18,13%)	82000	HOSP211	03	4 a 6	3	375,00	B	G458	GP06	81500106
1 (8,07%)	82000	HOSP272	03	1 a 3	4	625,00	M	A419	GP02	74500244
5(9,19%)	92000	HOSP211	01	1 a 3	6	375,00	M	A419	GP01	81500106
6 (21,29%)	102000	HOSP66	01	1 a 3	1	375,00	B	G458	GP06	81500106
3 (6,45%)	102000	HOSP211	03	1 a 3	6	625,00	M	A419	GP01	74300261
7 (7,99%)	112000	HOSP211	03	4 a 6	6	375,00	B	A419	GP01	81500106
2 (7,47%)	122000	HOSP211	03	4 a 6	6	375,00	B	A419	GP01	81500106

Legenda:
ESPEC: 03 – especialidade médica em clínica médica; 01 – especialidade médica em cirurgia geral.
DIAG_PRI: G458 – Outros acidentes isquêmicos cerebrais trans sindr corr; A419 – Septicemia neonatal; I509 – Insuficiência cardíaca NE.
GRUPOCID: GP06 – Doenças do sistema nervoso; GP09 – Doenças do aparelho circulatório; GP01 – Algumas doenças infecciosas e parasitárias; GP02 – Neoplasias (tumores).
PROC_REA: 81500106 – AVC agudo; 31000002– Cirurgia múltipla; 74500244 – Septicemia (clínica médica); 74300261 – Septicemia (pediatria).

Observa-se que os padrões mais freqüentes de hospitais PORTE 4 são bem variados, predominam internações do HOSP211 e os custos são um pouco mais elevados, se comparados com os custos de todas as AIHs bloqueadas (Tabela 5.14).

Os agrupamentos 6 e 1 podem ser vistos como *outliers*, uma vez que são constituídos por internações, todas do HOSP66 e HOSP272, respectivamente, e o primeiro forma o maior agrupamento de internações bloqueadas de hospitais PORTE 4 (Figura C do Anexo). As internações apresentadas em janeiro/2001 (12001), referentes às AIHs de dezembro/2000, não tiveram valores modais significativos o bastante para formar um agrupamento e ficaram diluídas nos agrupamentos das outras apresentações, que possuem padrões semelhantes.

O valor do Condorcet para os agrupamentos gerados foi de 0,5318 e o tempo de execução foi de 2min42.

Os próximos modelos de mineração serão criados para cada tipo de problema de AIHs bloqueadas.

Mineração 2.4 – Análise das AIHs bloqueadas e liberadas com código novo de hospitais PORTE 4 no período de maio a dezembro/2000.

Para a criação deste modelo de mineração, foi utilizado o subconjunto de 429 internações bloqueadas e, posteriormente, liberadas com um outro código de procedimento, uma vez que a auditoria discordou do procedimento cobrado inicialmente. Este subconjunto foi extraído do conjunto de dados C2.

Os campos ativos foram APRES e HOSP e os campos suplementares foram ESPEC, FAIXADIAS, CRS, VALTOTAL, CUSTO_AIH, DIAG_PRI, GRUPOCID, PROC_REA, NOVO_COD e CODMOTIVO.

- **Resultado**

O resultado é mostrado na Tabela 5.18 e pode ser visualizado na Figura D do Anexo. Os padrões mais freqüentes encontrados são bastante variáveis, significando que as impropriedades sofrem alterações mensais. Os diagnósticos principais mais comuns são aqueles do grupo GP06 (doenças do sistema nervoso), que receberam o novo código de procedimento 77500121 (crise hipertensiva), em lugar do procedimento 81500106 (acidente vascular cerebral agudo), o período de internação de 1 a 3 dias mudou nos últimos meses para 4 a 6 dias, e custo total de internação é baixo (B: entre R$200,00 e R$500,00).

O critério AVC agudo foi implantado na apresentação de agosto de 2000. Este fato justifica as estatísticas do agrupamento 7, que possui o maior tamanho devido ao número de impropriedades capturadas por esse motivo. Observa-se que esse ainda apareceu como o procedimento realizado mais freqüente nos meses finais do ano, mas em hospitais diferentes.

TABELA 5.18 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas e liberadas com código novo de hospitais porte 4.

AGRUP	APRES	HOSP	ESPEC	FAIXA DIAS	CRS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPO CID	PROC_ REA	NOVO_ COD	COD MOTI VO
1 (3,96%)	62000	HOSP66	01	1 a 3	1	300,00	M	T009	GP19	31000002	77500130	07
8 (9,09%)	72000	HOSP41	03	1 a 3	5	500,00	B	A419	GP10	31000002	72500000	17
7 (24,94%)	82000	HOSP41	03	1 a 3	10	300,00	B	G458	GP06	81500106	77500121	06
3 (12,35%)	92000	HOSP287	03	1 a 3	7	300,00	B	G458	GP06	81500106	77500121	06
0 (10,49%)	102000	HOSP158	03	1 a 3	17	500,00	M	A419	GP01	74300261	77500121	01
5 (9,79%)	102000	HOSP251	01	1 a 3	1	300,00	B	I679	GP11	31000002	34008020	04
6 (11,89%)	112000	HOSP158	03	4 a 6	17	300,00	B	A419	GP06	81500106	77500121	06
4 (11,19%)	122000	HOSP276	03	4 a 6	10	300,00	B	A419	GP06	81500106	77500121	06
2 (6,29%)	12001	HOSP213	03	4 a 6	6	500,00	B	G458	GP06	81500106	77500121	06

Legenda:

ESPEC: 03 – especialidade médica em clínica médica; 01 – especialidade médica em cirurgia geral.

DIAG_PRI: T009 – Traumatismo superf mult NE; A419 – Septicemia neonatal; G458 – Outros acidentes isquêmicos cerebrais trans sindr corr; I679 – Doença cerebrovascular NE.

GRUPOCID: GP19 – Lesões, enven e algumas outr conseq de causas externas; GP10 - Doenças do aparelho respiratório; GP06 – Doenças do sistema nervoso; GP01 – Algumas doenças infecciosas e parasitárias; GP11 – Doenças do aparelho digestivo.

PROC_REA: 31000002– Cirurgia múltipla; 81500106 – AVC agudo; 74300261– Septicemia (pediatria).

NOVO_COD: 77500130– Hipertensão maligna; 72500000– Diagnóstico e/ou primeiro atendimento em clínica médica; 77500121– Crise hipertensiva; 34008020 – Colpoperineoplastia anterior e posterior.

CODMOTIVO: 07 – Homônimos; 17 – Homônimos e Politraumatizados; 06 – AVC agudo; 01 – Septicemia; 04 – Cirurgias múltiplas.

O agrupamento 5 foi constituído por internações, todas do HOSP251, realizadas no período estudado e, portanto, é considerado um *outlier*.

O valor do Condorcet para os agrupamentos gerados foi de 0,5822 e o tempo de execução foi de 1min18.

Mineração 2.5 – Análise das AIHs bloqueadas e liberadas com mesmo código de hospitais PORTE 4 no período de maio a dezembro/2000.

Para a criação deste modelo de mineração, foi utilizado o subconjunto de 3.587 internações bloqueadas e, posteriormente, liberadas com o mesmo código, em virtude da auditoria não ter conseguido identificar problemas nesse tipo de internação, concluindo que são cobranças adequadas. Este subconjunto foi extraído do conjunto de dados C2. Foram excluídos os registros referentes a bloqueios por ajuste de taxa de cesárea.

Os campos ativos foram APRES e HOSP e os campos suplementares foram ESPEC, FAIXADIAS, CRS, VALTOTAL, CUSTO_AIH, DIAG_PRI, GRUPOCID, PROC_REA e CODMOTIVO.

- **Resultado**

O resultado é mostrado na Tabela 5.19 e pode ser visualizado na Figura E do Anexo. Os padrões encontrados variam mês a mês, mas o motivo de bloqueio que mais ocorreu foi por homônimos.

TABELA 5.19 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas e liberadas com mesmo código de hospitais PORTE 4.

AGRUP	APRES	HOSP	ESPEC	FAIXA DIAS	CRS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPO CID	PROC_ REA	COD MOTIVO
8 (9,65%)	62000	HOSP211	01	4 a 6	6	375,00	B	A419	GP09	31000002	07
3 (7,33%)	72000	HOSP211	01	1 a 3	6	375,00	B	I509	GP09	38018012	17
0 (13,21%)	82000	HOSP211	03	4 a 6	6	375,00	B	G458	GP06	81500106	17
1 (10,65%)	82000	HOSP272	03	1 a 3	4	625,00	M	A419	GP02	74500244	07
7 (9,45%)	92000	HOSP211	01	4 a 6	6	375,00	M	A419	GP01	74500244	07
4 (23,53%)	102000	HOSP66	01	1 a 3	1	375,00	B	G458	GP19	81500106	07
5 (9,00%)	102000	HOSP41	03	1 a 3	6	375,00	M	A419	GP01	74300261	07
6 (8,67%)	112000	HOSP211	01	4 a 6	6	375,00	B	A419	GP09	81500106	07
2 (8,50%)	122000	HOSP211	03	4 a 6	6	375,00	B	A419	GP01	74500244	07

Legenda:
ESPEC: 03 – especialidade médica em clínica médica; 01 – especialidade médica em cirurgia geral.
DIAG_PRI: A419 – Septicemia neonatal; I509 – Insuficiência cardíaca NE; G458 – Outros acidentes isquêmicos cerebrais trans sindr corr.
GRUPOCID: GP09 – Doenças do aparelho circulatório; GP06 – Doenças do sistema nervoso; GP02 – Neoplasias (tumores); GP01 – Algumas doenças infecciosas e parasitárias; GP19 – Lesões, enven e algumas outr conseq de causas externas.
PROC_REA: 31000002– Cirurgia múltipla; 38018012 – Debridamento da fasceite necrotizante; 81500106 – AVC agudo; 74500244 – Septicemia (clínica médica); 74300261- Septicemia (pediatria).
CODMOTIVO: 07 – Homônimos; 17 – Homônimos e Politraumatizados.

Observa-se agora o maior agrupamento, o agrupamento 4, com internações, todas do HOSP66 (vide Figura E do Anexo). Há também o agrupamento 1, com internações somente do HOSP272, os quais são considerados *outliers*. As internações apresentadas em dezembro (12001) não tiveram valores modais significativos o bastante para formar um agrupamento e ficaram diluídas nos agrupamentos das outras apresentações, com padrões semelhantes. Predominam casos de AVC agudo e septicemia, mas bloqueados pelo critério de homônimos.

Comparando os dois últimos modelos, verifica-se que as AIHs liberadas com código novo são bloqueadas, na maioria das vezes, pelo motivo AVC agudo, enquanto que as liberadas com o mesmo código são bloqueadas pelo motivo homônimos. Foi

sugerido à auditoria avaliar a possibilidade de melhorar as regras de homônimos, o que diminuirá bastante esse tipo de problema nos bloqueios.

O valor do Condorcet para os agrupamentos gerados foi de 0,5462 e o tempo de execução foi de 3min06.

Mineração 2.6 – Análise das AIHs que permanecem bloqueadas de hospitais PORTE 4 no período de maio a dezembro/2000.

Para a criação deste modelo de mineração, foi utilizado o subconjunto de 391 internações que permaneceram bloqueadas, também chamadas de glosadas ou sustadas, por apresentarem irregularidades segundo as normas do SUS. Este subconjunto foi extraído do conjunto de dados C2.

Os campos ativos foram APRES e HOSP e os campos suplementares foram ESPEC, FAIXADIAS, CRS, VALTOTAL, CUSTO_AIH, DIAG_PRI, GRUPOCID, PROC_REA e CODMOTIVO.

- **Resultado**

O resultado é mostrado na Tabela 5.20 e pode ser visualizado na Figura F do Anexo. Os padrões encontrados são bastante variáveis.

TABELA 5.20 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs que permanecem bloqueadas de hospitais PORTE 4.

