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

Utilização de Técnicas de Mineração de Dados considerando Aspectos Temporais

por

ANELISE DE MACEDO LUCAS

Trabalho de Conclusão submetido à avaliação,
como requisito parcial para a obtenção
de Mestre em Informática

Prof. Dr. Luis Otavio Campos Alvares
Orientador

Profa. Dra. Nara Martini Bigolin
Co-orientadora

Porto Alegre, outubro de 2002.

CIP - CATALOGAÇÃO NA PUBLICAÇÃO

Lucas, Anelise de Macedo

Utilização de Técnicas de Mineração de Dados considerando os Aspectos Temporais / por Anelise de Macedo Lucas. - Porto Alegre: PPGC da UFRGS, 2002.

Trabalho de Conclusão (mestrado) - Universidade Federal do Rio Grande do Sul. Instituto de Informática. Programa de Pós-Graduação em Computação, Porto Alegre, BR - RS, 2002. Orientador: Alvares, Luis Otavio Campos; Co-orientadora: Bigolin, Nara Martini.

1. Mineração de Dados. 2. Descoberta de Conhecimento. 3. Inteligência Artificial. 4. Banco de Dados. 5. Aspectos Temporais. 6. Mineração de Dados Temporais. I. Alvares, Luis Otavio Campos. II. Bigolin, Nara Martini. III. Título.

Agradecimentos

A minha família.

Ao Prof. Dr. Luis Otavio Campos Alvares, pela motivação e orientação.

À Profa. Dra. Nara Martini Bigolin, pelo incentivo e orientação.

Aos amigos e colegas.

À SES.

À UFRGS.

We are drowning in information,
But starving for knowledge
- John Naisbett

Sumário

Lista de Abreviaturas

AIH	Autorização de Internação Hospitalar
AM	Aprendizado de Máquina
API	Application Programming Interface
BD	Banco de Dados
CIAH	Comunicado de Internação Hospitalar
CID	Cadastro Internacional de Doenças
DCBD	Descoberta de Conhecimento em Bancos de Dados
DM	Data Mining
IA	Inteligência Artificial
IM	Intelligent Miner da IBM ©
KDD	Knowledge Discovery in Databases
MD	Mineração de Dados
MDT	Mineração de Dados Temporais
MGL	Modelo da Gramática da Linguagem
OLAP	On Line Analytical Processing
RNA	Redes Neurais Artificiais
SES	Secretaria Estadual da Saúde do Rio Grande do Sul
SGBD	Sistema Gerenciador de Banco de Dados
SINAN	Sistema Nacional de Agravos Notificáveis

Lista de Figuras

Lista de Tabelas

Resumo

Atualmente, o enorme volume de informações armazenadas em bancos de dados de organizações ultrapassa a capacidade dos tradicionais métodos de análise dos dados baseados em consultas, pois eles se tornaram insuficientes para analisar o conteúdo quanto a algum conhecimento implícito e importante na grande massa de dados. A partir disto, a mineração de dados tem-se transformado em um tópico importante de pesquisa, porque provê um conjunto de técnicas e ferramentas capazes de inteligente e automaticamente assistir o ser humano na análise de uma enorme quantidade de dados à procura de conhecimento relevante e que está encoberto pelos demais dados.

O presente trabalho se propõe a estudar e a utilizar a mineração de dados considerando os aspectos temporais. Através de um experimento realizado sobre os dados da Secretaria da Saúde do Estado do Rio Grande do Sul, com a aplicação de uma metodologia para a mineração de dados temporais, foi possível identificar padrões seqüenciais nos dados. Este experimento procurou descobrir padrões seqüenciais de comportamento em internações médicas, objetivando obter modelos de conhecimento dos dados temporais e representá-los na forma de regras temporais. A descoberta destes padrões seqüenciais permitiu comprovar tradicionais comportamentos dos tratamentos médicos efetuados, detectar situações anômalas, bem como, acompanhar a evolução das doenças existentes.

Palavras-Chave: Mineração de Dados. Inteligência Artificial. Bancos de Dados. Aspectos Temporais. Mineração de Dados Temporais.

TITLE: "USE OF DATA MINING TECHNIQUES CONSIDERING TEMPORAL ASPECTS"

Abstract

Currently, the enormous volume of data stored in organizations' databases exceeds the capacity of traditional query-based data analysis methods, since such methods became insufficient to analyze content as for some important and implicit knowledge in great amounts of data. Therefore data mining has become an important research topic, because it provides a set of techniques and tools capable of intelligently and automatically support the human being in the analysis of enormous amounts of data, in search of significant knowledge hidden by other irrelevant data.

The present work intends to study and to use data mining considering temporal aspects. Through an experiment carried out over data from the Health Department of the State of Rio Grande do Sul ("Secretaria da Saúde do Estado do Rio Grande do Sul"), in Brazil, by applying a temporal data mining methodology, it was possible to identify sequential patterns in the data. This experiment meant to discover sequential behavior patterns in medical internments, in order to obtain knowledge models from the temporal data and to represent those in the form of temporal rules. The discovery of such sequential patterns allowed to confirm usual behaviors of the effected medical treatments, to detect anomalous situations, as well as to follow the evolution of existing diseases.

Keywords: Data Mining. Artificial Intelligence. Databases. Temporal Aspects. Temporal Data Mining.

1 Introdução

Historicamente, a comunidade de Banco de Dados desenvolveu a tecnologia dos Sistemas Gerenciadores de Banco de Dados (SGBD) para tratar de forma eficiente e reutilizável volumes de dados que excediam o tamanho da memória física dos computadores. Os avanços desta tecnologia integram linguagens de programação convencional com sistemas de Banco de Dados funcionando com consultas, transações, segurança e distribuição. Os requisitos em termos de especialização, planejamento, translação de dados e custos monetários são grandes para muitas aplicações deste tipo. Um modelo SGBD é uma grande tecnologia de banco com o legado particular de técnicas de perfil de modelagem de dados, linguagens de consulta, otimização e evolução de consultas, visões baseadas em estados, gerenciamento de dados, transações, sistemas distribuídos e sistemas escaláveis objetivando prover rapidez no armazenamento e na manipulação de grande quantidade de informação [BLA 96; SIL 96; STO 96; ÖZS 96].

A partir do crescimento do volume de informações que as corporações manipulam, gera-se a necessidade urgente de técnicas e ferramentas que transformem dados em conhecimento útil de forma inteligente e automática. A solução para esta necessidade das organizações de obterem conhecimento de grandes volumes de dados está na utilização de técnicas de mineração de dados para extrair as informações implícitas existentes nos Bancos de Dados destas organizações.

Assim, a comunidade de Inteligência Artificial, especificamente de aprendizado de máquina, interessou-se pela extração de conhecimento e a aprendizagem a partir de uma quantidade reduzida de informações. É uma sinergia de aprendizado de máquina, análises estatísticas e tecnologias de BD, onde descobre conhecimento através de métodos como *clustering*, classificação e regras associativas entre outras. Estes métodos podem ser vistos em consultas *online* para novas famílias de linguagem de consulta. Entretanto, tais consultas requerem execução em máquinas com algoritmos que aprendem por indução grandes Bancos de Dados [BIG 2000; FAY 97a; PRA 98].

A combinação dessas duas abordagens deu origem a uma nova tecnologia chamada Descoberta de Conhecimento em Banco de Dados (DCBD), em inglês *Knowledge Discovery in Databases* (KDD). A DCBD utiliza técnicas de Mineração de

Dados (MD) ou Extração de Conhecimento[1] para extrair regras e informações implícitas a partir de grandes Bancos de Dados.

A partir de então, verifica-se que a DCBD está sendo empregada, tanto em aplicações comerciais, quanto científicas. Um exemplo desta utilização foi efetuada para a astronomia, na qual uma das aplicações primárias da DCBD foi realizada para a classificação de objetos voadores [FAY 93]. No comércio, aplicações DCBD incluem áreas como: *marketing*, onde identifica padrões de comportamento dos consumidores, encontra características dos consumidores de acordo com a região demográfica e prevê quais consumidores serão atingidos pelas campanhas publicitárias; finanças, onde detecta padrões de fraudes no uso dos cartões de crédito, identifica os consumidores que estão tendendo a mudar a companhia do cartão de crédito, identifica situações de estoque a partir dos dados do mercado; transporte, onde determina a distribuição dos horários entre os vários percursos e analisa padrões de sobrecarga; agricultura, em que prevê riscos que comprometerão uma lavoura visando o financiamento por parte das companhias de seguro agrícola; planos médicos, na qual determina quais procedimentos médicos serão requisitados ao mesmo tempo e identifica comportamentos fraudulentos; saúde, onde caracteriza o comportamento dos pacientes para prever novas consultas médicas e descobrir tratamentos de sucessos para diferentes doenças [FEL 97; RAI 2001].

A motivação para este trabalho está no grande volume de dados que é coletado a cada dia no decorrer do tempo de um evento. Estes dados se constituem em valiosas fontes de informação que podem ser analisadas em uma freqüência de determinados eventos, ou conjuntos de eventos relacionados a particularidades temporais. Observa-se também, que estes tipos de análises podem ser muito úteis para derivar a informação implícita dos dados crus, e também para predizer o comportamento futuro de um processo monitorado. Desta forma, acredita-se que a utilização de técnicas de mineração de dados considerando os aspectos temporais possa descobrir conhecimento de maior qualidade para serem utilizados na previsão de decisões e no enriquecimento de interesses diversos [KOU 2001].

Assim, o objetivo geral deste trabalho é o estudo da mineração de dados temporais. Para tanto, o presente documento está estruturado da seguinte forma:

[1] Neste trabalho, utiliza-se o termo mineração de dados como sinônimo de extração de conhecimento.

Neste capítulo, é apresentado a motivação, a proposta da dissertação e a estrutura deste trabalho de conclusão.

No segundo capítulo, é estudada a etapa de MD dentro do processo de DBCD com suas etapas, justificativa do método e técnica escolhida.

No terceiro capítulo, é descrita a mineração de dados temporais, na qual são considerados os aspectos temporais dos dados em um BD, as diferentes tarefas existentes, as principais técnicas empregadas e os algoritmos utilizados.

No quarto capítulo, é apresentada uma metodologia para efetuar a MDT .

No quinto capítulo, é exibido o domínio de uma aplicação prática e o emprego da técnica de padrão temporal sobre um caso real.

No sexto capítulo, são apresentadas as considerações resultantes com este trabalho e as contribuições futuras.

2 A Descoberta de Conhecimento em Bancos de Dados

A DCBD é o processo não-trivial de identificar padrões implícitos, previamente desconhecidos, e potencialmente utilizáveis dos dados [FRA 91; FAY 97a].

Ela se constitui em um processo, porque envolve múltiplos passos que vão desde a aquisição dos dados, o tratamento, a extração de padrões até a sua interpretação e incorporação em bases de conhecimento.

A DCBD se caracteriza por ser não-trivial, pois apresenta buscas autônomas durante a fase de MD e inferência no decorrer da descoberta de conhecimento.

Assim, a DCBD se propõe a identificar novos padrões implícitos e úteis dos dados através da MD a partir de grandes volumes de dados armazenados.

2.1 O Processo da DCBD

A DCBD abrange todo o processo de descoberta de conhecimento a partir dos dados, incluindo como os dados são armazenados e acessados, como os algoritmos podem ser escaláveis para grandes massas de dados e ainda serem eficientes, como o resultado pode ser interpretado e visualizado, e como a interação homem-máquina pode ser modelada e suportada.

O processo da DCBD compreende a preparação dos dados, a procura por padrões nos dados, a avaliação do conhecimento e o refinamento através da repetição destes passos com interação do usuário.

Após a avaliação dos padrões obtidos, ou do conhecimento descoberto, pode-se realizar a consolidação deste conhecimento, incorporando-o ao sistema para ações futuras. Isto se baseia na premissa de que qualquer exploração de dados é um processo interativo e iterativo, ou seja, pode apresentar vários ciclos em qualquer passo do processo da descoberta, pois maior interesse surge, a medida em que se aprende mais sobre um domínio específico.

O aspecto básico do processo DCBD é o de permitir que os usuários do sistema façam melhor utilização das informações implícitas dos BD de suas corporações, e explorem maiores benefícios, na forma de conhecimento, proporcionado com a riqueza das informações descobertas.

O processo de DCBD identifica o que é conhecimento, através da avaliação e interpretação dos padrões obtidos pelos algoritmos de MD, para poder ser utilizado em diferentes aplicações práticas.

O processo de DCBD utiliza os resultados crus da MD que são: a extração de tendências e padrões dos dados, transformando-os em conhecimento útil. Este conhecimento não é tipicamente a recuperação por técnicas de modelagem, mas é a descoberta de conhecimento através da utilização de técnicas de IA [WRI 2001].

Conforme [FEL 97], a DCBD utiliza algoritmos de MD para extrair o que é conhecimento tendo por base as especificações, medidas e limites atestados nos padrões obtidos pelas análises dos dados. Assim, a DCBD envolve a avaliação e interpretação dos padrões de dados para decidir sobre o que constitui conhecimento, enquanto que a MD é um componente do processo de DCBD que se encarrega da extração e enumeração destes padrões [FAY 97a].

Segundo [ZYT 91], um grande número de hipóteses pode ser extraído de um BD e muitas destas hipóteses podem não ser interessantes para o usuário. Portanto, a verificação do conhecimento útil ou interessante nos padrões descobertos, depende da subjetividade do julgamento humano. Desta forma, pode-se identificar como padrão, o conhecimento que excede os limites de interesse do usuário. A medida global do grau de interesse de um padrão combina validade, novidade, utilidade e simplicidade dos dados [FAY 97b]. Os ajustes adequados dos limites destes interesses podem enfatizar padrões mais precisos ou mais úteis em relação aos demais, o que faz também, com que muitos sistemas definam este grau de interesse indiretamente, através da ordenação dos padrões descobertos.

Assim, a DCBD é um processo que enfatiza a interação humana sobre o total de automação, identificando novos padrões válidos, úteis e simples dos dados. Ela expande a autonomia de descobertas artificiais, orientada pelos resultados práticos, da MD combinando intervenção humana com técnicas automatizadas.

O princípio básico do processo DCBD é mapear um baixo nível de dados em formas mais compactas, abstratas e úteis, denominadas de padrões. A DCBD é uma atividade que enfatiza o descobrimento e compreensão destes padrões que podem ser interpretados como conhecimento útil. Neste contexto, o conhecimento expressa o relacionamento e os padrões entre os elementos de dados [ADR 97].

A seguir, a figura 2.1 exibe os seis passos da DCBD:

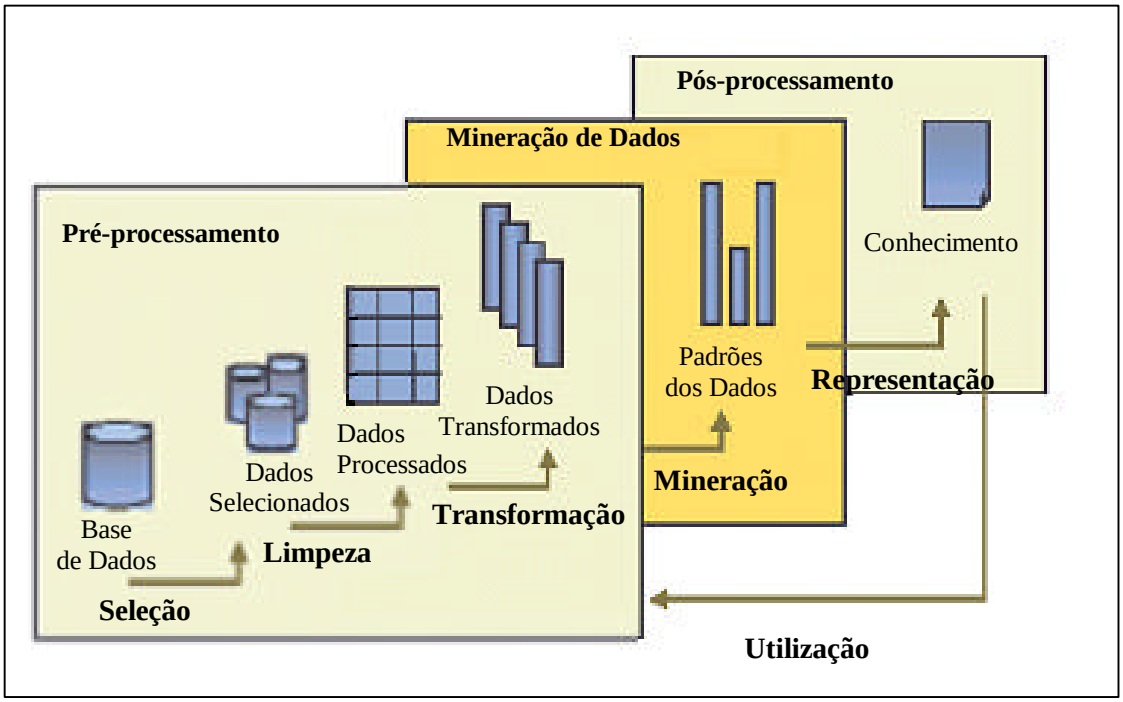

FIGURA 2.1- Processo da DCBD.

Fonte: FAYYAD, 1996. P.10 com adaptações.

O primeiro passo constitui na seleção dos dados. A seleção dos dados necessita do entendimento do problema através da compreensão do domínio da aplicação e do conhecimento prévio relevante. Como o processo de DCBD depende do domínio específico, ou seja, da aplicação em questão, é necessário o conhecimento deste domínio para que se possa extrair informações relevantes como, por exemplo, a definição e o entendimento do que se pretende com a utilização do sistema de DCBD. A partir disto, os conjuntos de dados são selecionados e focalizado o subconjunto de variáveis ou exemplos de dados onde a descoberta será efetuada.

O segundo passo refere-se a limpeza dos dados. Neste passo, são realizadas operações básicas para a remoção de ruídos, decisão das estratégias para manipulação de atributos errôneos ou inexistentes, e tratamento das alterações da informação ao longo do tempo.

O terceiro passo trata da transformação que consiste em simplificar a estrutura complexa dos dados armazenados em BD para obter um formato mais apropriado às técnicas de MD. Este passo objetiva descobrir modelos úteis para representar os dados através da redução ou transformação do número de variáveis existentes.

O quarto passo compõe a escolha de um método de MD como, por exemplo, *clustering*, classificação ou regras associativas entre outros. Estes métodos são apresentados na seção 2.3.1. Após esta escolha, são selecionados os algoritmos de MD para serem utilizados na formulação dos padrões de dados. Como existem diversos algoritmos que podem realizar aproximadamente a mesma tarefa, deve-se escolher qual o mais adequado para a aplicação em questão, pois os resultados irão repercutir em um significativo efeito na qualidade dos padrões extraídos. A partir disto, efetua-se a MD propriamente dita. A MD procura por padrões de conhecimento interessantes. Estes padrões de conhecimento estão presentes em um conjunto de representação como, por exemplo, redes neurais, árvores de decisão, algorítmos genéticos, extensões do algoritmo *Apriori* entre outros que são detalhados na seção 2.3.2. Após, o usuário pode adicionar métodos de MD para corrigir a performance dos passos precedentes.

O quinto passo consiste na representação dos padrões descobertos através de um modelo de conhecimento como, por exemplo, grupos, regras, grafos, gráficos e demais outros que permitem a visualização dos padrões extraídos e dos dados obtidos durante a extração. Este passo envolve a interpretação dos padrões descobertos, possivelmente retornado a algum passo a partir do primeiro para maior iteração.

Finalmente, o sexto passo efetua a utilização do conhecimento. Este passo incorpora o conhecimento para melhorar a sua performance, ou simplesmente documenta e comprova os seus interesses. Este processo inclui também a validação para solução de conflitos através de ações e verificação do conhecimento extraído.

2.2 Resumo dos Seis Passos da DCBD

A partir da descrição do processo de DCBD produziu-se a tabela 2.1 composta pelo significado, facilidade e dificuldade de cada um dos seis passos deste processo.

TABELA 2.1- Tabela resumo dos seis passos do processo DCBD.