AGRUP	APRES	HOSP	ESPEC	FAIXA DIAS	CRS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPO CID	PROC_ REA	COD MOTIVO
4 (5,63%)	72000	HOSP41	03	7 a 9	5	350,00	B	N111	GP10	80500072	17
2 (38,36%)	82000	HOSP66	03	4 a 6	1	350,00	B	G458	GP06	81500106	07
3 (14,32%)	82000	HOSP287	03	4 a 6	7	250,00	B	G458	GP06	81500106	06
0 (8,95%)	82000	HOSP186	03	4 a 6	2	250,00	B	G458	GP06	81500106	06
5 (11,00%)	92000	HOSP158	03	1 a 3	17	550,00	M	A419	GP01	74300261	07
7 (5,88%)	92000	HOSP251	03	1 a 3	1	150,00	B	K810	GP11	31000002	07
1 (4,35%)	112000	HOSP213	03	4 a 6	7	250,00	B	A419	GP01	81500106	07
8 (5,12%)	122000	HOSP213	03	4 a 6	7	250,00	B	G458	GP15	81500106	07
6 (6,39%)	12001	HOSP262	03	4 a 6	3	150,00	B	A419	GP15	75500272	07

Legenda:
ESPEC: 03 – especialidade médica em clínica médica; 01 – especialidade médica em cirurgia geral.
DIAG_PRI: N111 – Pielonefrite obstrutiva crônica; G458 – Outros acidentes isquêmicos cerebrais trans sindr corr; A419 – Septicemia neonatal; K810 – Colecistite aguda.
GRUPOCID: GP10 - Doenças do aparelho respiratório; GP06 – Doenças do sistema nervoso; GP01 – Algumas doenças infecciosas e parasitárias; GP11 – Doenças do aparelho digestivo; GP15 – Gravidez, parto e puerpério.
PROC_REA: 80500072 - Pielonefrite; 81500106 – AVC agudo; 74300261– Septicemia (pediatria); 31000002– Cirurgia múltipla; 75500272 – Colecistite aguda.
CODMOTIVO: 17 – Homônimos e Politraumatizados; 07 – Homônimos; 06 – AVC agudo.

O número de internações que permaneceram bloqueadas aumentou significativamente na apresentação de agosto/2000. O motivo se explica, em particular, pelo novo critério de bloqueio que iniciou em julho/2000 (AVC agudo) segundo consta no Relatório Anual da SES/2000 [RIO 2000]. A quantidade de AIHs glosadas foi diminuindo no decorrer do tempo.

Os agrupamentos 2, 0 e 7, são formados por registros os quais são todos pertencentes aos hospitais HOSP66, HOSP186 e HOSP251, respectivamente, e são vistos como *outliers*.

O valor do Condorcet para os agrupamentos gerados foi de 0,6058 e o tempo de execução foi de 3min32.

Mineração 2.7 – Análise das AIHs sem resposta do auditor de hospitais PORTE 4 no período de maio a dezembro/2000.

Para a criação deste modelo de mineração, foi utilizado o subconjunto de 413 internações bloqueadas e que ficaram sem resposta do auditor. Este subconjunto foi extraído do conjunto de dados C2.

Os campos ativos foram APRES e HOSP e os campos suplementares foram ESPEC, FAIXADIAS, CRS, VALTOTAL, CUSTO_AIH, DIAG_PRI, GRUPOCID, PROC_REA e CODMOTIVO.

- **Resultado**

O resultado é mostrado na Tabela 5.21 e pode ser visualizado na Figura G do Anexo. Os padrões encontrados são bastante variáveis.

O valor do Condorcet para os agrupamentos gerados foi de 0,7066, e se apresenta mais elevado em relação aos anteriores. Isto ocorre por que o campo ativo APRES apresenta agora somente três meses de apresentação, facilitando a formação dos agrupamentos. O tempo de execução foi de 1min21.

TABELA 5.21 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs sem resposta do auditor de hospitais PORTE 4.

AGRUP	APRES	HOSP	ESPEC	FAIXA DIAS	CRS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPO CID	PROC_ REA	COD MOTIVO
8 (5,08%)	62000	HOSP158	03	4 a 6	17	250,00	M	G458	GP19	39000001	07
3 (1,21%)	62000	HOSP186	03	0	1	1250,00	A	Z949	GP21	62001000	05
4 (0,48%)	62000	HOSP287	01	13 a 15	7	750,00	A	K929	GP11	31000002	04
0 (0,24%)	62000	HOSP186	01	13 a 15	2	1750,00	A	S822	GP19	31000002	04
7 (10,17%)	72000	HOSP251	01	1 a 3	1	250,00	B	K810	GP11	77500113	18
5 (45,52%)	82000	HOSP221	01	4 a 6	3	250,00	B	G458	GP09	80500072	17
1 (18,40%)	82000	HOSP220	03	4 a 6	3	250,00	B	G458	GP11	81500106	07
2 (14,53%)	82000	HOSP325	03	7 a 9	10	250,00	B	J960	GP10	81500106	18
6 (4,36%)	82000	HOSP262	03	1 a 3	3	250,00	M	I509	GP09	77500113	20

Legenda:

ESPEC: 03 – especialidade médica em clínica médica; 01 – especialidade médica em cirurgia geral.

DIAG_PRI: G458 – Outros acidentes isquêmicos cerebrais trans sindr corr; Z949 – Órgão e tec NE transplantado; K929 – Doença do aparelho digestivo SOE; S822 – Frat da diafise da tíbia; K810 – Colecistite aguda; J960 – Insuf respirat aguda; I509 - Insuf cardíaca NE.

GRUPOCID: GP19 – Lesões, enven e algumas outr conseq de causas externas; GP21 – Fat que infl o est de saúde e o contato c/ os ser de saúde; GP11 – Doenças do aparelho digestivo; GP09 – Doenças do aparelho circulatório; GP10 - Doenças do aparelho respiratório.

PROC_REA: 39000001 – Politraumatizado; 62001000 – Busca ativa de doador de órgão; 31000002 - Cirurgia múltipla; 77500113 – Insuficiência cardíaca; 80500072 – Pielonefrite; 81500106 – AVC agudo; 77500113 - Insuficiência cardíaca.

CODMOTIVO: 07 – Homônimos; 05 – Transplante; 04 - Cirurgias múltiplas; 18 – Homônimos/duplicidade e reapresentada que permanece bloqueada; 17 – Homônimos e Politraumatizados; 20 – Reapresentada que permanece bloqueada.

Esses conjunto de bloqueios é constituído de casos incomuns, cuja conclusão é mais demorada. Por esse motivo, observa-se a formação de pequenos agrupamentos, com características diferentes. Os bloqueios desse tipo só ocorreram, para hospitais desse porte, até a apresentação de agosto/2000.

Até aqui, foi visto que os padrões encontrados nos registros não bloqueados apresentam um comportamento regular no decorrer do tempo. Enquanto que os registros bloqueados apresentam um comportamento variável no decorrer do tempo. Os agrupamentos das minerações realizadas precisam ser refinados, para que se possa obter detalhes mais específicos sobre as internações.

O próximo passo foi a construção de modelos mais direcionados, verificando-se o perfil, para cada tipo de problema de bloqueios, dos hospitais com o maior percentual de internações: liberadas com código novo (HOSP158), liberadas com o mesmo código (HOSP211) e o glosadas (HOSP66), no período estudado. Os resultados serão apresentados a seguir, agrupados pelo tipo de problemas de bloqueio, para que se possa comparar os perfis de comportamento dos hospitais relacionados. Foram criados modelos por mês de apresentação para cada um desses hospitais e foram escolhidos os agrupamentos de maior tamanho, os quais compõem as linhas das tabelas que serão apresentadas a seguir.

Mineração 2.8 – Análise das AIHs bloqueadas e liberadas com código novo do HOSP158.

Os campos ativos foram CODMOTIVO, PROC_REA, GRUPOCID e CUSTO_AIH e os campos suplementares foram ESPEC, FAIXADIAS, VALTOTAL, DIAG_PRI E NOVO_COD. Foram criados modelos para cada mês de apresentação e são apresentados na Tabela 5.22 os valores do maior agrupamento de cada mês.

TABELA 5.22 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas e liberadas com código novo do HOSP158.

% AGRUP MAIS FREQ	No. DE AIHs	APRES	PROC_ REA	ESPEC	FAIXA DIAS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPO CID	NOVO COD	COD MOTI VO
-	-	62000	-	-	-	-	-	-	-	-	-
100,00	1	72000	77500164	03	7 a 9	459,24	B	J81	GP10	77500113	17
40,00	15	82000	74300261	07	7 a 9	536,76	M	A419	GP01	76300080	17
-	-	92000	-	-	-	-	-	-	-	-	-
85,00	20	102000	74300261	07	4 a 6	528,81	M	A419	GP01	76500071	01
85,71	14	112000	74300261	07	7 a 9	542,06	M	A419	GP01	76400085	01
100,00	3	122000	74300261	03	10 a 12	518,21	M	A419	GP01	76500063	01
-	-	12001	-	-	-	-	-	-	-	-	-

Legenda:
ESPEC: 03 – especialidade médica em clínica médica; 07 – especialidade médica em pediatria.
DIAG_PRI: J81– Edema pulmonar NE de outr form; A419 – Septicemia neonatal.
GRUPOCID: GP10 – Doenças do aparelho respiratório; GP01 – Algumas doenças infecciosas e parasitárias.
PROC_REA: 77500164 - Edema agudo de pulmão; 74300261 - Septicemia (pediatria).
NOVO_COD: 77500113 – Insuficiência cardíaca; 76300080 - Broncopneumonia; 76500071– Broncopneumonia; 76400085 – broncopneumonia em lactente; 76500063 – Pneumonia não especificada.
CODMOTIVO: 17 – Homônimos e Politraumatizados; 01 – Septicemia.

Os resultado da Tabela 5.22 indicam que as internações do HOSP158 liberadas com código novo ocorreram, na maior parte das vezes, por doenças diagnosticadas como doenças infecciosas e parasitárias, que eram cobradas como septicemia neonatal, e tiveram que mudar o procedimento para outros procedimentos desse grupo de doenças. O número dessas internações foi diminuindo e não houve nenhuma no último mês do ano.

Mineração 2.9 – Análise das AIHs liberadas com código novo do HOSP211 no período de maio a dezembro/2000.

Os campos ativos foram CODMOTIVO, PROC_REA, GRUPOCID e CUSTO_AIH e os campos suplementares foram ESPEC, FAIXADIAS, VALTOTAL, DIAG_PRI E NOVO_COD. Foram criados modelos para cada mês de apresentação e são apresentados na Tabela 5.23 os valores do maior agrupamento de cada mês.

Os resultados da Tabela 5.23 mostram que as internações do HOSP211 que foram liberadas com código novo apresentaram poucos casos, e com padrões bem diferentes uns dos outros.

TABELA 5.23 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas e liberadas com código novo do HOSP211.

% AGRUP MAIS FREQ	No. DE AIHs	APRES	PROC_ REA	ESPEC	FAIXA DIAS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPO CID	NOVO COD	COD MOTIVO
50,00	2	62000	77500202	03	13 a 15	795,18	M	I739	GP09	32039042	07
100,00	1	72000	39006123	01	7 a 9	1640,03	A	Y839	GP20	38018012	17
83,33	6	82000	81500106	03	4 a 6	275,55	B	G458	GP06	81500076	06
-	-	92000	-	-	-	-	-	-	-	-	-
-	-	102000	-	-	-	-	-	-	-	-	-
-	-	112000	-	-	-	-	-	-	-	-	-
100,00	1	122000	40200000	01	> 18	1037,90	A	S069	GP19	38027011	07
100,00	1	12001	74300261	07	13 a 15	557,96	M	A419	GP01	76300080	01

Legenda:

ESPEC: 03 – especialidade médica em clínica médica; 01 – especialidade médica em cirurgia geral; 07 – Especialidade médica em pediatria.

DIAG_PRI: I739 – Doenças vasculares periféricas NE; Y839 - Intervenção cirúrgica NE; G458 – Outros acidentes isquêmicos cerebrais trans sindr corr; A419 – Septicemia neonatal.