PASSO	SIGNIFICADO	FACILIDADE	DIFICULDADE
Seleção	Seleciona os dados relevantes e identifica o objetivo da DCBD.	Utilizar especialistas da área para adquirir o conhecimento do domínio, bem como, linguagens de consulta tradicionais.	Clareza na exposição do domínio, problema e objetivo da descoberta. Identificar informações relevantes.
Limpeza	Efetua a limpeza, a consistência e os testes de validade dos dados.	Aumentar a qualidade dos dados.	As alterações temporais de renomeação dos campos, semântica especial e infidelidade temporal dos dados.
Transformação	Simplifica a estrutura dos dados para a MD.	Utilizar técnicas de generalização e transformação dos dados.	Selecionar modelos úteis para a representação dos dados.
Mineração de Dados	Seleciona um método que atenda ao objetivo da DCBD. Escolhe um algoritmo que implemente o método selecionado.Aplica a MD.	Escolher um método que melhor atenda as necessidades da descoberta. O método sugere o algoritmo a ser escolhido. Encontrar o padrão de conhecimento para os dados.	Adequar o método à solução do problema. Distância entre o algoritmo e o problema apresentado. Proporcionar eficiência na MD.
Representação	Verificação e validação do padrão encontrado.	Retornar aos passos anteriores para otimizar o processo de DCBD.	Interpretação de dados que sofreram manutenções com o decorrer do tempo e que mascararam os resultados e análises deste passo.
Utilização	Incorporação do conhecimento.	Incorporar o conhecimento ao sistema de DCBD.	Resolução de conflitos entre os conhecimentos anteriores e adquiridos.

A seguir, a MD será discutida com maior detalhe, porque ela será o enfoque deste trabalho.

2.3 A Mineração de Dados

A MD é uma etapa do processo de DCBD que consiste em efetuar análises nos dados através da executação de algoritmos que, sob aceitáveis limitações de eficiência computacional, produzem uma enumeração de padrões sobre um conjunto de dados [FAY 97b].

A MD se caracteriza por ser um conjunto de técnicas que envolve métodos matemáticos, algoritmos e heurística para descobrir padrões e regularidades em grandes conjuntos de dados. A MD se preocupa em ajustar modelos ou determinar padrões a partir dos dados observados. Ela pode ser vista como uma forma de selecionar, explorar e modelar grandes conjuntos de dados para detectar padrões de comportamento. Os padrões ajustados representam o conhecimento inferido, o que a torna uma poderosa ferramenta de auxílio à tomada de decisão.

Existem dois tipos de conhecimento extraído através da prática da MD que são: previsão e descrição. A previsão utiliza algumas variáveis ou campos de um BD para prever valores futuros ou variáveis de interesse. Na descrição, os sistemas extraem padrões de representação que descrevem os dados no formato do entendimento humano.

O objetivo da MD a partir de BD é gerar uma organização autônoma de aprendizado que faça uso da informação gerada de forma ótima [ADR 97].

A seguir, são apresentados os métodos da MD.

2.3.1 Os Métodos da MD

As comunidades ligadas à IA, os estatísticos e os físicos desenvolveram um conjunto de métodos lógicos utilizados pela MD. Os métodos da MD permitem que seja possível descobrir uma representação otimizada da estrutura de um BD a partir de um conjunto dos seus dados.

Os métodos da MD utilizam processos indutivos de aprendizado para efetuar a aquisição de conhecimento dos BD. A aprendizagem indutiva se divide em dois tipos: aprendizagem supervisionada e aprendizagem não supervisionada.

A aprendizagem supervisionada consiste em examinar as características de um novo objeto a partir de um conjunto pré-definido para associá-lo à uma classe[2]. Alguns exemplos de métodos da MD que utilizam o tipo de aprendizado supervisionado são: classificação e regras associativas.

A aprendizagem não supervisionada é uma técnica utilizada para se encontrar uma classe a partir de uma ou várias características de um conjunto de objetos. A classe destes objetos não é conhecida inicialmente. Um exemplo de método da MD que utiliza

[2] Classe é um atributo que determina um conjunto de dados em um BD.

o aprendizado não supervisionado é o *clustering*. A seguir, serão estudados os métodos da MD de *clustering*, classificação e regras associativas.

2.3.1.1 *Clustering*

O *Clustering* [ENG 2001; PRA 98] constitui-se em um método descritivo que identifica um conjunto finito de classes ou *clusters* para descrever os dados. Ele é também conhecido como identificação de classes, segmentação, ou agrupamento automático.

O método de *clustering* significa agrupar indivíduos conforme sua semelhança, formando sub-conjuntos de dados que representam grupos. A partir disto, pode-se obter um resumo das características dos indivíduos de cada grupo, ou ainda aplicar alguma outra ferramenta de MD para descrevem regras de conhecimento para cada um dos grupos encontrados.

Neste método, as classes não são conhecidas e os objetos são agrupados pelas suas propriedades. Desta forma, o método de *clustering* explora diferentes alternativas e detecta padrões dos dados, pois não é informado a que classe pertence cada uma das entidades. Após detectados os padrões dos dados, são descritos os seus conceitos. Por isto, o método de *clustering* é considerado como uma forma de aprendizado não-supervisionado. A análise de *clustering* ajuda na construção de agrupamentos significativos de um grande conjunto de objetos, pois decompõe um sistema grande em componentes menores para simplificar a sua estrutura.

O método de *clustering* identifica *clusters*, ou regiões densamente povoadas, de acordo com algumas medidas de distância, em um conjunto multidimensional e grande de dados. Entretanto, verifica-se que o espaço dos dados não é uniformemente ocupado pelos pontos de dados. O *clustering* identifica agrupamentos escassos e aglomerados através de modelos de distribuição representados por conjuntos de dados. Os agrupamentos obtidos podem ser mutuamente exclusivos e exaustivos ou consistirem de uma representação rica, tal como, uma hierarquia ou sobreposição de agrupamentos.

O método de *clustering* pode ser realizado por diversas técnicas, geralmente baseadas em redes neurais que são descritas na seção 2.3.2. A figura 2.2 apresenta um exemplo do método *clustering* empregado sobre os pacientes de uma base em potencial. O objetivo seria descobrir os pacientes mais indicados a desenvolverem um tratamento preventivo para uma doença futura.

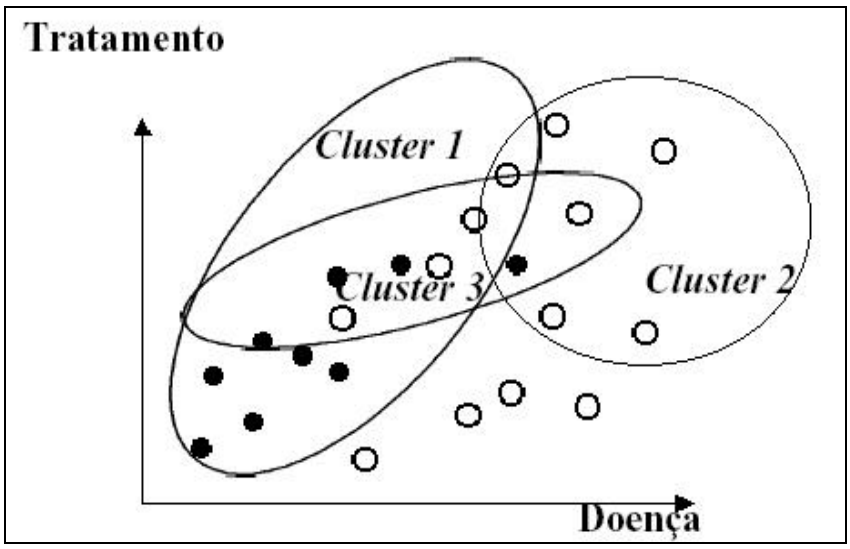

FIGURA 2.2 - *Cluster* sobre os pacientes de um BD.

A utilização do método de *clustering* possibilita descobrir quais agrupamentos ou segmentos gerados apresenta as seguintes características:

Cluster 1: 3,5% do total de pacientes em potencial, 90% dos membros do segmento apresentam características para sofrerem um infarto do miocárdio, 70% são do sexo masculino.

Cluster 2: 4,2% dos pacientes em potencial, 60% apresentam características para sofrerem um infarto do miocárdio, de ambos os sexos.

Cluster 3: 7,2% dos pacientes em potencial, 30% apresentam características para sofrerem um infarto do miocárdio, 60% são do sexo feminino.

A partir das características dos *cluster* gerados, um ou mais são selecionados para utilizar o tratamento, da mesma forma, ou de forma diferenciada. Os resultados de pesquisa devem ser monitorados para que se saiba qual o segmento em que o tratamento deve ser efetivamente dirigido, realimentando o processo.

A vantagem deste método é a facilidade de utilização de valores numéricos, textuais ou categorias. A desvantagem reside na dificuldade de interpretação dos resultados.

2.3.1.2 Classificação

A classificação [HAN 2001] consiste em descobrir conhecimento capaz de prever situações ou eventos futuros. Para tanto, uma característica dos dados

apresentados é definida como "objetivo". A classificação se enquadra no aprendizado supervisionado pois executa uma função que mapeia ou classifica um item em uma ou mais classes de um dado conjunto. Ela utiliza este conjunto de exemplos pré-classificados para desenvolver um modelo que possa classificar uma população de registros. A utilização do método de classificação inicia-se com este conjunto de treinamento formado por exemplos pré-classificados. Este conjunto de treinamento pode incluir registros completos das atividades determinadas sobre o BD, do qual se procede a classificação de registro a registro. A classificação é considerado o método de MD mais freqüentemente empregado.

A vantagem deste método é a facilidade da utilização de classes pré-determinadas. A desvantagem reside na dificuldade em se obter um conjunto de exemplos bastante diversificado de modo a não produzir resultados tendenciosos.

A figura 2.3 apresenta uma ilustração do método de classificação utilizado para descobrir regras que descrevem pacientes com probabilidade de receberem um leito para internação hospitalar. O eixo vertical corresponde aos hospitais da rede pública com leitos disponíveis, e o horizontal, os pedidos de internação. Os pacientes estão representados com o símbolo ✍, para os pacientes com enfermidade leve e o símbolo ✍, para os pacientes com enfermidade grave.

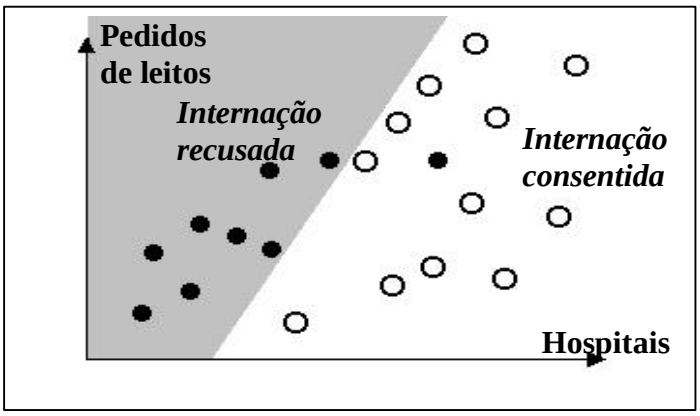

FIGURA 2.3 - Classificação para leitos de pacientes.

2.3.1.3 Regras Associativas

As regras associativas [HAN 2001] são utilizadas para encontrar uma descrição compacta para um subconjunto de dados. As regras associativas consistem em encontrar um modelo que descreve as dependências significativas entre as variáveis. As regras associativas utilizam dois tipos de variáveis: qualitativas e quantitativas. As variáveis

qualitativas, também, chamadas estruturais, categóricas ou nominais, especificam a dependência local das variáveis e em que ordem. As variáveis quantitativas especificam as dependências em uma escala numérica.

O exemplo de regras associativas da figura 2.4 apresenta a propriedade da dependência funcional onde o estado de aspecto geral de prostação, ocasiona a perda de tonus muscular. A dependência funcional é a formalização de uma regra a partir de suas ocorrências em um determinado evento.

SE aspecto geral de prostação ENTÃO perda de tonus muscular

FIGURA 2.4 - Exemplo de regras associativas.

Em uma regra de associação onde "Se X então Y", os fatores de suporte e confiança são expressos conforme a figura 2.5:

$$\text{Suporte} = \frac{\text{Número de registros com X e Y}}{\text{Número total de registros}} \qquad \text{Confiança} = \frac{\text{Número de registros com X e Y}}{\text{Número de registros com X}}$$

FIGURA 2.5 - Fatores de suporte e confiança das regras associativas.

O fator de confiança determina o percentual de acerto da mesma em ocorrências semelhantes, como em sentenças do tipo em que se pode afirmar que “80% de todos os quadros clínicos observados nas quais malária é constatada, também apresentam desnutrição”, e isto é comprovado com 95% de acerto. Desta maneira, pode-se descobrir regras associativas considerando-se todas as regras com suporte e grau de certeza mínimo especificado.

Algumas ferramentas de regras associativas podem descobrir regras mais complexas ou com mais elementos no lado SE conforme a figura 2.6.

SE aspecto geral de prostação E perda de tonus muscular E tonalidade amarela da pele ENTÃO avaliar disfunção hepática

FIGURA 2.6 - Exemplo de regra de associação complexa.

A vantagem deste método é a facilidade de se encontrar uma descrição compacta para um subconjunto de dados. As regras associativas são úteis quando se necessita de um primeiro conhecimento, ou uma idéia do quê se procura. A desvantagem reside na

dificuldade da derivação de regras mais sofisticadas por envolverem as relações funcionais entre muitas variáveis e grandes BD.

A tabela 2.2 abaixo, apresenta os principais métodos de MD, a representação do conhecimento, a vantagem e desvantagem e a comparação com os demais métodos.

TABELA 2.2 - Tabela comparativa dos métodos de MD.

Método	**Representação do conhecimento**	**Vantagem**	**Desvantagem**	**Comparação com os demais métodos**
Clustering	*Cluster*	Valores numéricos, textuais e categorizados	Interpretar resultados	Aprendizado não supervisionado
Classificação	Regras de classificação	Classes pré-determinadas	Exemplo diversificado	Mais empregado dos métodos
Regras Associativas	Regras de associação	Descrição compacta dos dados	Derivação de regras sofisticadas em relação a um grande número de variáveis	Baseado em dependências funcionais.

2.3.2 As Técnicas de MD

Existem muitas abordagens que têm sido empregadas com o intuito de se descobrir conhecimento em BD. A seguir, são apresentadas as técnicas de MD de redes neurais, árvores de decisão, algoritmo genético e *Apriori*.

2.3.2.1 Redes Neurais

As Redes Neurais Artificiais (RNA) [ENG 2001] são representadas como grafos nos quais cada nó é denominado um elemento de processamento. Cada arco deste grafo pode representar: a entrada de um sinal em um dos elementos de processamento a partir do meio externo, a comunicação de um sinal entre dois elementos de processamento, ou a saída de um sinal de um elemento de processamento para o meio externo. Uma RNA interage com o meio externo recebendo estímulos e gerando respostas a estes estímulos. A comunicação dentro de um elemento de processamento é feita a partir dos valores das entradas e de seus pesos associados às conexões (sinapses) dos arcos. Por isto, as RNA também são chamadas de conexionistas.

Uma das principais características das RNA é a capacidade de aprendizagem automática. O processo de aprendizagem, chamado de treinamento da rede, inicia com a separação dos dados existentes sobre o problema em dois conjuntos: treinamento e

atuação. O treinamento é utilizado para treinar a rede, isto é, ajustar os seus pesos. No treinamento os valores das sinapses estão sendo ajustados a partir dos efeitos das entradas. A atuação é utilizada para validação do conhecimento descoberto. Na atuação, a RNA é aplicada a novos casos. Este parâmetro refere-se à existência ou não de sinais de saída pré-definidos para a rede, e classificam-se em supervisionado ou não-supervisionado (auto-aprendizado). Os algoritmos de treinamento dependem da topologia da rede que pode assumir mais de uma forma com camadas escondidas. Por esta razão, apesar das RNA serem ferramentas de alto poder de modelagem, são relativamente mais complexas comparada as árvores de decisão [FAY 97a].

Existem diversas formas de RNA que executam diferentes estratégias de aprendizado, porém as três que se sobressaem são: perceptrons, redes de propagação e mapas de organização de Kohonen.

A RNA perceptron consiste em uma rede simples de três níveis com unidades de entrada chamadas de foto-receptores, unidades intermediárias chamadas de associadores, e unidades de saída chamadas respondedores.

As redes de propagação apresentam camadas ocultas entre as redes e atribuem pesos randomicamente às sinapses para caso exista uma diferença entre a solução correta e a obtida, os pesos dos nodos individuais e sinapses ajustam a rede.

O mapa de organização de Kohonen é composto por uma coleção de neurônios ou unidades que contém um fator que está relacionado a um espaço da estrutura investigada onde são atribuídos vetores randomicamente a cada unidade que são incrementalmente ajustados de modo a chegar a melhor cobertura para o espaço analisado [ADR 97].

A vantagem principal da utilização de RNA é a versatilidade e o resultado satisfatório em áreas complexas com entradas incompletas ou imprecisas, pois estas são mais dificilmente tratadas com a IA tradicinal baseada na lógica, representação explícita do conhecimento ou busca de grafos entre outras. As RNA têm excelente desempenho em problemas de classificação e reconhecimento de padrões como para o reconhecimento de caracteres, de imagens, de voz, na identificação de impressões digitais, análise de crédito e diagnóstico médico.

As desvantagens existentes dizem respeito a solução final que depende das condições iniciais estabelecidas na rede pois os resultados dependem dos valores

aprendidos. Outra desvantagem consiste na apresentação de uma caixa preta que não contém informação que justifique as conclusões chegadas. As RNA não podem provar uma teoria a partir do que aprenderam. Elas são simples caixas pretas que produzem respostas mas não demonstram claramente o desenvolvimento de como chegaram aos resultados [ADR 97].

2.3.2.2 Árvore de decisão

A árvore de decisão, também chamada de indução de árvores de decisão, constitui-se na técnica mais utilizada pelo método da classificação pois organiza classes de objetos [ADR 97].

As árvores de decisão utilizam a estratégia do dividir para conquistar. Esta estratégia consiste em um problema complexo ser dividido em um conjunto simplificado de sub-problemas. Recursivamente, aplica-se a mesma estratégia em cada um dos sub-problemas. A solução para cada um dos sub-problemas é combinada até se obter uma solução para o problema original. Para isto, esta técnica consiste em construir uma árvore, na qual, cada nodo "não-folha" representa um atributo, os arcos expressam assertivas sobre os atributos, e as "folhas" representam as classes associadas a um determinado percurso da árvore. A partir de cada nodo da árvore, aparecem decisões, baseadas em atributos do BD. Cada nodo representa um valor daquele atributo. Conforme a figura 2.7, um objetivo é classificado da raiz até as "folhas" enquanto as suas características atenderem as ligações. Estes algoritmos dividem de forma recursiva os dados, gerando uma árvore, até o ponto em que cada "folha" contenha dados de uma única classe ou principalmente, de uma classe.

As vantagens desta técnica está na capacidade de dividir um problema grande em sub-problemas. Assim, cada sub-problema está adequado a diferentes funções. Também, as árvores de decisão apresentam uma forma de representação simples, gerando modelos de inferência adequados para a compreensão do usuário. As árvores de decisão são boas para classificar ou fazer predições dos dados através do atributo mais informativo. Algoritmos de indução de árvores de decisão trabalham bem com grandes conjuntos de dados e são adequadas tanto para dados qualitativos quanto quantitativos. Outra vantagem, é que elas possibilitam verdadeiros *insights* para a natureza do

processo de decisão. Adicionalmente, os resultados das árvores de decisão produzem algoritmos que podem ser utilizados diretamente pelos usuários [ADR 97].

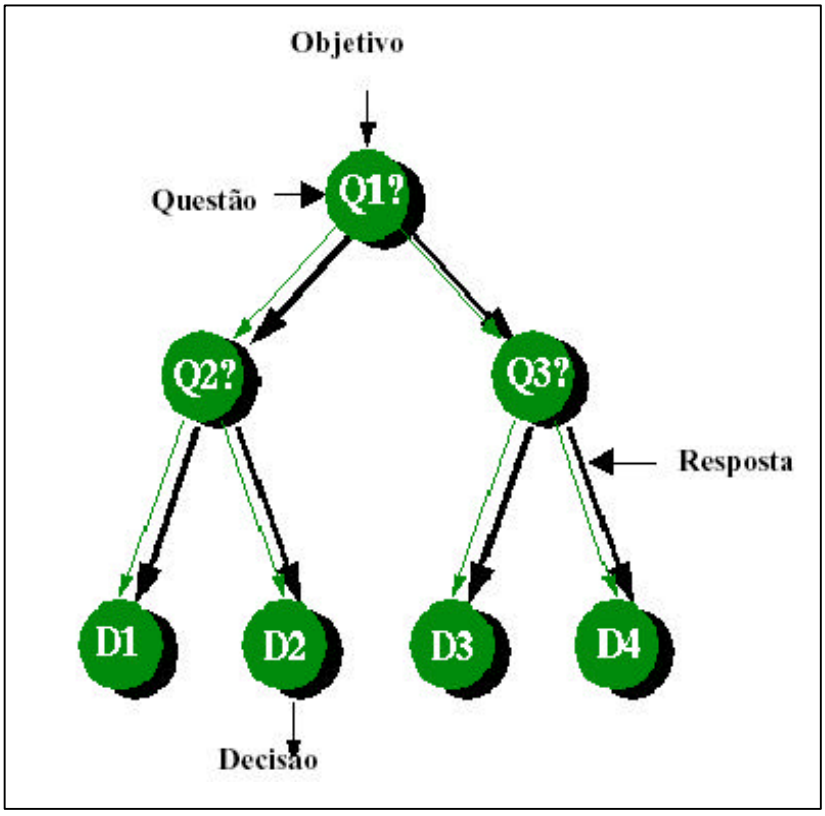

FIGURA 2.7 - Estrutura de uma árvore de decisão.