GRUPOCID: GP09 - Doenças do aparelho circulatório; GP20 – Causas externas de morbidade e mortalidade; GP06 – Doenças do sistema nervoso; GP19 - Lesões,enven e algumas outr conseq de causas externas.

PROC_REA: 77500202 – Vasculopatia periférica; 39006123 - Desarticulação da articulação coxo femoral; 81500106 – AVC agudo; 40200000 – Tratamento conservador do traumatismo cranioencefálico; 74300261– Septicemia (pediatria).

CODMOTIVO: 07 – Homônimos; 17 – Homônimos e Politraumatizados; 06 – AVC agudo; 01 - Septicemia.

Mineração 2.10 – Análise das AIHs liberadas com código novo do HOSP66 no período de maio a dezembro/2000.

Os campos ativos foram CODMOTIVO, PROC_REA, GRUPOCID e CUSTO_AIH e os campos suplementares foram ESPEC, FAIXADIAS, VALTOTAL, DIAG_PRI E NOVO_COD. Foram criados modelos para cada mês de apresentação e são apresentados na Tabela 5.24 os valores do maior agrupamento de cada mês.

Os resultados da Tabela 5.24 mostram que as internações do HOSP66 liberadas com código novo apresentaram a maior parte dos casos de bloqueio por AVC agudo. No último mês do ano, no entanto, não foi registrado nenhum caso.

TABELA 5.24 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas e liberadas com código novo do HOSP66.

% AGRUP MAIS FREQ	No. DE AIHs	APRES	PROC_ REA	ESPEC	FAIXA DIAS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPO CID	NOVO COD	COD MOTI VO
50,00	4	62000	39000001	01	7 a 9	689,75	M	T009	GP19	31000002	03
50,00	2	72000	39000001	01	7 a 9	966,44	M	T009	GP19	38025019	03
100,00	1	82000	74300261	07	4 a 6	518,21	M	A419	GP01	74300270	19
40,00	5	92000	81500106	03	1 a 3	435,00	B	G458	GP06	81500076	06
33,00	6	102000	81500106	03	1 a 3	275,55	B	G458	GP06	77500130	06
83,00	6	112000	81500106	03	4 a 6	362,30	B	G458	GP06	77500130	06
100,00	6	122000	81500106	03	1 a 3	362,30	B	G458	GP06	77500121	06
-	-	12001	-	-	-	-	-	-	-	-	-

Legenda:
ESPEC: 01 – Especialidade médica em cirurgia geral; 07 – Especialidade médica em pediatria; 03 – especialidade médica em clínica médica.
DIAG_PRI: T009 – Traum superf mult NE; A419 – Septicemia neonatal; G458 – Outros acidentes isquêmicos cerebrais trans sindr corr.
GRUPOCID: GP19 - Lesões,enven e algumas outr conseq de causas externas; GP01 – Algumas doenças infecciosas e parasitárias; GP06 – Doenças do sistema nervoso.
PROC_REA: 39000001 – Politraumatizado; 74300261 – Septicemia (pediatria); 81500106 – AVC agudo.
NOVO_COD: 31000002– Cirurgia múltipla; 38025019 - Perda de substancia cutânea - lesões extensas planos superficial; 74300270 – Entero infecções (pediatria); 81500076– Epilepsias; 77500130 - Hipertensão maligna; 77500121 – Crise hipertensiva.
CODMOTIVO: 03 – Politraumatizados; 19 – Homônimos/duplicidade e septicemia; 06 – AVC agudo.

Comparando-se os três hospitais, verifica-se que apresentam comportamentos diferentes para os bloqueios liberados com código novo. O HOSP158 apresenta casos mais freqüentes de internações por doenças infecciosas e parasitárias. O HOSP211 não apresenta uma tendência específica e o HOSP66 apresenta casos mais freqüentes de doenças do sistema nervoso.

Os auditores recomendaram observar casos como o que aparece na Tabela 5.22, em que não houve bloqueios para a apresentação 92000 do HOSP158. Já houve casos em que o sistema não detectou as impropriedades, que se modificaram no decorrer do tempo. Em casos como esse, a análise de agrupamentos pode ser usada de forma preditiva, para identificar padrões de comportamentos semelhantes, conforme mostra o exemplo da próxima mineração.

Mineração 2.11 – Análise das AIHs bloqueadas e liberadas com mesmo código do HOSP158 na apresentação de 9/2000 para a identificação de padrões semelhantes aos das AIHs liberadas com código novo na apresentação de 8/2000 desse hospital.

O IM possibilita a utilização da pesquisa de agrupamento no modo de aplicação. Desta forma, foi realizada uma pesquisa de agrupamento que gerou um modelo para identificar os padrões das internações liberadas com código novo no mês de apresentação 8/2000 do HOSP158. Como não houve internações desse tipo na apresentação do mês seguinte, esse modelo foi aplicado sobre os dados das internações liberadas com mesmo código da apresentação 9/2000, com a finalidade de identificar

padrões semelhantes, ou seja, internações que poderiam ter sido bloqueadas e que poderiam ter seu custo reduzido. Das 63 internações liberadas com mesmo código em 9/2000, foram identificados 41 internações com padrões semelhantes com uma taxa de acerto de 70%. Esses casos são passíveis de investigação.

Mineração 2.12 – Análise das AIHs bloqueadas e liberadas com mesmo código do HOSP158 no período de maio a dezembro/2000.

Os campos ativos foram CODMOTIVO, PROC_REA, GRUPOCID e CUSTO_AIH e os campos suplementares foram ESPEC, FAIXADIAS, VALTOTAL, DIAG_PRI E NOVO_COD. Foram criados modelos para cada mês de apresentação e são apresentados na Tabela 5.25 os valores do maior agrupamento de cada mês.

Os resultados da Tabela 5.25 indicam que as internações do HOSP158 liberadas com o mesmo código ocorreram, na maior parte das vezes, também por doenças diagnosticadas como doenças infecciosas e parasitárias, com procedimento realizado mais freqüente de septicemia (clínica médica). Esse tipo de AIH bloqueada ocorreu durante todo o período estudado, exceto na apresentação de 6/2000.

TABELA 5.25 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas e liberadas com mesmo código do HOSP158.

% AGRUP MAIS FREQ	No. DE AIHs	APRES	PROC_ REA	ESPEC	FAIXA DIAS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPOCID	COD MOTIVO
-	-	62000	-	-	-	-	-	-	-	-
20,00	10	72000	40205002	01	10 a 12	628,13	M	I674	GP09	17
12,60	71	82000	77500113	03	4 a 6	500,00	M	I509	GP09	17
39,60	63	92000	74300261	07	4 a 6	534,11	M	A419	GP01	01
36,36	22	102000	74500244	03	7 a 9	518,21	M	A419	GP01	01
32,26	31	112000	85500755	04	> 18	1754,10	A	G328	GP06	02
48,89	45	122000	74500244	03	10 a 12	518,21	M	A419	GP01	01
36,11	36	12001	74500244	03	7 a 9	518,21	M	A419	GP01	01

Legenda:

ESPEC: 01– Especialidade médica em cirurgia geral; 03 – Especialidade médica em clínica médica; 07 – Especialidade médica em pediatria; 04 – Especialidade médica em crônico e FTP.

DIAG_PRI: I674 – Encefalopatia hipertensiva; I509 - Insuf cardíaca NE; A419 – Septicemia neonatal; G328 – Outr transt degener espec sist nerv doen COP.

GRUPOCID: GP09 – Doenças do aparelho circulatório; GP01 – Algumas doenças infecciosas e parasitárias; GP06 – Doenças do sistema nervoso.

PROC_REA: 40205002 - Tratamento conservador da hipertensão intracraniana; 77500113 – Insuficiência cardíaca; 74300261 - Septicemia (pediatria); 74500244 - Septicemia (clinica medica); 85500755 – Paciente sob cuidados prolongados por enfermidades neurológicas.

CODMOTIVO: 17 – Homônimos e Politraumatizados; 01 – Septicemia; 02 - Cuidados prolongados.

Mineração 2.13 – Análise das AIHs liberadas com mesmo código do HOSP211 no período de maio a dezembro/2000.

Os campos ativos foram CODMOTIVO, PROC_REA, GRUPOCID e CUSTO_AIH e os campos suplementares foram ESPEC, FAIXADIAS, VALTOTAL, DIAG_PRI E NOVO_COD. Foram criados modelos para cada mês de apresentação e são apresentados na Tabela 5.26 os valores do maior agrupamento de cada mês.

Os resultado da Tabela 5.26 mostram que as internações do HOSP211 que foram liberadas com mesmo código apresentaram padrões variados e foram bloqueadas na maioria das vezes pelo motivo de homônimos.

TABELA 5.26 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas e liberadas com mesmo código do HOSP211.

% AGRUP MAIS FREQ	No. DE AIHs	APRES	PROC_ REA	ESPEC	FAIXA DIAS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPOCID	COD MOTIVO
30,21	96	62000	31008011	01	4 a 6	750,00	M	N200	GP19	07
17,71	96	72000	76500225	03	4 a 6	250,00	B	J449	GP10	17
25,00	108	82000	81500106	03	4 a 6	250,00	B	G458	GP06	06
20,16	129	92000	39011119	01	1 a 3	750,00	M	M869	GP13	07
20,00	60	102000	33016119	01	1 a 3	625,00	M	K929	GP11	07
16,39	122	112000	33022119	01	4 a 6	250,00	B	K469	GP11	07
25,33	129	122000	32019041	01	7 a 9	500,00	M	I509	GP09	07
29,23	138	12001	38018012	01	7 a 9	250,00	B	J180	GP10	07

Legenda:
ESPEC: 01 – especialidade médica em cirurgia geral; 03 – especialidade médica em clínica médica.
DIAG_PRI: N200 – Calculose do rim; J449 – Doença pulmonar obstrutiva crônica NE; G458 – Outros acidentes isquêmicos cerebrais trans sindr corr; M869 – Osteomielite NE; K929 – Doença do aparelho digestivo SOE; K469 – Hérnia abdominal NE s/obstrução ou gangrena; I509 – Insuf cardíaca NE; J180 – Broncopneumonia NE.
GRUPOCID: GP19 - Lesões,enven e algumas outr conseq de causas externas; GP10 - Doenças do aparelho respiratório; GP06 – Doenças do sistema nervoso; GP13 – Doenças sist osteomuscular e tecido conjuntivo; GP11 – Doenças do aparelho digestivo; GP09 – Doenças do aparelho circulatório.
PROC_REA: 31008011 – Nefrolitotomia; 76500225 – Doença pulmonar obstrutiva crônica; 81500106 – AVC agudo; 39011119 – Tratamento cirúrgico da osteomielite da pelve; 32019041 – Bypass ou endarterectomia femoro popliteia; 38018012 - Debridamento da fasceite necrotizante.
CODMOTIVO: 07 – Homônimos; 17 – Homônimos e Politraumatizados; 06 – AVC agudo.

Mineração 2.14 – Análise das AIHs liberadas com mesmo código do HOSP66 no período de maio a dezembro/2000.

Os campos ativos foram CODMOTIVO, PROC_REA, GRUPOCID e CUSTO_AIH e os campos suplementares foram ESPEC, FAIXADIAS, VALTOTAL, DIAG_PRI E NOVO_COD. Foram criados modelos para cada mês de apresentação e são apresentados na Tabela 5.27 os valores do maior agrupamento de cada mês.

TABELA 5.27 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas e liberadas com mesmo código do HOSP66.

% AGRUP MAIS FREQ	No. DE AIHs	APRES	PROC_ REA	ESPEC	FAIXA DIAS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPOCID	COD MOTIVO
31,19	109	62000	77500202	03	7 a 9	250,00	B	I739	GP11	07
34,38	96	72000	38025019	01	4 a 6	450,00	B	T009	GP19	17
24,74	97	82000	81500106	03	4 a 6	300,00	B	G458	GP06	06
34,26	108	92000	38025019	01	1 a 3	300,00	B	T009	GP19	07
17,91	134	102000	81500106	03	1 a 3	300,00	B	G458	GP14	07
25,81	93	112000	33004080	01	4 a 6	500,00	M	W199	GP20	07
31,25	96	122000	81500106	01	1 a 3	350,00	B	G458	GP06	06
27,93	111	12001	81500106	03	1 a 3	350,00	B	G458	GP06	06

Legenda:
ESPEC: 03 – especialidade médica em clínica médica; 01 – especialidade médica em cirurgia geral.
DIAG_PRI: I739 – Doenças vasculares periféricas NE ; T009 – Traum superf mult NE; G458 – Outros acidentes isquêmicos cerebrais trans sindr corr; W199 – Local NE.
GRUPOCID: GP11 – Doenças do aparelho digestivo; GP19 - Lesões,enven e algumas outr conseq de causas externas; GP06 – Doenças do sistema nervoso; GP14 – Doenças do aparelho geniturinário; GP20 – Causas externas de morbidade e mortalidade.

PROC_REA: 77500202 – Vasculopatia periférica; 38025019 - Perda de substancia cutânea - lesões extensas planos superficial; 81500106 – AVC agudo; 33004080 – Colecistectomia.
CODMOTIVO: 07 – Homônimos; 17 – Homônimos e Politraumatizados; 06 – AVC agudo.

Os resultados da Tabela 5.27 mostram que as internações mais freqüentes do HOSP66 liberadas com mesmo código apresentaram nos últimos meses do ano, bloqueio por AVC agudo. Este fato chama a atenção, tendo em vista que, aparentemente, as regras do critério de AVC agudo não foram suficientes para filtrar esses casos apresentados pelo HOSP66.

Comparando-se os três hospitais, verifica-se que apresentam comportamentos diferentes para os bloqueios liberados com mesmo código novo. O HOSP158 apresenta casos mais freqüentes de internações por doenças infecciosas e parasitárias. O HOSP211 e o HOSP66 não apresentaram uma tendência específica.

Mineração 2.15 – Análise das AIHs que permaneceram bloqueadas do HOSP158 no período de maio a dezembro/2000.

Os campos ativos foram CODMOTIVO, PROC_REA, GRUPOCID e CUSTO_AIH e os campos suplementares foram ESPEC, FAIXADIAS, VALTOTAL, DIAG_PRI E NOVO_COD. Foram criados modelos para cada mês de apresentação e são apresentados na Tabela 5.28 os valores do maior agrupamento de cada mês.

Os resultados da Tabela 5.28 mostram que as internações do HOSP158 que foram glosadas por apresentarem impropriedades segundo as normas do SUS ocorreram, na maior parte das vezes, também por doenças diagnosticadas como doenças infecciosas e parasitárias, com procedimento realizado mais freqüente de septicemia (clínica médica). Porém, não mais ocorreram nos últimos meses do ano.

TABELA 5.28 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs que permaneceram bloqueadas do HOSP158.

% AGRUP MAIS FREQ	No. DE AIHs	APRES	PROC_ REA	ESPEC	FAIXA DIAS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPOCID	COD MOTIVO
-	-	62000	-	-	-	-	-	-	-	-
-	-	72000	-	-	-	-	-	-	-	-
40,00	5	82000	76500233	03	1 a 3	425,10	B	J960	GP10	17
66,67	27	92000	74300261	07	4 a 6	518,21	M	A419	GP01	01
75,00	4	102000	74500244	03	4 a 6	518,21	M	A419	GP01	01
50,00	2	112000	74500244	03	4 a 6	518,21	M	A419	GP01	01
-	-	122000	-	-	-	-	-	-	-	-
-	-	12001	-	-	-	-	-	-	-	-

Legenda:
ESPEC: 03 – Especialidade médica em clínica médica; 07 – Especialidade médica em pediatria.
DIAG_PRI: J960 – Insuf respirat aguda; A419 – Septicemia neonatal.
GRUPOCID: GP10 - Doenças do aparelho respiratório; GP01 – Algumas doenças infecciosas e parasitárias.
PROC_REA: 76500233 - Insuficiência respiratória aguda; 74300261– Septicemia (pediatria); 74500244 - Septicemia (clinica medica).
CODMOTIVO: 17 – Homônimos e Politraumatizados; 01 – Septicemia.

Mineração 2.16 – Análise das AIHs que permaneceram bloqueadas do HOSP211 no período de maio a dezembro/2000.

Os campos ativos foram CODMOTIVO, PROC_REA, GRUPOCID e CUSTO_AIH e os campos suplementares foram ESPEC, FAIXADIAS, VALTOTAL, DIAG_PRI E NOVO_COD. Foram criados modelos para cada mês de apresentação e são apresentados na Tabela 5.29 os valores do maior agrupamento de cada mês.

Os resultados da Tabela 5.29 mostram que as internações do HOSP211 que foram glosadas apresentaram poucos casos durante o ano, com padrões mais freqüentes bem diferentes nos meses em que ocorreram.

TABELA 5.29 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs que permaneceram bloqueadas do HOSP211.

% AGRUP MAIS FREQ	No. DE AIHs	APRES	PROC_ REA	ESPEC	FAIXA DIAS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPOCID	COD MOTIVO
50,00	2	62000	33021066	01	7 a 9	896,52	M	K566	GP11	07
100,00	1	72000	76400085	07	10 a 12	441,00	B	J180	GP10	17
-	-	82000	-	-	-	-	-	-	-	-
-	-	92000	-	-	-	-	-	-	-	-
-	-	102000	-	-	-	-	-	-	-	-
-	-	112000	-	-	-	-	-	-	-	-
-	-	122000	-	-	-	-	-	-	-	-
100,00	2	12001	77500202	03	> 18	425,27	B	I739	GP09	07

Legenda:
ESPEC: 01 – especialidade médica em cirurgia geral; 07 – especialidade médica em pediatria; 03 – Especialidade médica em clínica médica.
DIAG_PRI: K566 – Outr form de obstrução intestinal e as NE; J180 – Broncopneumonia NE; I739 – Doenças vasculares periféricas NE.
GRUPOCID: GP11 – Doenças do aparelho digestivo; GP10 - Doenças do aparelho respiratório; GP09 – Doenças do aparelho circulatório.
PROC_REA: 33021066 – Enterectomia; 76400085 – Broncopneumonia em lactente; 77500202 - Vasculopatia periférica.
CODMOTIVO: 07 – Homônimos; 17 – Homônimos e Politraumatizados.

Mineração 2.17 – Análise das AIHs que permaneceram bloqueadas do HOSP66 no período de maio a dezembro/2000.

Os campos ativos foram CODMOTIVO, PROC_REA, GRUPOCID e CUSTO_AIH e os campos suplementares foram ESPEC, FAIXADIAS, VALTOTAL, DIAG_PRI E NOVO_COD. Foram criados modelos para cada mês de apresentação e são apresentados na Tabela 5.30 os valores do maior agrupamento de cada mês.

Os resultados da Tabela 5.30 mostram que os padrões mais freqüentes se modificam a cada mês.

Comparando-se os três hospitais, verifica-se que apresentam comportamentos diferentes para os bloqueios que permaneceram bloqueados. O HOSP158 apresentou a mesma tendência dos casos anteriores. O HOSP211 e o HOSP66 não apresentaram uma tendência específica.

TABELA 5.30 – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs que permaneceram bloqueadas do HOSP66.

% AGRUP MAIS FREQ	No. DE AIHs	APRES	PROC_ REA	ESPEC	FAIXA DIAS	VAL TOTAL	CUSTO AIH	DIAG PRI	GRUPOCID	COD MOTIVO
30,00	10	62000	76500225	03	4 a 6	392,05	B	J449	GP10	07
33,33	9	72000	75500272	03	4 a 6	148,02	MB	K810	GP11	17
50,85	59	82000	81500106	03	4 a 6	362,30	B	G458	GP06	06
28,57	14	92000	77500202	03	4 a 6	199,39	MB	I739	GP09	07
20,00	25	102000	77500202	03	4 a 6	199,39	MB	I739	GP09	07
23,08	13	112000	75500272	03	4 a 6	146,62	MB	K810	GP11	07
43,75	16	122000	77500237	03	7 a 9	199,39	B	I809	GP09	07
25,00	4	12001	31002021	01	7 a 9	503,03	M	N289	GP14	07

Legenda:
ESPEC: 03 – especialidade médica em clínica médica; 01 – especialidade médica em cirurgia geral.
DIAG_PRI: J449 – Doença pulmonar obstrutiva crônica NE; K810 – Colecistite aguda; G458 – Outros acidentes isquêmicos cerebrais trans sindr corr; I739 – Doenc vasculares periféricas NE; I809 – I80.9 Flebite e tromboflebite de localiz NE; N289 – Transt NE do rim e do ureter.
GRUPOCID: GP10 - Doenças do aparelho respiratório; GP11 – Doenças do aparelho digestivo; GP06 – Doenças do sistema nervoso; GP09 – Doenças do aparelho circulatório; GP14 – Doenças do aparelho geniturinário.
PROC_REA: 76500225 – Doença pulmonar obstrutiva crônica; 75500272 – Colecistite aguda; 81500106 – AVC agudo; 77500202 – Vasculopatia periférica; 77500237 – Tromboflebites profundas; 31002021 – Tratamento cirúrgico da ureterocele.
CODMOTIVO: 07 – Homônimos; 17 – Homônimos e Politraumatizados; 06 – AVC agudo.

5.4.5 Conclusões e validação do Experimento 2

Com a pesquisa de agrupamento, utilizando-se o agrupamento demográfico do IBM Intelligent Miner, foi possível visualizar diversos padrões de comportamento das internações hospitalares bloqueadas do SIH/SUS sob a gestão da SES/RS.

Muitas questões têm sido discutidas sobre esses bloqueios pela auditoria e estão relacionadas com a eficiência dos bloqueios técnicos. O grande número de AIHs liberadas com o mesmo código de procedimento para pagamento tem dado a idéia de pouca eficácia e eficiência nas auditorias.

Os auditores, ao analisarem este experimento, comentaram que estes mostraram muitas informações úteis, que levam a diversas ações, baseadas em questões, como por exemplo: Por que tal prestador tem certo comportamento? Por que tal procedimento foi mais apresentado? Os gráficos de visualização gerados e relatórios que a ferramenta disponibiliza, após explicação de como deveriam ser interpretados, passaram a ser facilmente entendidos. Algumas observações que representaram novidade foram: a predominância de bloqueios de homônimos nas AIHs liberadas com mesmo código de hospitais PORTE 4. Segundo observações documentadas no início do trabalho, acreditava-se que os casos de homônimos eram característicos de hospitais pequenos. Além disso, são o que os diferenciam dos bloqueios liberados com novo código. A sugestão de estabelecer regras mais eficientes para esse tipo de bloqueio foi considerada interessante.

Com relação aos agrupamentos que foram mencionados como *outliers*, somente os especialistas poderão avaliar se esses casos merecem investigação. Esses agrupamentos e os últimos modelos gerados no Experimento 2, requisitariam uma análise muito demorada por parte dos especialistas e ficaram de ser observados futuramente.

5.5 Avaliação dos resultados

Os experimentos foram apresentados aos especialistas do domínio da aplicação. Alguns resultados foram validados, conforme foi mencionado nas conclusões do experimentos, outros exigem análise mais profunda, mas deram origem a novos questionamentos.

A análise de agrupamentos aqui apresentada mostrou, de forma superficial, o potencial de utilização desta técnica em aplicações reais. Se bem refinada, e com o conhecimento do domínio que os especialistas possuem, poderá mostrar resultados de padrões bem interessantes.

6 Considerações Finais

Este trabalho apresentou o processo de mineração de dados com a utilização de aprendizado não-supervisionado pela análise de agrupamentos, suas técnicas e métodos, itens relacionados e o uso de uma metodologia aplicada ao estudo de caso sobre dados reais da área da saúde.

O objetivo geral foi explorar a aplicação desta tecnologia em um banco de dados complexo do mundo real e avaliar se os resultados desta aplicação atendem aos critérios de validade, interpretabilidade, utilidade, originalidade e validação de hipóteses, almejados pela mineração de dados na descoberta de padrões interessantes.