As desvantagens desta técnica estão na instabilidade para pequenas variações no conjunto de treinamento. Uma das limitações para este tipo de técnica é a difícil visualização em conjuntos de exemplos com muitos atributos, com muitos valores possíveis, o que é comum em BD reais. Outra desvantagem existente para a representação particular de árvores e regras podem resultar em restrições de acordo com a forma ou do modelo funcional. A figura 2.8 apresenta um exemplo de árvore de decisão aplicado a um domínio médico.

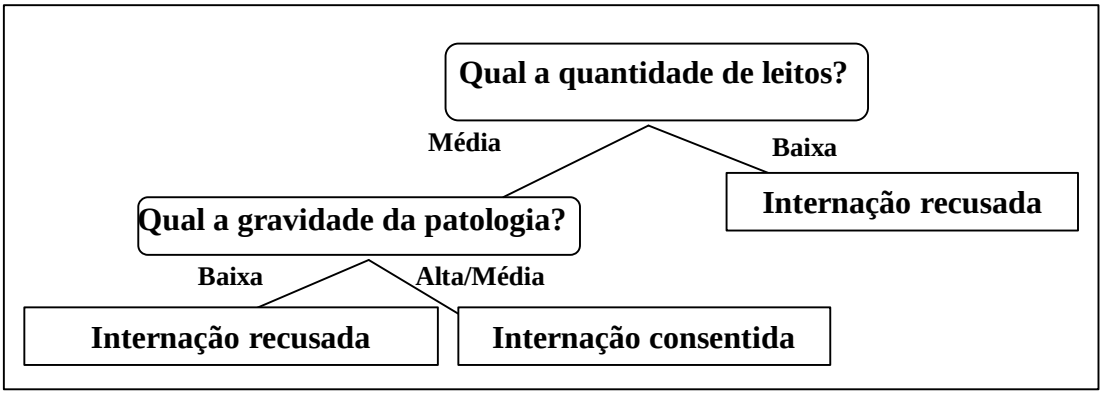

FIGURA 2.8 - Árvore de decisão para concessão de leitos hospitalares.

Existem numerosos algoritmos baseados nesta técnica capazes de classificar corretamente um mesmo conjunto de dados. O algoritmo CART [BRE 84] baseia-se em um arquivo de treinamento com dados previamente rotulados. Em cada nodo, os casos são separados em função de apenas um atributo. O atributo a ser testado em um nodo é escolhido como aquele que gera grupos com a predominância de uma única classe. O algoritmo C4.5 [QUI 93] produz árvores com número de ramos variável, no qual, cada valor de um dado categórico, gera um ramo. O algoritmo ID3 baseia-se em uma árvore de decisão em que, cada nodo, deve estar associado ao atributo mais informativo entre os atributos ainda não considerados no caminho a partir da raiz.

O caminho da raiz até uma folha pode ser expresso diretamente como uma regra, mas caso houver muitas folhas, ou a árvore de decisão for muito profunda, o conjunto resultante de regras pode não ser compreensível. Desta forma, a partir da árvore gerada pode-se extrair regras, e algumas das regras geradas a partir da árvore anterior estão exibidos na figura 2.9.

SE quantidade de leitos média E gravidade da patologia baixa
ENTÃO internação recusada

SE quantidade de leitos média E gravidade da patologia média
ENTÃO internação consentida

SE quantidade de leitos média E gravidade da patologia alta
ENTÃO internação consentida

SE quantidade de leitos baixa
ENTÂO internação recusada

FIGURA 2.9 - Exemplo das regras de conhecimento obtidas a partir da árvore de decisão da figura 2.8.

Uma solução existente para os problemas das árvores de decisão é através da sua poda através da retiranda de ramos que fornecem pouco poder de previsão por folha. Isto é feito analisando-se a freqüência de casos representados pelo ramo e a taxa de erro que incorre quando poda-se este ramo. A taxa de erro de uma folha representa a razão entre o número de casos com classificação errada e o número de casos classificados corretamente pelo ramo. Neste caso, inicia-se pelo fundo da árvore e examina-se cada sub-árvore. Se a substituição desta árvore por uma folha ou pelo ramo mais freqüente ocasionar uma taxa de erro pequena, então efetua-se a poda [ENG 2001].

2.3.2.3 Algoritmo Genético

O algoritmo genético objetiva um desempenho similar à escolha da adaptação da natureza a um padrão artificial. A principal utilização dos algoritmos genéticos está em encontrar os dados relevantes dentro de um grande conjunto de dados.

Os algoritmos genéticos utilizam indivíduos e populações. Cada indivíduo contém algumas informações genéticas codificadas na forma de genes, representados com valores numéricos que são característicos de cada indivíduo. Um conjunto de indivíduos forma uma população. Uma geração é um estado particular de uma população.

O esquema geral de um algoritmo genético baseia-se na representação de um problema através de um conjunto de indivíduos que são soluções potenciais para o problema em questão. Através de processos de seleção, reprodução e mutação, obtêm-se uma nova geração de indivíduos. Após um certo número de gerações espera-se convergir para uma geração de elite que corresponda a uma solução ótima ou quase ótima para o problema [ALV 99; ZHO 99].

O estudo da adaptação envolve o estudo do sistema adaptável a seu ambiente. Em termos gerais, é um estudo de como os sistemas podem gerar procedimentos permitindo ajustes eficientes em seus ambientes. Se a adaptabilidade não deve ser restringida arbitrariamente ao conjunto de saída, o sistema para se adaptar, deve poder gerar todo o método ou o procedimento capaz de uma adaptação eficaz na definição do problema. Isto está baseado na seleção diferencial de apredizagem supervisionada. Esta é a questão do "mais bem sucedido" que é o predominante em termos de abilidade em produzir soluções para os problemas.

A seguir, são apresentas as desvantagens e vantagens da utilização dos algoritmos genéticos.

Uma desvantagem está na grande produção de indivíduos existentes ou do número de avaliações necessárias, pois de acordo com [QUI 93], um classificador genético necessita de um número muito maior de exemplos de treinamento para alcançar os mesmos resultados das árvores de decisão. Desta forma, os algoritmos genéticos são melhores utilizados em pequenos BD, onde nenhum conhecimento do domínio esteja disponível.

Outra desvantagem existente é que todo o processo tem um pequeno propósito, pois a evolução das espécies dependem de fatores de chances. As espécies podem se envolver em mutações randômicas de genes, mas as chances de que a mutação realmente provoque uma alteração significativa é bastante pequena.

As vantagens da seleção natural são solidez e paralelismo herdado, e a certeza de que se uma solução existe, um algoritmo genético irá encontrá-la.

Entretanto, a grande vantagem prática existente está em fornecer soluções próximas à solução ótima, mesmo ignorando outros métodos que solucionem diretamente o problema. O algoritmo genético não exige nenhum conhecimento sobre a maneira de resolver o problema, pois somente, é necessário avaliar a qualidade de uma solução [ALV 99]. Desta maneira, esta técnica só oferecerá vantagens quando nenhum conhecimento do domínio estiver disponível, pois o conhecimento do domínio pode ser incorporado aos algoritmos genéticos, modificando operadores, escolhendo uma população particular inicial ou modificando a função de qualidade.

Portanto, os algoritmos genéticos permitem a obtenção de soluções para problemas que não têm um método de solução descrito precisamente. Os algoritmos genéticos servem também para problemas cuja solução exata é muito difícil de ser encontrada em um tempo razoável como, por exemplo, quando restrições múltiplas e complexas existentes devem ser satisfeitas simultaneamente.

2.3.2.4 Algoritmo *Apriori*

A técnica utilizada pelo algoritmo *Apriori* para encontrar padrões freqüentes de comportamento em um BD é apresentado na figura 2.10.

```
Procedure Apriori()
    begin
        L1 := {Conjunto de elementos freqüentes}
            for (k:=2; Sk-1 ? ? ; k++) do {
                Ck := apriori_gen(Lk-1) ; // Nova candidata
                    forall transações t ? D do {
                        Ct := subconjunto(Ck,t);
                            forall candidata c ? Ck do
                                c.count++;
                            }
                    Lk :={c ? Ck | c.count ? sup_min };
            }
        Seqüência := ? k Lk;
    end
```

FIGURA 2.10 - Algoritmo *Apriori*.

Este algoritmo se baseia no suporte mínimo da uniformidade apresentado pelos dados. O suporte mínimo da uniformidade é o percentual mínimo exigido para que os dados apresentem um padrão freqüente de comportamento.

O algoritmo *Apriori* percorre o BD para gerar um conjunto de elementos freqüentes a constituirem-se candidatos a uma seqüência de elementos. Este algoritmo verifica se o conjunto de elementos candidatos satisfazem o suporte mínimo estabelecido para a seqüência.

Na primeira passada dos dados pelo algoritmo, o suporte para cada elemento individual é obtido. Os elementos que satisfazem este suporte são selecionados como elementos freqüentes. Na segunda passada, os elementos candidatos à seqüência são obtidos através da junção dos conjuntos de elementos mais freqüentes encontrados. O algoritmo prossegue até que o conjunto de elementos restante seja um conjunto vazio.

Verifica-se que o algoritmo *Apriori* apresenta a propriedade de que se um conjunto de elementos é freqüente, então todos os seus subconjuntos também são freqüentes [AGR 95b]. A partir disto, observa-se que o algoritmo *Apriori* pode ser utilizado para descobrir diversos tipos de padrões de conhecimento como, por exemplo, padrões que representem as relações de associação, correlação, causalidade, seqüência, periodicidade, restrição, classificação e demais padrões emergentes dos BD.

Isto é comprovado, através de várias extensões que foram desenvolvidas para adaptar o algoritmo *Apriori* aos diversos métodos de MD existentes. Estas extensões foram denominadas de algoritmos *AprioriLike* como, por exemplo, regras de associação cíclicas, episódios, *icebergs* entre outros descritos em [ÖZD 98; MAN 95; HUG 2000].