Os resultados obtidos da aplicação são considerados positivos, uma vez que se conclui que a tecnologia estudada pode ser de grande ajuda para a solução de problemas como os deste estudo de caso. Mas, são positivos também por mostrarem as dificuldades enfrentadas e as falhas que costumam ocorrer durante a realização do processo, as quais muitas vezes podem levar a frustrações e até à desistência de se utilizar tal tecnologia. Experiências como esta reforçam, com novos exemplos práticos, os cuidados que devem ser tomados ao se utilizar o processo de MD e objetivam contribuir para o aperfeiçoamento desta tecnologia.

6.1 Conclusões

As conclusões extraídas desta experiência são enumeradas a seguir:

1) Sobre o processo de mineração de dados:

Na prática, foi possível constatar as possibilidades de ganhos que a MD pode proporcionar em decisões estratégicas. No entanto, não é um processo simples de ser realizado em bases de dados reais, conforme pode parecer inicialmente.

No decorrer da pesquisa, as maiores dificuldades se referem às fases da metodologia empregada de compreensão do domínio, compreensão dos dados e preparação de dados. Essas fases, importantíssimas para o sucesso da aplicação, representaram, na realidade, um grande "entrave" para o alcance da fase principal do trabalho, que é a modelagem, seguida da avaliação e aplicação.

Essas dificuldades ocorreram porque, na maior parte do tempo, não houve a real compreensão do domínio da aplicação. Alguns dos problemas encontrados estão relacionados à complexidade do sistema estudado, o qual se encontra ainda em evolução e não apresenta uma documentação detalhada e organizada que facilite o seu entendimento. Apesar da boa vontade que os especialistas do domínio da aplicação e demais técnicos da SES demonstraram em colaborar no processo, estes apresentaram dificuldades para explicar os dados e o funcionamento do sistema. A experiência mostrou que informações não documentadas ou documentadas de forma incompleta ficaram sujeitas a interpretações diversificadas de seus usuários, que acabaram transmitindo uma visão distorcida do problema que deveria ser analisado. Além disto, a

equipe envolvida foi se modificando com o passar do tempo e outras visões sobre os objetivos que deveriam ser alcançados para a otimização das atividades da auditoria mudaram os rumos da aplicação.

Por exemplo, boa parte dos esforços desta pesquisa foi direcionada no sentido de avaliar o comportamento das internações que permaneceram bloqueadas, pois a leitura de um relatório e a entrevista com alguns especialistas conduziram à idéia de que estas representavam o objetivo principal na investigação de AIHs bloqueadas. Somente em uma das validações mais recentes, um dos auditores esclareceu que o maior objetivo da auditoria não é a investigação dessas AIHs, mas sim a investigação das AIHs que são liberadas com o mesmo código, as quais representam trabalho extra para os auditores médicos, visto que são bloqueadas inutilmente. Então, significa que os critérios de bloqueios técnicos precisam ser melhorados para diminuir a quantidade de bloqueios desse tipo. Isto levou ao redirecionamento dos objetivos e a novos esforços de MD.

Com esta experiência, foi observado que é muito importante que as técnicas utilizadas no processo atendam aos requisitos de MD, mas também é indispensável que se consiga perceber os problemas do domínio da aplicação e que haja um forte entrosamento entre os profissionais envolvidos no trabalho. Como a comunicação entre os analistas de dados e os especialistas do domínio da aplicação muitas vezes resulta em mal entendidos, vê-se que é preciso dar mais atenção ao uso de instrumentos que possam facilitar e proporcionar a coleta de informações sobre o sistema de forma mais eficiente. Além disto, o sistema de banco de dados deve estar muito bem organizado e documentado, permitindo a extração de conhecimento interessante e confiável para a criação de modelos de mineração de dados realmente válidos e úteis.

Na preparação de dados, detectou-se que muita coisa ainda precisa ser melhorada. A literatura menciona as vantagens no uso de *data warehouses*, em que os dados já se encontram limpos, transformados e integrados, além de sumarizados e consolidados sob uma perspectiva histórica, facilitando em muito a aplicação do processo de MD. No entanto, na impossibilidade de utilização dessa tecnologia, uma alternativa é a utilização de ferramentas que facilitem o pré-processamento de dados.

O *Intelligent Miner*, ferramenta utilizada nesta pesquisa, disponibiliza uma série de funções de pré-processamento, mas para utilizá-las é preciso que os dados estejam em uma base de dados do SGBD DB2. Seria bastante desejável que houvesse uma integração maior entre sistemas gerenciadores de bancos de dados, ferramentas de pré-processamento e ferramentas de mineração de dados, em que se pudesse realizar o processo completo de MD em um só ambiente. A autora acredita que isto representa uma evolução natural para a utilização de MD.

2) Sobre o uso de aprendizado não-supervisionado pela descoberta de agrupamentos:

Com relação à descoberta de agrupamentos, foi constatado que essas técnicas são bastante adequadas para casos como o estudado neste trabalho. Os experimentos mostram que sua utilização, mesmo que em um nível superficial, foi bem sucedida na identificação de padrões médios de procedimentos de interesse e na identificação de desvios ou *outliers*. Se mais refinada e com o conhecimento de fundo dos especialistas,

a análise de agrupamentos poderá revelar conhecimento muito mais interessante, útil, novo e válido.

Para os novos usuários desta técnica, a maior dificuldade se refere à configuração de parâmetros para a formação dos agrupamentos, apesar de que os algoritmos estão realizando avanços com relação a esse fato e, na maioria das vezes, a utilização dos parâmetros padrões oferecidos pela ferramenta conduz a resultados satisfatórios. Entretanto, a configuração de parâmetros será melhor aproveitada conforme o conhecimento que se tiver dos dados.

3) Sobre a metodologia utilizada:

O uso efetivo de MD requer a integração do conhecimento do domínio da aplicação com as funções de MD, de forma que não se deve esperar que sistemas genéricos de MD sejam bem sucedidos em sistemas inteligentes de forma semelhante aos bancos de dados tradicionais. Assim, a metodologia utilizada neste trabalho permitiu o mapeamento de um modelo genérico para um modelo específico, integrando as características próprias do sistema estudado com o uso de técnicas de agrupamentos de dados.

A única observação se refere à tarefa de seleção de dados, tendo em vista que, ante à complexidade do sistema estudado, é natural que após a fase de avaliação, haja a necessidade de seleção de novos atributos, como ocorreu neste estudo de caso. Dessa forma, a tarefa de seleção de dados deve ser deslocada para após a limpeza, construção, integração e formatação de dados, evitando um retorno desnecessário para essas etapas.

4) Sobre a ferramenta utilizada:

Algumas vantagens observadas na utilização do IM foram as seguintes: é de fácil manuseio; suporta grandes quantidades de dados, por exemplo, alguns experimentos não apresentados neste trabalho foram realizados com sucesso sobre uma base de dados contendo 2.152.123 registros; realiza as principais tarefas de MD, mais especificamente, a tarefa de agrupamento de dados, de forma satisfatória, permitindo uma visualização acessível dos resultados.

Sobre a visualização dos resultados, ressalta-se que houve, por parte dos especialistas do domínio da aplicação, a facilidade em entender os gráficos dos agrupamentos gerados pelo IM, os quais aparentavam ser de difícil compreensão para novos usuários. Esta facilidade se deu em decorrência do conhecimento que esses profissionais possuem em estatística médica, o que permitiu a assimilação e o interesse pela análise de agrupamentos.

Algumas falhas detectadas no IM envolvem: operações de filtrar atributos da base de dados, como por exemplo, aos se estabelecer um filtro para obter internações pelo CGC de um hospital, a ferramenta retornava erro na execução; o uso da tarefa de regras associativas cuja opção se restringe à análise unidimensional; as regras de classificação que apresentam difícil visualização. Além disto, os relatórios estatísticos dos agrupamentos são gerados em formato *PostSript*, o que dificulta a transposição dos resultados para tabelas como as que foram apresentadas nesta dissertação.

Uma limitação que a ferramenta apresentou para a versão disponibilizada diz respeito ao hardware, que deve possuir uma boa capacidade de processamento e armazenamento, bem como a necessidade de uso de um sistema operacional cliente/servidor, o que nem sempre está ao alcance da maioria dos usuários.

6.2 Limitações da pesquisa

Um dos fatores que limitaram a avaliação de resultados mais precisos desta pesquisa foi a falta de respostas dos especialistas sobre diversas situações apontadas, e que precisariam ser investigadas na base de dados de 2000.

6.3 Contribuições da pesquisa

As contribuições desta pesquisa, dentro da especificidade do estudo de caso realizado, foram a descoberta de alguns padrões interessantes, a saber:

1) A constatação, pelos resultados apresentados, de que novos critérios de bloqueio técnico para o sistema da saúde podem ser criados ou que os critérios existentes podem ser melhorados com a observação dos resultados obtidos pela análise de agrupamentos. Alguns casos identificados foram:

- Que a maioria das internações realizadas no ano 2000 apresentou baixa permanência, fato que foi identificado rapidamente com a análise dos agrupamentos gerados pela mineração. Os auditores, no ano de 2001, chegaram a esse resultado mediante a análise estatística dos dados e estabeleceram um novo critério de bloqueio para internações de baixa permanência, mas comentaram que a visualização dos agrupamentos revela com clareza e de forma bem interessante o comportamento geral dos dados.

- Que a maior parte das internações liberadas com o mesmo código são internações cujo procedimento realizado foi septicemia ou AVC agudo, mas que foram bloqueadas por homônimos. Significa que essas internações não apresentaram motivos para serem bloqueadas pelas regras de septicemia e AVC agudo, mas apresentaram motivos para serem bloqueadas pelas regras de homônimos. No entanto, a auditoria veio a comprovar que eram cobranças adequadas e as liberou com o mesmo código. Portanto, uma melhora nas regras do critério de homônimos poderá filtrar esses casos e evitar o tempo gasto pela auditoria para examiná-los e depois comprovar que não estão ocorrendo impropriedades.

2) Uma observação que representou novidade foi a predominância de bloqueios pelo motivo de homônimos em hospitais aqui denominados como Porte 4. As observações documentadas no início da pesquisa se referem aos casos de homônimos como característicos de hospitais pequenos, isto é, com até 50 leitos.

3) O conhecimento proporcionado aos auditores médicos sobre as possibilidades de avaliar tendências das internações hospitalares e desvios em relação a essas tendências. Os auditores médicos viram que a utilização de agrupamentos possui

um alcance razoável para casos em que a auditoria não conseguiu ainda identificar um relacionamento aparente. As probabilidades de sucesso se baseiam no fato de a aplicação levar em conta o comportamento histórico do sistema, em que é possível visualizar perfis regulares nos dados e alterações de comportamentos delineadas no decorrer do tempo.

Os experimentos realizados provavelmente revelam um número maior de padrões interessantes, uma vez que muitos questionamentos feitos pelos auditores surgiram com relação aos resultados aqui apresentados. Porém, as limitações de tempo para o final desta pesquisa não permitiram uma análise mais aprofundada.

As contribuições desta pesquisa para o uso de aprendizado não-supervisionado pela análise de agrupamentos na área de mineração de dados foram:

1) A exploração desta tecnologia com a aplicação de uma metodologia e a utilização da uma ferramenta de MD, em que falhas e acertos resultaram em ganhos na compreensão do processo.

2) A parametrização para o tratamento de *outliers* na pesquisa de agrupamento do IM, mostrada no Experimento 1, que explora as possibilidades de aplicação desse recurso.

3) A utilização de agrupamento como uma tarefa preditiva, conforme demonstrado na Mineração 2.11.

6.4 Trabalhos futuros

Como trabalhos futuros para o sistema de auditoria médica da SES, recentemente, foram detectados novos problemas que podem gerar novos ciclos do processo de mineração de dados iniciado neste estudo de caso. Um deles é identificar o perfil do comportamento de auditores médicos ante as ações de bloqueio, como forma de avaliar a eficiência do sistema. Um outro problema consiste na observação de que as AIHs que permanecem bloqueadas têm sido reapresentadas com outra numeração, três ou quatro meses depois de terem sido bloqueadas, escapando ao controle da auditoria.