Entretanto, estes algoritmos são deficientes para efetuar a descoberta de seqüências através de um grande número de atributos com diferentes níveis de suporte que são apresentados por algumas seqüências, e também por ocasionarem a formação de gargalos na geração de elementos candidatos a seqüência. Isto é devido ao crescimento do número de candidatas à seqüência que causam o aumento do tempo de execução do algoritmo. Este número elevado de geração de candidatas à seqüência também provocam uma drástica deteriorização da performance do algoritmo em função do *swapping* entre a memória e o disco quando as seqüências candidatas são lidas do disco para cada transação.

Uma solução existente para reduzir a geração de elementos candidatos a seqüência chama-se *Adaptative Apriori*. Esta estratégia, utiliza o suporte da restrição

para especificar o suporte mínimo mais indicado para o conjunto de elementos candidatos à seqüência. O suporte da restrição especifica os requisitos exigidos para um conjunto de elementos freqüentes atender ao suporte mínimo. Assim, somente os conjuntos de elementos com suporte específico serão gerados [HAN 2000].

A solução adotada pelo padrão de crescimento freqüente [PEI 2000] procura por seqüências sem a geração de conjuntos de itens candidatos. Esta solução elimina ou reduz substancialmente o número de conjuntos candidatos a serem gerados durante a execução do algoritmo para a descoberta da seqüência, diminui o volume de dados do BD e aumenta a performance do algoritmo.

O algoritmo *Partition* [SAV 95, PLA 2001] é outra estratégia existente que segue com o mesmo objetivo de diminuir o número de leituras da totalidade do BD. Ela utiliza a filosofia de dividir para conquistar e efetua a descoberta de seqüências por partição do BD. Esta estratégia é formalizada da seguinde maneira: se um conjunto F é freqüente em relação a totalidade da BD, chamada freqüência global, então F é freqüente em relação a uma partição chamada de freqüência local, e possui suporte maior ou igual ao mínimo dentro desta partição. Este algoritmo divide-se em duas fases: na primeira, conforme a estratégia *Apriori*, são gerados os conjuntos freqüentes locais candidatos em um único acesso a totalidade da BD; na segunda fase, todas as transações do BD são percorridas para verificar quais freqüentes locais ou candidatos globais são também, freqüentes globais. Entretanto, dependendo do número de candidatos gerados, os freqüentes locais que não serão candidados globais, isto pode comprometer o desempenho deste algoritmo.

O algoritmo *AprioriAll* [AGR 94b] procura por grandes conjuntos de elementos. Na primeira passada dos dados pelo algoritmo, são geradas as candidatas a grandes seqüências e, em uma segunda passada, são medidos os níveis de suporte. Ao final do algoritmo, o suporte das candidatas é utilizado para determinar as seqüências máximas[3] que atendam ao suporte mínimo.

Os algoritmos *AprioriSome* e *DynamicSome* procuram por seqüências com um certo comprimento. O *AprioriSome* gera candidatas em uma passagem pelo algoritmo sobre as grandes seqüências obtidas no passo anterior e após, procura pelo nível de suporte destas seqüências em uma nova passagem. O *DynamicSome* gera candidatas a grandes seqüências no mesmo instante em que as procura no BD [AGR 94b].

[3] Seqüência máxima é uma seqüência que não esteja contida em qualquer outra seqüência.

Outra solução, utiliza o algoritmo Padrões Seqüenciais Generalizados (PSG) que incorpora os conceitos de restrições temporais entre elementos adjacentes em um padrão, janela deslizante[4] e a taxonomia[5] aos padrões temporais descobertos [AGR 96a].

Neste trabalho, conforme as técnicas descritas acima, a técnica de MD a ser utilizada será o algoritmo *Apriori*.

Tabela comparativa entre as técnicas de MD

A tabela 2.3 a seguir, apresenta as principais técnicas de MD, a vantagem e desvantagem da sua utilização em comparação com as demais técnicas.

TABELA 2.3 - Tabela comparativa das técnicas de MD.

Técnica	**Vantagem**	**Desvantagem**	**Comparação com demais técnicas**
Redes Neurais	Versatilidade e resultado satisfatório em áreas complexas	Valores iniciais da base de aprendizado	Mais difícil compreensão
Árvore de decisão	Dividir problema	Instabilidade de pequenas variações e grandes BD	Mais fácil compreensão
Algoritmo genético	Maior qualidade nas gerações de indivíduos	Necessita de um número grande de exemplos para alcançar os resultados das árvores de decisão	Em pequenos BD onde se conheça pouco o domínio
Apriori	Descobrir padrões de comportamento freqüentes nos dados	Deficiente na detecção de seqüências com diversos níveis de suporte	Indicado para atender aos diversos interesses da MD.

A seguir, são apresentadas as tarefas relacionadas as técnicas mais indicadas de MD, os critérios para a seleção de uma potencial aplicação de MD e a direção tomada para este trabalho.

2.4 Considerações Finais

Observa-se que a utilização de uma técnica de MD é indicada em situações onde o volume de dados é muito grande, muito complicado, ou de menor extensão em comparação ao processamento manual em que especialistas humanos não estão disponíveis para prover conhecimento. A escolha de uma técnica de MD implica na investigação das suas técnicas e na sua extensão aplicada a novos domínios ou

[4] Janela deslizante é o intervalo temporal associado a um evento.
[5] Taxonomia é uma hierarquia ou grade de associações entre categorias diferentes de um item.

problemas [RAI 2001]. O relacionamento entre os diferentes tipos de problemas e a variedade de técnicas de MD existentes envolvem a aplicabilidade de certas técnicas relacionadas a certos tipos de métodos.

A figura 2.11 procura enquadrar as quatro técnicas de MD apresentadas na seção 2.3.2 nos três tipos de métodos de conhecimento diferentes da seção 2.3.1.

Na figura 2.11, observa-se que os diferentes métodos pressupõem a utilização de diferentes técnicas de MD. As técnicas das RNA apresentam-se mais relacionadas ao método da classificação do que a técnica dos algoritmos genéticos. Ao passo que, os algoritmos genéticos são mais adequados ao método das regras associativas. Os algoritmos *Apriori* são mais adaptáveis ao método das regras associativas do que as árvores de decisão. Entretanto, as árvores de decisão organizam-se mais próximas do método da classificação do que as RNA. As RNA estão mais próximas do método de *clustering* do que as árvores de decisão.

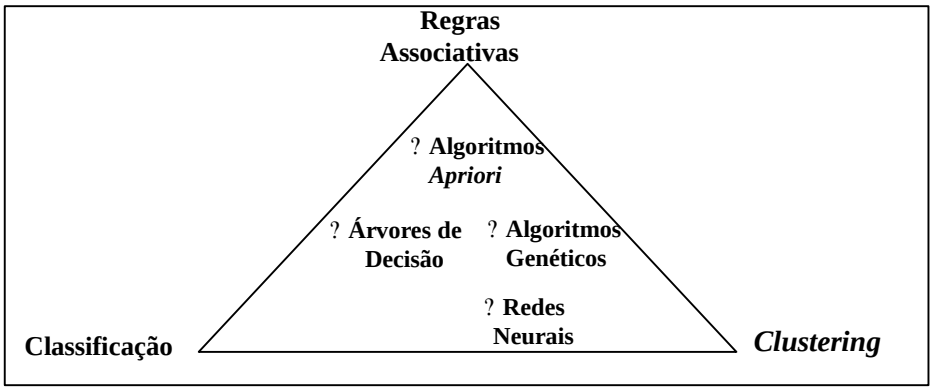

FIGURA 2.11 - Técnicas de MD indicadas pelos métodos.

Adicionalmente, observa-se que um ambiente de MD deve suportar estes tipos de tarefas diferentes de forma híbrida. Por exemplo, se o objetivo da descoberta for detectar fraudes em um ambiente de MD, por esta ser uma tarefa bastante complexa que engloba os métodos das regras associativas, classificação e *clustering*, deve-se utilizar as diferentes técnicas para a análise do conjunto de dados.

A escolha da técnica mais adequada de MD implica na capacidade em interpretar os resultados obtidos pelo algoritmo utilizado, apresentando eficiência computacional e efetividade na descoberta de um modelo geral para os dados existentes em um BD.

A literatura de MD, coloca a efetividade de um algoritmo como a característica mais importante a se observar, objetivando a minimização da taxa de erro da descoberta. Ela está relacionada a aplicação dos algoritmos a um BD específico. A partir de então, toma-se porções de dados para treinamento e teste com base em algum método de amostragem. Adicionalmente, observa-se que a característica da eficiência, visando minimização do tempo de processamento, é colocada em segundo plano [FRE 96; PRA 98].

Outro ponto importante para a seleção de um algoritmo de MD em BD pode ser a análise de qualidade da solução final. Isto é, se o algoritmo é capaz ou não de aprender as regras de conhecimento descobertas. Para se selecionar um algoritmo de MD, deve-se verificar se ele é capaz de aprender incrementalmente, pois novas informações se tornam disponíveis a cada MD. Isto significa que, não se necessita partir sempre do processo inicial de aprendizagem, pois o conhecimento descoberto pode ser incrementado a cada MD.

Finalmente, o aspecto da avaliação do algoritmo em relação a sua performance geral que objetiva a sua eficiência. A eficiência de um algoritmo é dividida em duas situações que são estágio de aprendizado e estágio atual de aplicação do algoritmo. A escolha da técnica de MD convém na utilização de um algoritmo que garanta a sua melhor performance sem comprometimento na qualidade do conhecimento descoberto.

Os critérios de seleção para uma aplicação de MD está dividido em prática e técnica. O critério prática é similar as outras aplicações de tecnologia avançada, já a técnica é mais específica para a MD.

O critério prática inclui considerar o potencial de impacto significativo de uma aplicação. Nas aplicações comerciais, o potencial de impacto pode ser medido como uma grande novidade, preço baixo, alta qualidade, ou salvamento da atividade a tempo. Nas aplicações científicas o impacto pode ser medido por novidade e qualidade no conhecimento descoberto e pelo aumento do acesso aos dados através da automação nas análises dos processos. Outra consideração da prática é a segurança na descoberta do conhecimento. Um exemplo disto seria quando o usuário apresenta interesse na atividade de MD e existe suporte organizacional para utilização da nova tecnologia na empresa. Entretanto, existe a necessidade de potencial legal e decisões privadas que garantam o descobrimento de padrões legais e éticos sem invasão de privacidade.

O critério técnico inclui considerar a avaliabilidade dos dados e casos suficientes para a MD. O número de exemplos ou casos existentes são avaliados para a inferência dos padrões úteis dos dados e para atender ao objetivo de uma aplicação particular. Em geral, quanto mais campos existem no BD, mais complexos são os padrões descobertos, e consequentemente, mais dados são necessários para a MD. Entretanto, pode-se reduzir o conteúdo informacional através dos casos mais significativos e dos atributos relevantes para a descoberta, pois nem todos necessitam ser previsivos.

O nível de ruídos ou erros dos dados também é levado em consideração, pois grandes quantidades de ruído tornam difícil identificar padrões nos dados, a menos que, um grande número de casos com erros randômicos possa agregar novos padrões de conhecimento aos dados.

Os intervalos de confiança do conhecimento extraído são importantes, porque em muitas aplicações é difícil atribuir intervalos de confiança e produzir previsões para o sistema de MD.

O conhecimento relevante é imprescindível para o conhecimento do domínio, dos campos, relacionamentos, funcionamentos e identificação de quais padrões já são conhecidos. O conhecimento prioritário pode significantemente reduzir as buscas dos dados na MD e demais passos da DCBD.

A partir disto, pode-se demarcar um ambiente potencial para uma aplicação de utilização de técnicas de MD considerando-se velocidade, complexidade, repetição e custos de implantação. Os custos da implantação podem ser justificados pela comparação da realização humana para a mesma tarefa [ADR 97].

Após, deve-se selecionar o conhecimento descoberto da organização em questão para que ela possa se auto-gerir, ou seja, se atualizar e utilizar o novo conhecimento em seu benefício.

Com base nestas observações e, primeiramente, no estudo efetuado dos métodos e técnicas da MD foi centralizado o método da classificação e técnica de árvores de decisão sobre a ferramenta SIPINA-W © [ZIG 85]. Entretanto, de acordo com a bibliografia lida e revisada, estas escolhas não se enquadram perfeitamente à MD considerando os aspectos temporais.

Assim, a solução adotada foi a mudança do método para a descoberta de padrões temporais descritos na seção 3.2 e da técnica para o algoritmo Apriori descrito na seção 4.2. Esta mudança, também refletiu na alteração da ferramenta de MD para o *Intelligent Miner* da IBM versão 6.1 [IBM 99]. A seguir, será apresentado o estado da arte da mineração de dados temporais que é o ponto central deste trabalho.

3 A Mineração de Dados Temporais

A Mineração de Dados Temporais (MDT) constitui-se na descoberta de padrões temporais implícitos, novos e úteis entre seqüências e sub-seqüências de eventos [AGR 95a; POL 99; SPI 2001a; OLI 2001]. Os padrões temporais são conhecimentos obtidos após a tarefa de MDT. As seqüências e sub-seqüências de eventos são conjuntos de registros de ocorrências importantes nos dados dos BD ordenados em relação ao tempo. A partir desta definição, verifica-se que a ordenação dos eventos relacionados leva a formação de seqüências temporais, pois os eventos podem estar ordenados em uma ou mais dimensões do tempo. Desta forma, múltiplas dimensões do tempo são admissíveis para um mesmo evento, caso um sistema de MDT acomode diversas linhas de tempo[6] apresentando informações armazenadas conforme o tempo de validade, o tempo de transação ou o tempo de decisão [EDE 94; EDE 98].

A figura 3.1 apresenta a MDT dentro do processo de descoberta temporal. Este processo é composto pelo conjunto de dados de entrada e o conjunto de saída que são, respectivamente, as seqüências temporais de eventos e os padrões temporais obtidos como resultados desta extração de conhecimento.

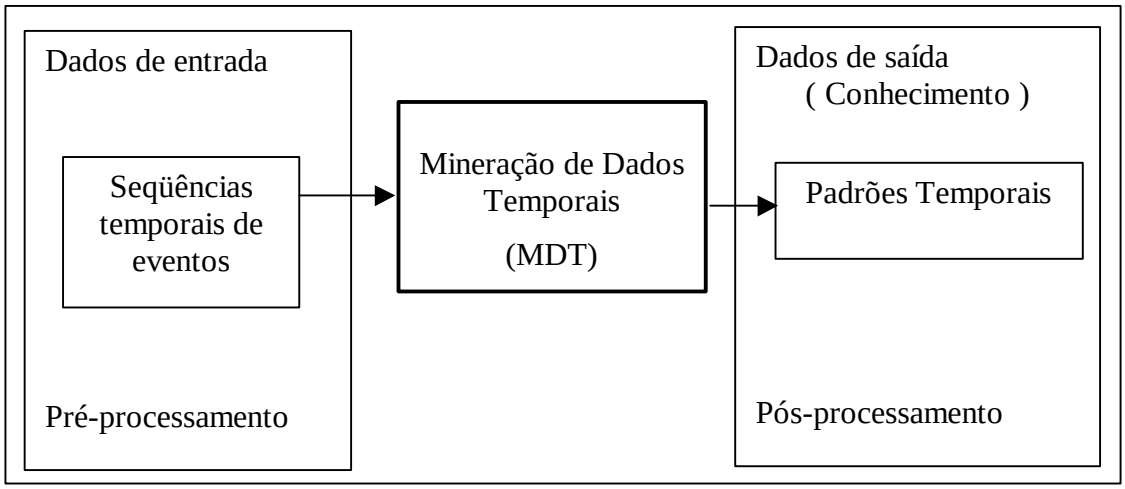

FIGURA 3.1 - A MDT dentro do processo de descoberta temporal.

Comparando-se a MD e a MDT, observa-se que a diferença existente entre elas está no tratamento das seqüências dos eventos. Enquanto as técnicas da MD tratam dos dados como coleções desordenadas de eventos, ignorando o seu aspecto temporal, a

[6] Linha de tempo é uma seqüência discreta de pontos consecutivos.

MDT depende das relações entre as seqüências e sub-seqüências dos eventos [SPI 2001a].

3.1 O Formalismo para uma Seqüência Temporal

O modelo formal para a representação de uma seqüência temporal é expresso conforme [AGR 93a] como uma lista de eventos ordenados de acordo com o tempo dos eventos, sendo que cada evento (E_n) corresponde a um conjunto de elementos [AGR 93a]. Desta forma, a seqüência temporal = {conjunto de elementos (E_1), conjunto de elementos (E_2), conjunto de elementos (E_3),...conjunto de elementos (E_n)}. O problema existente está em descobrir todas as seqüências dos dados para os registros do BD e neste conjunto de seqüências procurar por um padrão freqüente nos dados.

Observa-se, que o conjunto de elementos dos eventos é um conjunto de elementos que existem, ou seja, elementos não-vazios e não-nulos. A partir disto, têm-se que uma seqüência corresponde a uma lista de elementos ordenados. Conforme a teoria dos conjuntos, um dado conjunto A é subconjunto de B, se e somente se, todo elemento de A pertence também a B, representado por: A ? B. Assim, uma seqüência_a = {a_1, a_2, a_3...,a_n} está contida em outra seqüência_b = {b_1, b_2, b_3...b_n} se, na seqüência_b existirem elementos, tais que, $e_1 < e_2 < e_3 < ... < e_n$, de forma que, a_1 ? b_1, a_2 ? b_2, a_3 ? b_3, ...,a_n ? b_n [AGR 94b].

A partir de então, observa-se que a seqüência temporal carateriza-se por apresentar três componentes fundamentais que são: duração, simultaneidade e intervalo temporal nas relações entre os eventos [WIT 2000].

A duração refere-se ao intervalo de tempo observado para a descoberta, como, por exemplo, o experimento efetuado com as AIH do ano de 2000 no capítulo 5.

A simultaneidade trata de eventos que ocorrem em um espaço temporal específico e são analisados como ocorrências simultâneas. Um exemplo para esta característica é verificada ao observar-se que, um paciente com dor no braço esquerdo pouco tempo depois tem um ataque cardíaco. Estes dois eventos são tratados como ocorrências simultânteas em análises de seqüência.

O intervalo temporal constitui-se em encontrar eventos relacionados à seqüência. Por exemplo, a freqüência obedecida pela seqüência dos procedimentos de um paciente

que foi internado com Colecistite Aguda[7] é seguida por uma internação de Colecistectomia[8], porque eles se constituem em dois eventos consecutivos que ocorrem em uma seqüência.

3.2 Padrões Temporais

Os padrões temporais procuram por comportamentos freqüentes levando em consideração a dimensão temporal dos eventos.

Pode-se discernir o estudo da descoberta de padrões temporais conforme a descoberta que se pretende efetuar, o tipo de dado escolhido e a variação temporal[9] em duas direções que são: seqüencias similares e padrões seqüenciais.

As seqüências similares (SS) estudam as semelhanças ou diferenças entre as seqüências de um BD. As SS da MDT utilizam padrões pré-especificados, como os utilizados no processamento de sinais ou *strings*. Na descoberta de séries similares, os dados dispõem-se em uma seqüência contínua de elementos de valores reais que estão dispostos em uma escala da medida de proporção ou intervalo conhecidos por séries temporais. Assim, é possível analisar as similaridades entre os componentes de diferentes objetos no mesmo período. Um exemplo de uma série temporal são as médias de temperatura atmosférica de uma determinada região = {23.9, 24.4, 26.1, 26.2, 26.3, 26.5}.

Os padrões seqüenciais (PS) procuram por relações de dependência nos dados dos BD onde existe uma associação temporal nos eventos em uma ordem determinada. Estas relações de dependência são relativas a descoberta das relações causais acerca da orientação temporal dos eventos. Esta direção enfoca a descoberta da causalidade entre os relacionamentos dos eventos ordenados pelo tempo. Na descoberta de padrões seqüenciais, os dados são discretos ou categorizados e a extração de conhecimento efetuada é realizada sobre uma seqüência temporal. Um exemplo de uma seqüência temporal é representado na definição dos símbolos de uma seqüência de doenças diagnosticadas como, por exemplo, em {Malária, Leptospirose, Meningite}. O modelo de conhecimento extraído a partir de um padrões seqüenciais é representado por uma

[7] Colecistite Aguda é a inflamação na vesícula biliar.
[8] Colecistectomia é o procedimento médico de retirada da vesícula biliar.
[9] Variação temporal é a diversidade temporal encontrada em uma representação do tempo de um evento.

regra temporal. Em uma regra temporal, a causa de um evento sempre ocorre após o resultado da ordenação dos eventos de uma seqüência.

Adicionalmente, ambos os padrões temporais, mas sobretudo, as séries temporais estão mais sujeitas ao aumento do volume dos dados, e por esta razão, ao problema da alta-dimensionalidade dos dados. Para solucionar este problema, a MDT necessita que todos os dados estejam amostrados e convertidos em uma seqüência de dados discretos. Desta maneira, é escolhida uma certa granularidade temporal[10] mínima, isto permite diminuir a quantidade dos dados através da conversão de toda a massa de dados para esta granularidade. Assim, os padrões temporais são discretizados e designados de seqüências temporais.