Com relação a análise de agrupamentos, propõem-se a exploração dos outros parâmetros que o *Intelligent Miner* oferece para a mineração de dados, permitindo uma compreensão maior das possibilidades oferecidas pelos algoritmos estudados. Ou a exploração de parâmetros oferecidos por outras ferramentas que realizem agrupamento.

Anexo - Modelos de mineração do Experimento 2

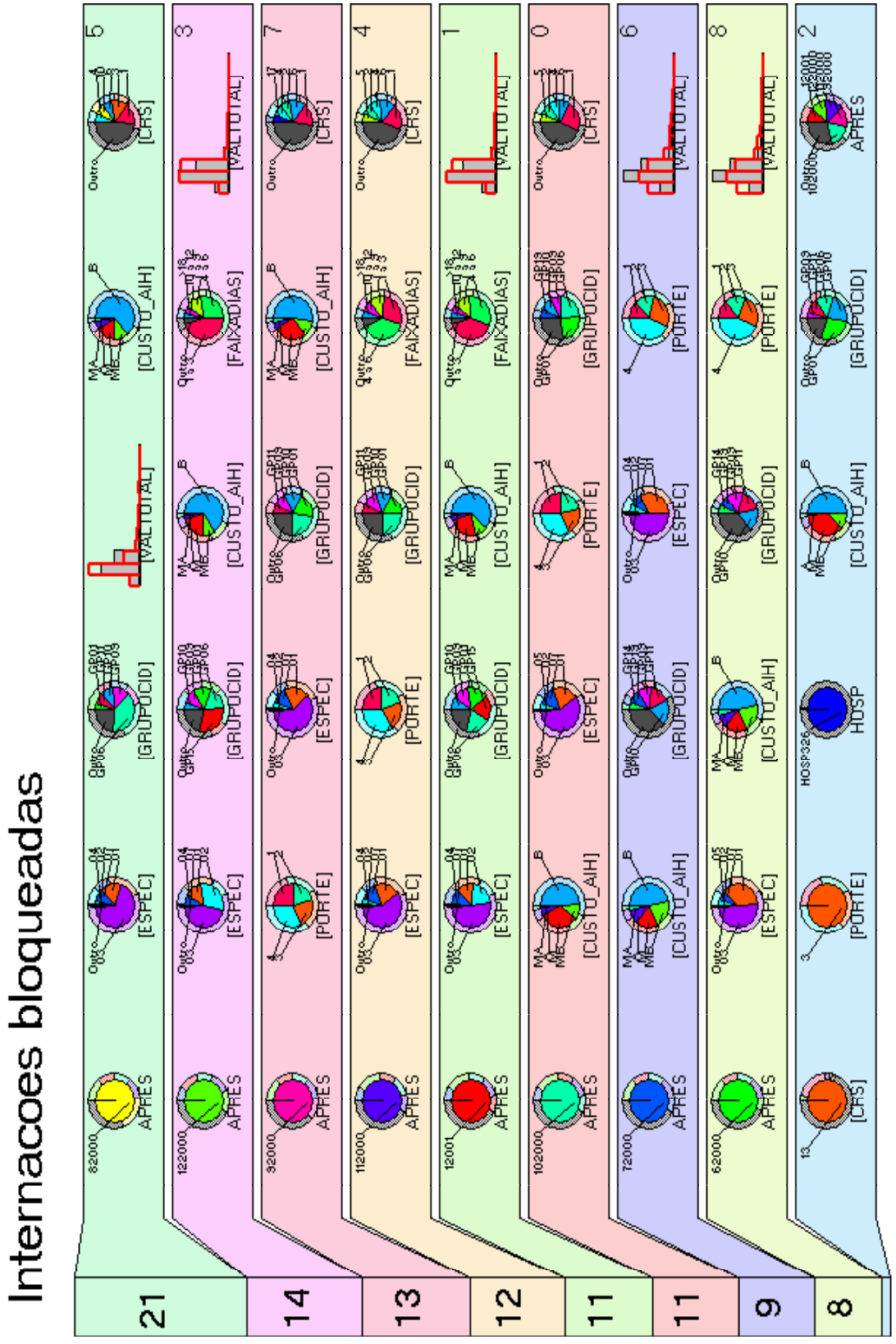


FIGURA A – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas.

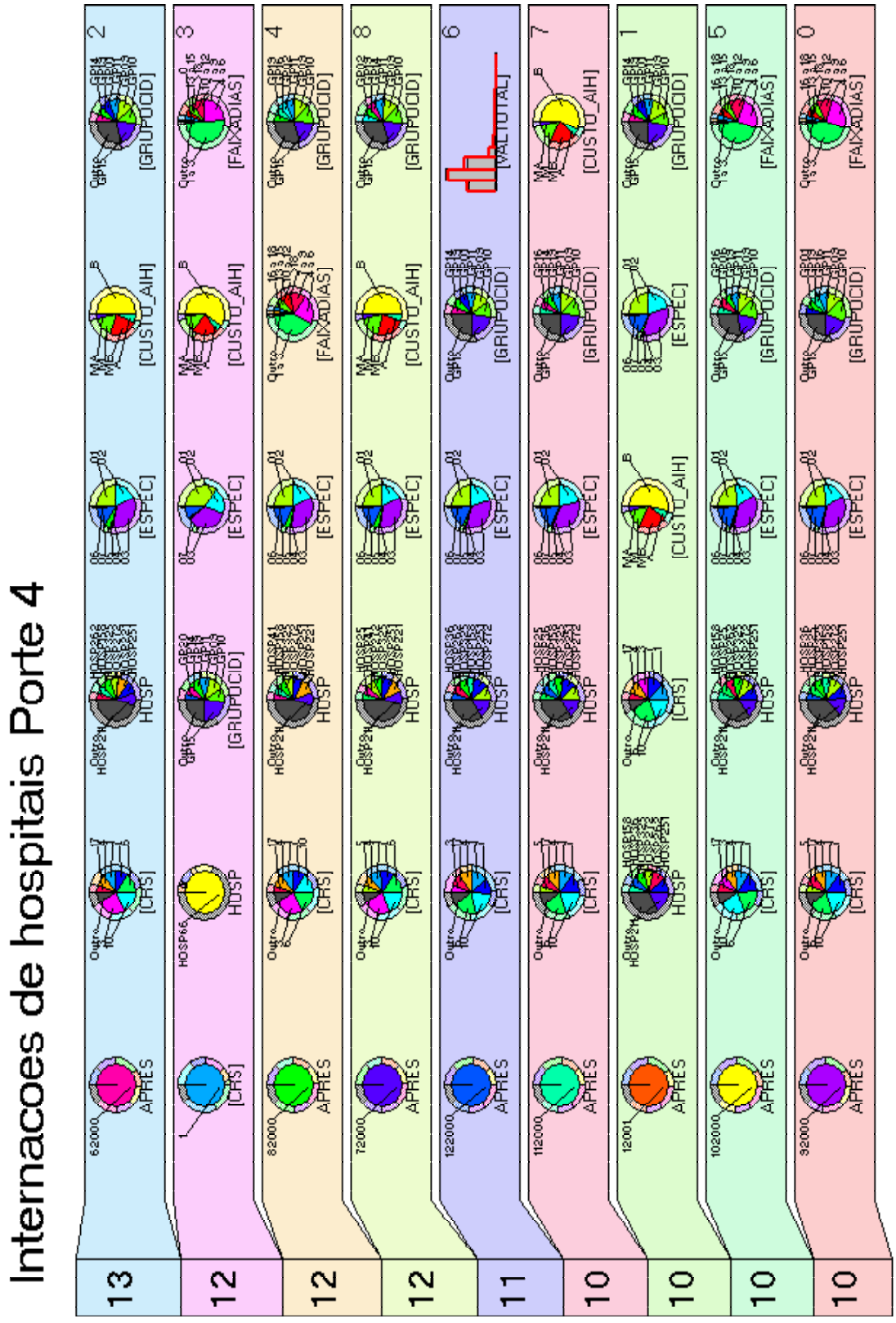


FIGURA B – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs de hospitais Porte 4.

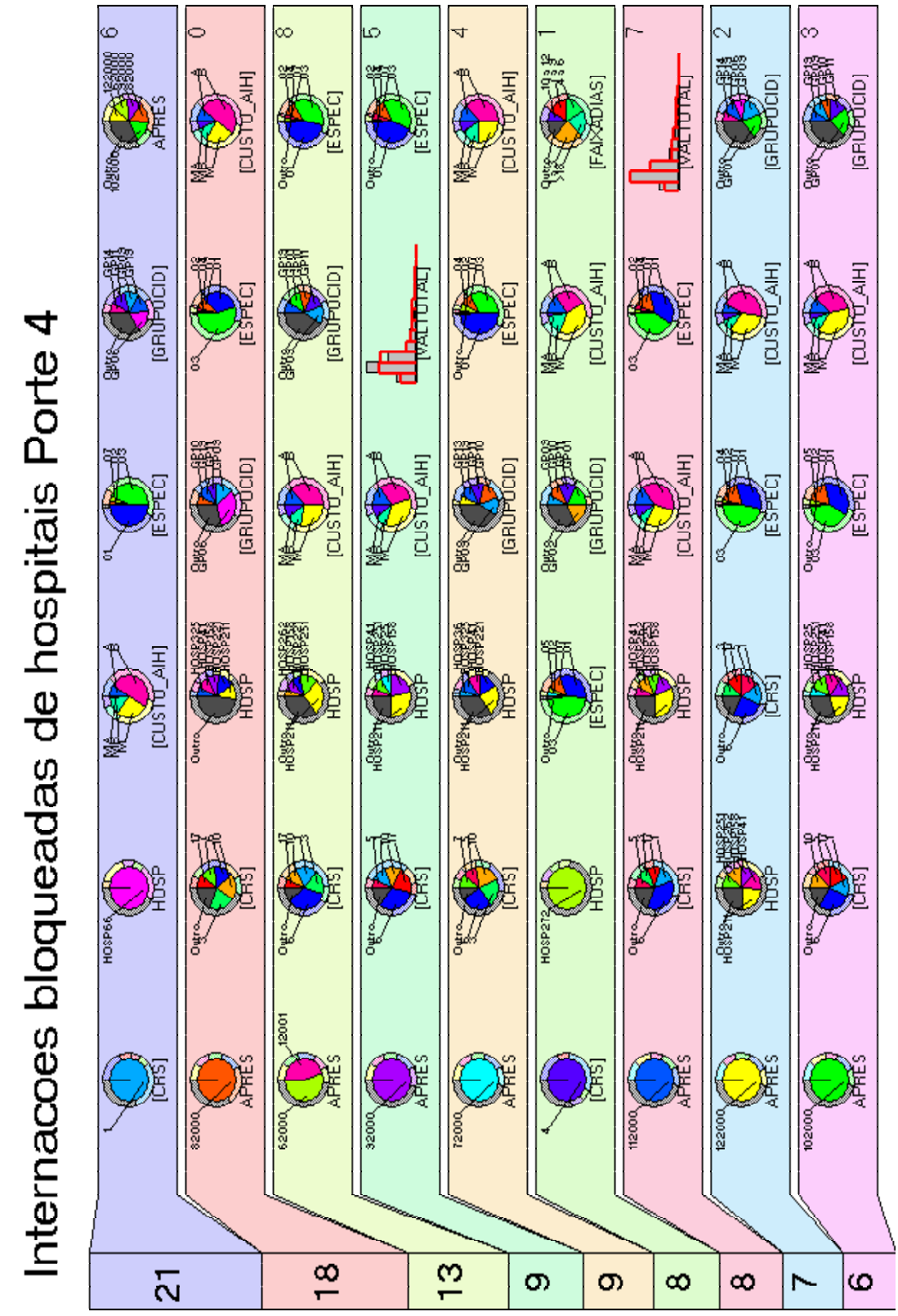


FIGURA C – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs bloqueadas de hospitais Porte 4.

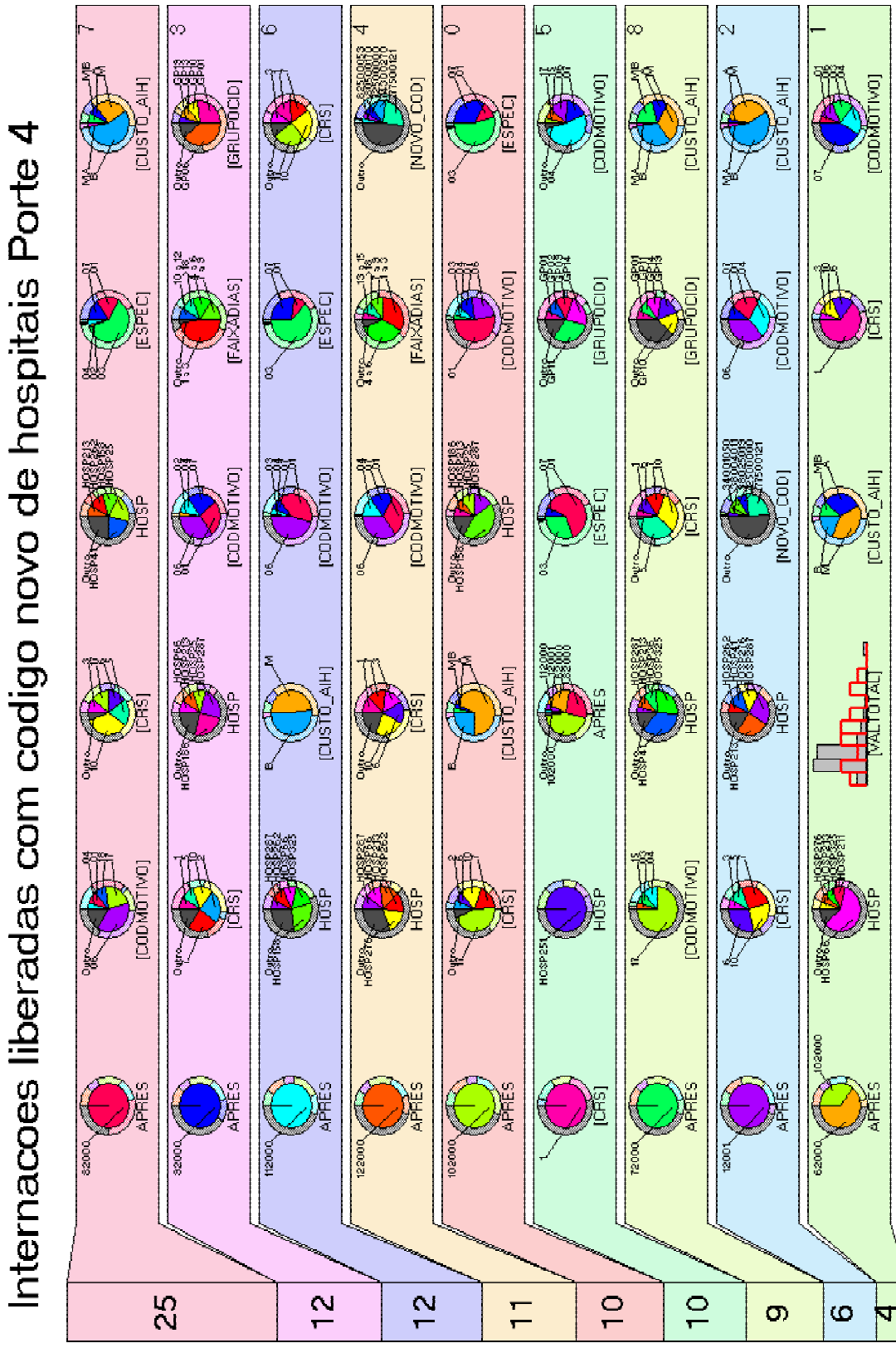


FIGURA D – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs liberadas com código novo de hospitais Porte 4.

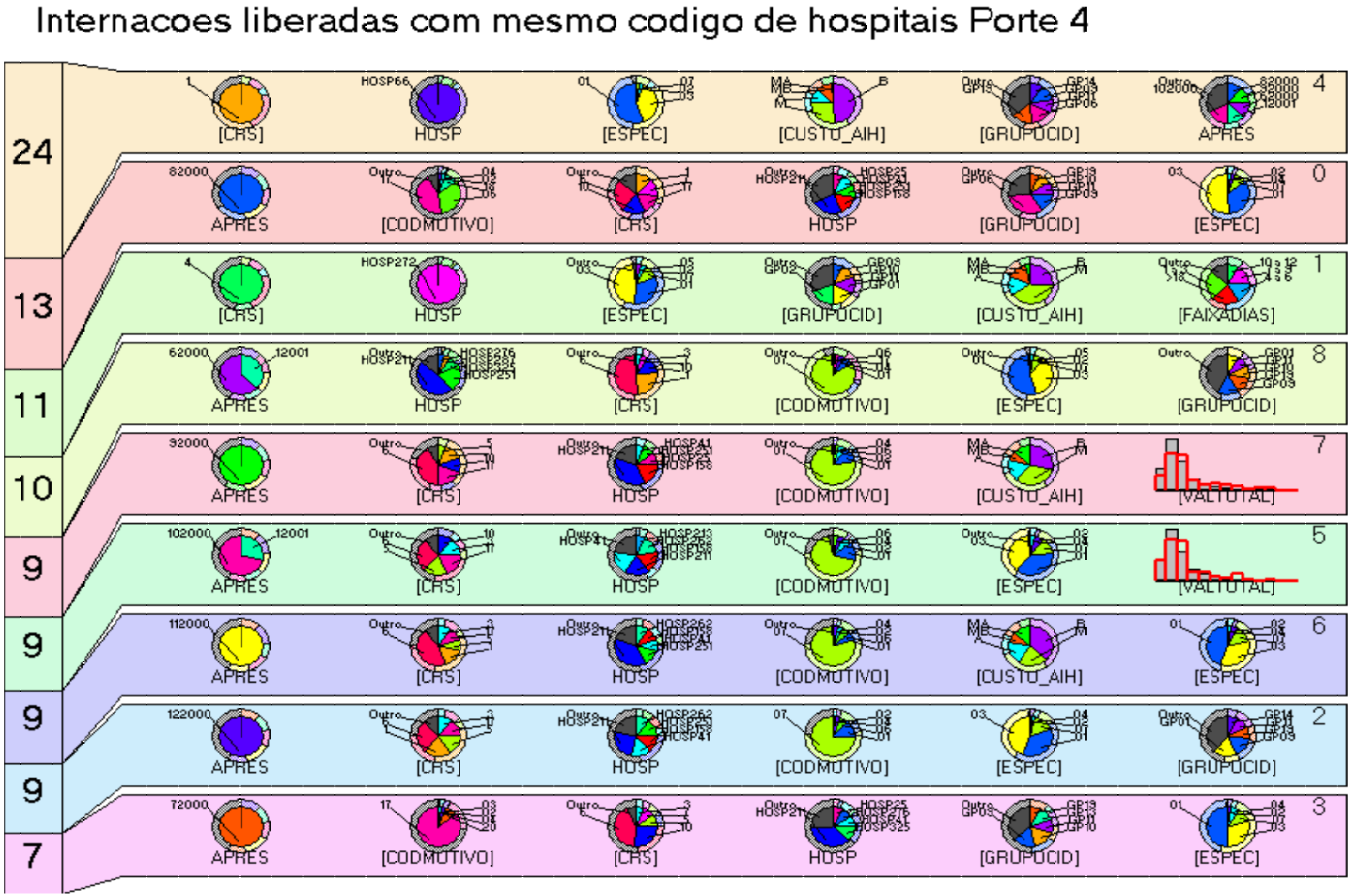


FIGURA E – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs liberadas com mesmo código de hospitais Porte 4.

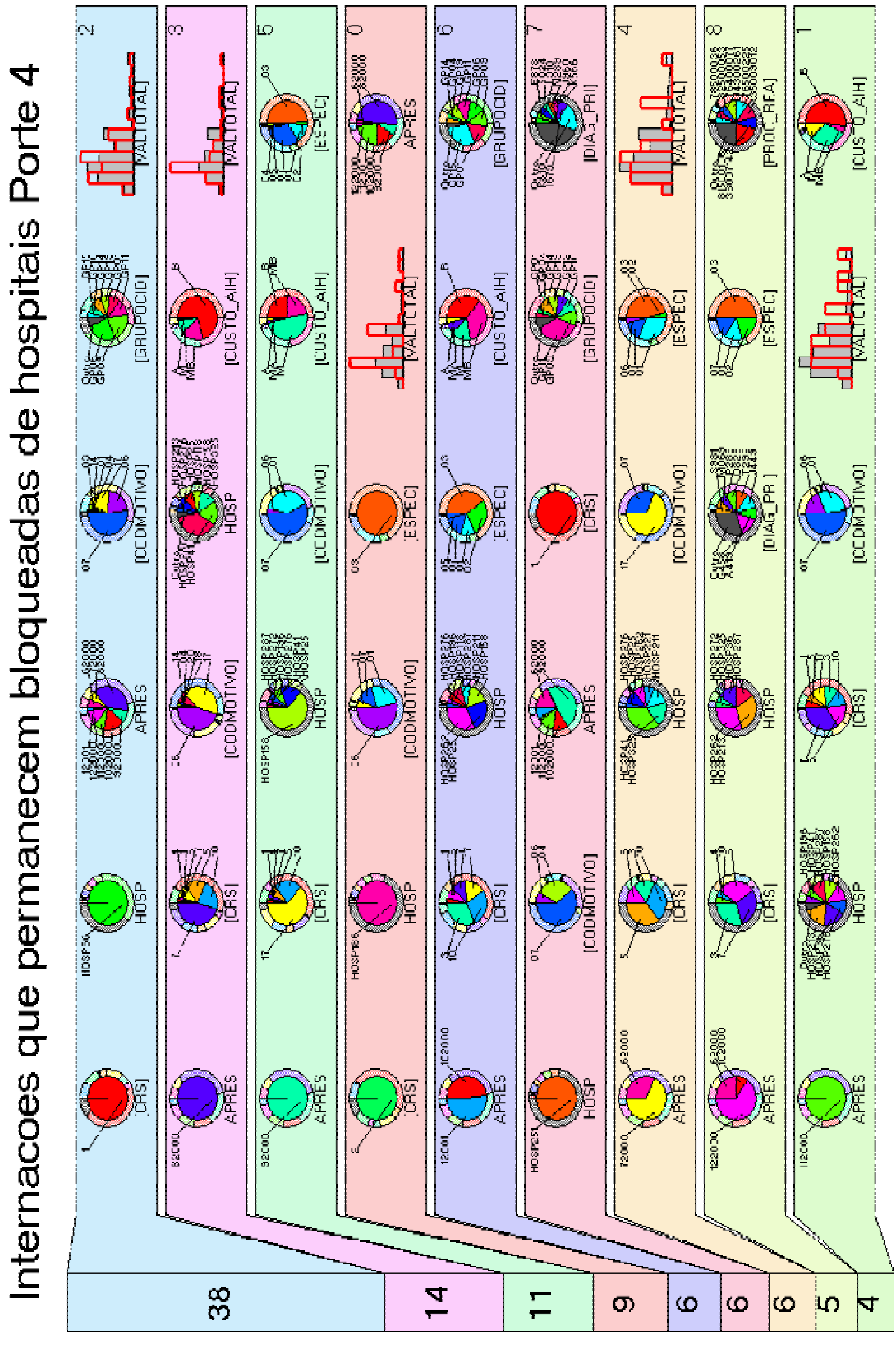


FIGURA F – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs que permanecem bloqueadas de hospitais Porte 4.

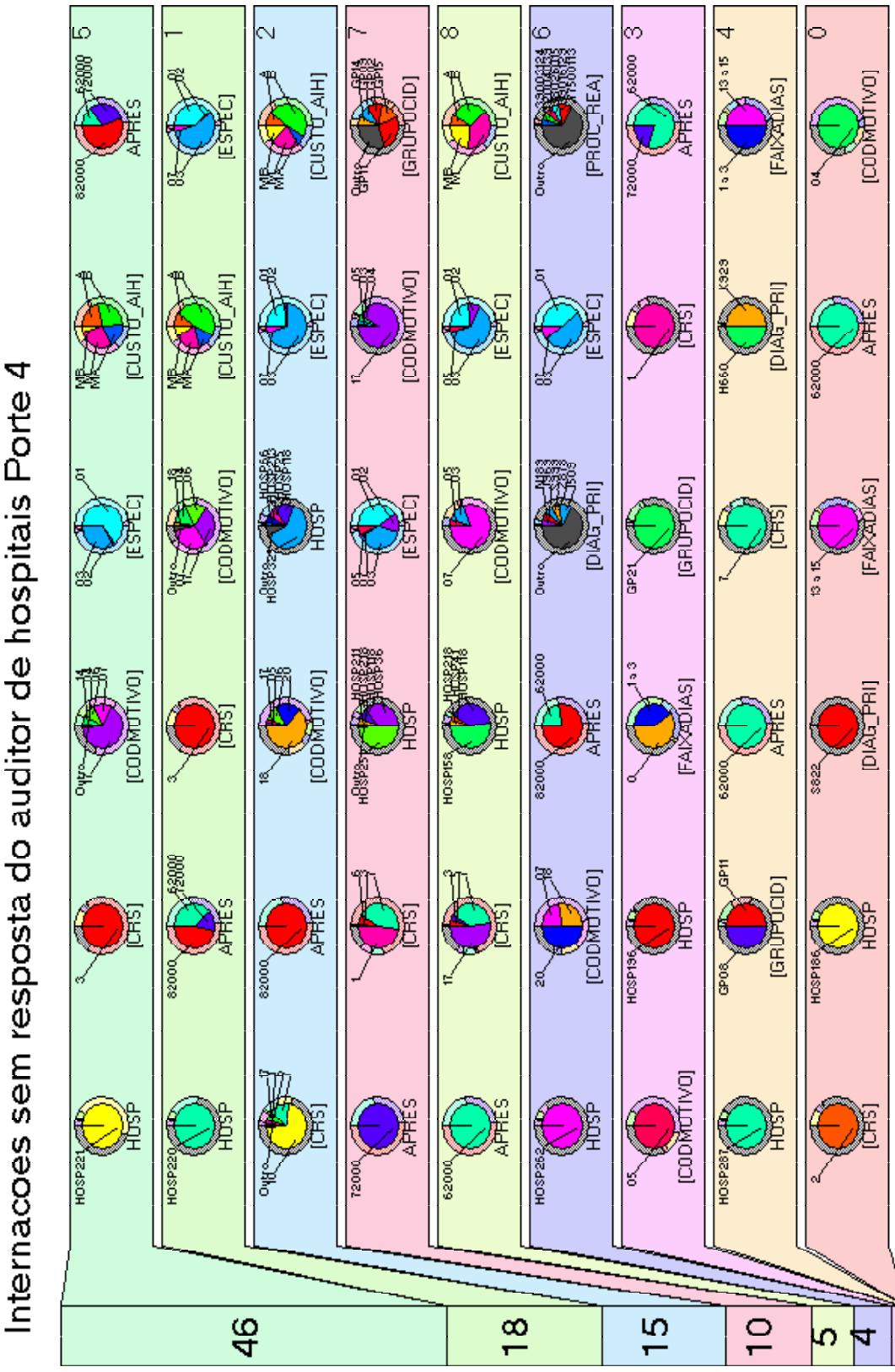


FIGURA G – Modelo de mineração com a pesquisa de agrupamento demográfico sobre os dados das AIHs sem resposta do auditor de hospitais Porte 4.

Bibliografia

[AGR 98] AGRAWAL, Rakesh; GEHRKE, Johannes; GUNOPULOS, Dimitrios; RAGHAVAN, Prabhakar. Automatic Subspace Clustering of High Dimensional Data dor Data Mining Applications. In: ACM SIGMOD INT. CONFERENCE MANAGEMENT OF DATA, 1998. **Proceedings...** Seattle, WA: ACM, 1998. p.94-105.

[ALS 98] ALSABTI, Khaled; RANKA, Sanjay; SINGH, Vineet. An Efficient K-means Clustering Algorithm. In: WORKSHOP ON HIGH PERFORMANCE DATA MINING, 1., 1998. **Proceedings...** Orlando: [S.n.], 1998. Disponível em: <www.cise.ufl.edu/~ranka/>. Acesso em: maio 2001.

[BRA 98] BRADLEY, P. S.; FAYYAD, Usama M. Refining Initial Points for K-means Clustering. In: INT. CONF. ON MACHINE LEARNING, 15., 1998. **Proceedings...** San Francisco: Morgan Kaufmann, 1998. p. 91-99.

[CAB 97] CABENA, Peter; HADJINIAN, Pablo; STADLER, Rolf; VERHEES, Jaap; ZANASI, Alessandro. **Discovering data mining:** from concept to implementation. Upper Saddle River: Prentice-Hall PTR, 1997.

[CAE 2002] CAETANO, Tibério. **Introdução ao Reconhecimento de Padrões**. Disponível em: <http://www.inf.ufrgs.br/~silvia/ipg/slidesRecPadroes.PDF>. Acesso em: dez. 2002.

[CHA 99] CHAPMAN, Pete; KERBER, Randy; CLINTON, Julian; KHABAZA, Thomas; REINARTZ, Thomas, WIRTH, Rüdiger. **The CRISP-DM Process Model**. CRISP-DM consortium, 1999. (Discussion Paper). Disponível em: <http://www.crisp-dm.org>. Acesso em: maio 2001.

[COS 2001] COSTA, Luciano da F.; MONTAGNOLI, Cristian. Máquinas tomam decisões: reconhecimento de padrões e mineração de dados. **Ciência Hoje**, São Paulo, v. 30, n. 176, p. 22-29, out. 2001. Disponível em: http://www.uol.com.br/cienciahoje/chmais/pass/ch176/maquinas.pdf>. Acesso em: fev. 2002.

[DUD 97] DUDA, Richard O. **Feature Selection and Clustering for HCI**. Disponível em: <http://www.engr.sjsu.edu/~knapp/HCIRDFSC/FSC_home.htm>. Acesso em: 02 jul. 2001.

[ENG 2000] ENGEL, Paulo Martins; ALVARES, Luis Otavio; GEYER, Cláudio F. R. et al. **Desenvolvimento de Metodologia para Extração de Conhecimento de Bases de Dados de Saúde do Estado para Avaliação e Planejamento**. Projeto interinstitucional envolvendo UFRGS, UCS e SES e com Apoio ao Desenvolvimento Científico e Tecnológico da Informática da Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul - FAPERGS. Edital jun. 2000.

[ENG 2001] ENGEL, Paulo M. **Sistemas de Informações Inteligentes:** notas de aula. Porto Alegre-RS: PPGC da UFRGS, 2001.

[FAS 99] FASULO, Daniel. **An Analysis of Recent Work on Clustering Algorithms**. Seattle, WA: University of Washington, 1999. (Technical Report 01-03-02).

[FAY 96] FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery: An Overview. In: FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic; UTHURUSAMY, Ramasamy. **Advances in Knowledge Discovery and Data Mining**. Menlo Park: MIT Press, 1996. 611 p. p. 1-34.

[GRA 98] GRABMEIER, Johannes; RUDOLPH, Andreas. **Techniques of cluster algorithms in data mining version 2.0**. Heidelberg: IBM Deutschland Informationssysteme GmbH, 1998. Disponível em: <http://www-3.ibm.com/software/data/iminer/fordata/clusttechn.pdf>. Acesso em: jan. 2002.

[GRI 2002] GRIVET, Marco. **Reconhecimento de Padrões**. Disponível em: <http://www.lncc.br/~biologia/downloads/ReconhecimentoPadroes.pdf> . Acesso em: dez. 2002.

[HAN 2001] HAN, Jiawei; KAMBER, Micheline. **Data mining:** concepts and techniques. San Francisco: Morgan Kaufmann, 2001.

[IBM 99] IBM. **Utilizando o Intelligent Miner for Data**. Versão 6. Release 1. Edição S517-6338-00. [S.l.], 1999. Disponível em: <ftp://ftp.software.ibm.com/software/data/iminer/fordata/docu/Br/idmu0mst.pdf>. Acesso em: out. 2001.

[KAS 97] KASKI, Samuel. **Data exploration using self-organizing maps**. Espoo, 1997. 57 p. (Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series n. 82). Disponível em: <http://www.cis.hut.fi/sami/thesis/node7.html>. Acesso em: maio 2001.

[KNO 2002] KNORR, Edwin. **Outliers and Data Mining:** Finding Exceptios in Data. 2002. Doctor of Philosophy thesis – Department of Computer Science, the University of British Columbia. Disponível em:

<http://www.cs.ubc.ca/grads/resources/thesis/May02/Ed_Knorr.pdf>. Acesso em: fev. 2003.

[LOP 99] LOPES, Carlos H. P. **Classificação de Registros em Bancos de Dados por Evolução de Regras de Associação Utilizando Algoritmos Genéticos**. 1999. Dissertação (Mestrado em Engenharia Elétrica) – Depto. de Engenharia Elétrica, PUC/Rio, Rio de Janeiro.

[MIC 97] MICHAUD, Pierre. Clustering techniques. **Future Generation Computer Systems**, [S.l.], v.13 n.2-3, p.135-147, nov. 1997.

[PEL 99] PELLEG, Dan; MOORE, Andrew. Accelerating Exact *k*-means Algorithms with Geometric Reasoning. In: ACM SIGKDD INT. CONF. ON KNOWLEDGE DISCOVERY AND DATA MINING, 5., 1999. **Proceedings...** San Diego: ACM, 1999. p. 277-281. Disponível em: <www.cs.cmu.edu/~dpelleg/>. Acesso em: maio 2001.

[PYL 99] PYLE, Dorian. **Data preparation for data mining**. San Francisco: Morgan Kaufmann, 1999.

[RIO 2000] RIO GRANDE DO SUL. SECRETARIA DA SAÚDE. **Relatório Anual de 2000**. Porto Alegre, 2000.

[SOF 2002] SOFYAN, Hizir; WERWATZ, Axel. **Analyzing XploRE download profiles with Intelligent Miner**. Disponível em: <jetta.math.uni-augsburg.de/symposium/papers/hizir.ps>. Acesso em: abr. 2002.

[SUS 2001] SISTEMA ÚNICO DE SAÚDE. **SIH/SUS: Sistema de Informações Hospitalares do Sistema Único de Saúde**. Brasília: Ministério da Saúde, SUS, 2001.

[VAS 99] VASCONCELOS, Germano; QUEIROZ, Fausto. **Sistema de aquisição, processamento e reconhecimento de padrões.** 1999. Disponível em: <http://www.cin.ufpe.br/~sapri/ReconhecimentoDePadroes.htm>. Acesso em: dez. 2002.

[VES 97] VESANTO, Juha. **Data Mining Techniques Based on the Self-Organizing Map**. 1997. Master's thesis (Master of Science in Engineering) - Department of Engineering Physics and Mathematics, Helsinki University of Technology, Espoo. Disponível em: <http://www.cis.hut.fi/projects/ide/publications/html/masterJV97/>. Acesso em: maio 2001.

[VES 2000] VESANTO, Juha. **Using SOM in Data Mining**. 2000. Licentiate's thesis (Licentiate of Science in Technology) – Department of Computer Science and Engineering , Helsinki University of Technology, Espoo. Disponível em:

<http://www.cis.hut.fi/projects/ide/publications/fulldetails.html#vesanto2000licentiate>. Acesso em: maio 2001.

[WIT 99] WITTEN, Ian H.; FRANK, Eibe. **Data mining:** practical machine learning tools and techniques with Java implementations. San Francisco: Morgan Kaufmann, 1999.

[YAM 2000] YAMANISHI, Kenji; TAKEUCHI, Jun-ichi; GRAHAM, Williams. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. In: ACM SIGKDD INT. CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 6., 2000. **Proceedings...** Boston: ACM, 2000. p.320-324.