Observa-se que tanto as séries similares quanto os padrões seqüenciais relacionam-se ao mesmo problema da descoberta de um padrão seqüencial freqüente. De acordo com [SIP 2001a], estes padrões temporais obtidos são diferentes terminologias empregadas para aludirem ao mesmo problema. Assim, os padrões temporais são empregados em vários domínios para se referir ao aspecto temporal dos eventos relacionados como, por exemplo, sistemas de telecomunicação, análises financeiras, previsões na agricultura e acompanhamentos de internações hospitalares entre outros.

3.3 Os Métodos da MDT

Os métodos da MDT procuram encontrar as maiores seqüências até as candidatas a seqüência se exaurirem. A geração de candidatas à seqüência da MDT é a tarefa de maior custo operacional, sendo que, para minimizar o custo desta tarefa é importante utilizar o método mais apropriado. Os métodos da MDT necessitam de algumas adaptações para manipular as seqüências temporais como, por exemplo, o fator de suporte. O fator de suporte é obtido através do número de seqüências que apresentam um padrão temporal sobre o número total de seqüências existentes.

A partir disto, a MDT apresenta métodos eficazes, capazes de descobrir padrões de eventos que ocorrem freqüentemente em uma seqüência de dados e que satisfazem a relacionamentos temporais específicos. A seguir, serão apresentados os principais métodos da MDT que são *clustering*, classificação e regras associativas.

[10] Granularidade temporal é a duração do tempo de um evento como, por exemplo, dias, semanas ou horas.

3.3.1 *Clustering*

O problema do *clustering* para seqüências temporais está em descobrir o número de *cluster* de um elemento da seqüência. Isto envolve duas questões diferentes como: a escolha do número de *cluster* e a inicialização dos seus parâmetros. Assim, a questão existente quando se trata de *clustering* para seqüências temporais é a medida da similaridade entre as seqüências. Por exemplo, dada uma seqüência qualquer, caso ela for gerada de acordo com um modelo probablístico, o *clustering* pode modelar a seqüência de dados como uma mistura das suas seqüências finitas. Já, os parâmetros do algoritmo podem estimar se cada grupo da seqüência corresponde a um *cluster* ou não [SMY 2000].

O método existente de *cluster* hierárquico para uma seqüência temporal utiliza o algoritmo COBWEB [FIS 87] que trabalha em duas etapas: a primeira, agrupando os elementos dentro de uma seqüência e a segunda, agrupando as seqüências entre elas. Por exemplo, considerando-se uma seqüência temporal, primeiramente, agrupa-se os elementos das seqüências e após, agrupa-se o quê é comum entre elas defindo-se o mecanismo de geração da seqüência.

3.3.2 Classificação

O método da classificação para seqüências temporais está baseado no operador *merge*. Este operador recebe duas seqüências e retorna a seqüência cujo formato é a união entre as duas seqüências originais. A idéia básica é a interatividade do *merge* entre uma classe típica com um exemplo positivo que constrói um modelo geral para a classe. A influência de um fator positivo implica na generalização de um seqüência, já um fator negativo, enfatiza as diferenças entre o positivo e o negativo.

A classificação descobre modelos probabilísticos e determinísticos para a geração de padrões nos dados desde que uma seqüência ocorra. Entretanto, algoritmos de classificação são dificilmente aplicados para exemplos de padrões temporais devido ao grande número de características existentes para descrever eficientemente cada evento [OGI 99]. Além das aplicações de classificação para a descoberta de padrões temporais estarem carentes de literatura [LAN 98].

Uma solução existente para reduzir o número das características dos dados é através da transformação dos padrões descobertos em um conjunto de características

binárias. Adicionando-se pesos a estas características, permite-se descobrir diferentes regras de conhecimento no BD. A partir disto, os padrões descobertos são combinados para classificar novos padrões. Desta maneira, são utilizados os paradigmas das seqüências temporais que procuram por padrões relacionados a classes; e da classificação que efetua descobertas associando pesos às características dos dados para a classificação de novos exemplos [OGI 99].

Uma ferramenta existente que efetua o método da classificação considerando dados temporais é a ferramenta *Easy Miner* da UMIST [THE 98]. Esta ferramenta seleciona os atributos de maior relevância acerca do BD como, por exemplo: a data de nascimento, o sexo e estado civil dos pacientes. Este tipo de conhecimento é útil para garimpar grandes BD que detém informação sobre muitos objetos. Entretanto, a informação do campo data de nascimento para representar o componente temporal desta transação é muito similar aos demais campos apresentados e poucas operações são necessárias em um processo de geração de árvore de decisão que envolva o aspecto temporal.

3.3.3 Regras Associativas

A descoberta de conhecimento através das regras associativas para seqüências temporais utiliza as sub-seqüências nas relações entre os eventos. Uma regra de associação para a MDT pode ser entendida como um regra de associação que inclue uma conjunção de um ou mais relacionamentos temporais entre os itens antecedente e conseqüente da regra. A partir disto, a representação do conhecimento obtido chama-se regras de associação temporais [RAI 99].

Uma proposta existente consiste em estender a noção da regra: X ? Y (onde: se X ocorre, então Y ocorre) para uma regra com um novo significado: X ? ^{T}Y (onde: se X ocorre, então Y irá ocorrer no tempo T) [DAS 98]. Este novo formato da regra permite controlar o impacto da ocorrência de um evento em relação a outro evento com um intervalo temporal específico. Um exemplo aplicado à área da saúde é dado pela regra de associação temporal representada na figura 3.2.

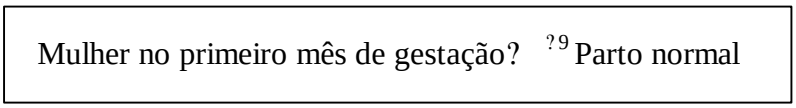

FIGURA 3.2 - Exemplo de regra de associação com extensão temporal.

Na regra da figura 3.2, verifica-se que a ocorrência de uma mulher no primeiro mês de gestação, implica na realização de um parto normal com um intervalo temporal de nove meses em média. Desta forma, a utilização de regras de associação com extensão temporal geram conseqüências que envolvem o tempo e mudanças no estado dos eventos.

A taxonomia de Allen é outra proposta existente que consiste em determinar as extensões das regras de associação por um conjunto de predicados temporais para descrever as relações entre os eventos. A taxonomia de Allen descreve os relacionamentos básicos entre os intervalos de dados utilizados na lógica temporal de predicados [ALL 83].

As regras associativas cíclicas ocorrem durante um intervalo regular de tempo [ÖZD 98]. Um exemplo de caso real na área da saúde que poderia ser empregado a regra de associação cíclica é representado na figura 3.3.

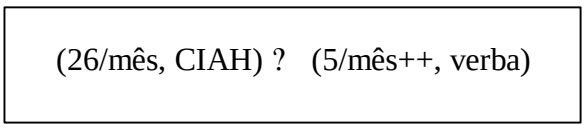

FIGURA 3.3 - Exemplo de regra de associação cíclica.

No antecedente da regra da figura 3.3, verifica-se a representação do evento que mensalmente ocorre todo dia 26, em que hospitais conveniados ao SUS entregam disquetes na SES correspondentes ao movimento mensal das internações efetuadas juntamente com o documento de Comunicado de Internação Hospitalar (CIAH). No consequente da regra, é representado o repasse da verba respectiva destinada a estes hospitais todo dia 5 do mês subsequente.

Observa-se que a descoberta de uma regra de associação cíclica envolve a ocorrência de uma regra repetidamente em um instante de tempo específico e durante uma pequena porção do tempo considerado atendendo a uma periodicidade. O método para se descobrir uma regra de associação cíclica consiste em aplicar um algoritmo de MDT que obtenha um conjunto de regras de conhecimento e que encontre os ciclos implícitos existentes nestas regras. Um outro método mais eficiente para se descobrir regras associativas cíclicas é invertendo este processo: primeiramente, obtendo-se ciclos em grandes conjuntos de elementos e depois generalizando-se as regras descobertas.

Uma extensão natural do processo de regras de associação cíclicas consiste em permitir a existência de diferentes granularidades de tempo como dias, semanas ou meses definidos em um calendário, e a partir de então, manipular os intervalos de tempo existentes. Estes tipos de regras descobertas são denominadas regras de associação calêndricas [RAM 98].

A seguir, são apresentados os tipos de padrões temporais encontrados a partir da MDT.

3.4 Os Tipos de Padrões Temporais Encontrados

A partir da descoberta de um padrão temporal através de um método da MDT, as demais relações entre sequências e subsequências de eventos são descritos por padrões seqüenciais extraídos a partir do primeiro padrão temporal encontrado como, por exemplo, tendência, ciclo, sazonalidade e irregularidade dos eventos.

O padrão temporal de tendência caracteriza-se pela variação no crescimento (crescimento ou decrescimento) linear ou exponencial de um evento em um período de tempo. Um exemplo de padrão temporal de tendência é o aumento dos números de internações hospitalares envolvendo casos de doenças respiratórias relacionados a diminuição brusca de temperatura associado ao aumento da umidade relativa do ar.

O padrão temporal de ciclo refere-se a repetição dos eventos acompanhando tendências, por exemplo, o ciclo do aumento de internações hospitalares por casos de emergência aos fins de semana e a diminuição de internações hospitalares durante a semana.

O padrão temporal de sazonalidade descreve padrões que se completam anualmente como, por exemplo, o aparecimento de doenças alérgicas, durante a estação da primavera, ocasionadas pelo contato com o pólen das plantas.

Os demais padrões temporais encontrados nos dados temporais podem ser movimentos destoantes de uma seqüência temporal que não seguem um padrão regular e que correspondem a inconsistências ou variações nos padrões temporais observados anteriormente, causadas pela ocorrência de ruído, valores nulos ou repetidos nos dados. Estas irregularidades, a princípio exceções, são muito freqüentemente denominadas

desvios como, por exemplo, a ocorrência de intrusões, falhas e anomalias entre outras [ARN 96; SAR 98; STO 99]. Contudo, observa-se que, muitas vêzes, estes desvios são inseridos intencionalmente nos BD para provocar a ocorrência de novas alterações, como uma renovação do conhecimento sobre o BD, ou para descobrir novos padrões seqüencias sobre a base de conhecimento.

3.5 A Representação dos Dados de Entrada

Antes de aplicar a MDT, os dados temporais necessitam ser representados, modelados e pré-processados. A representação dos dados temporais expressam o estado da seqüência temporal no BD. A seguir, é apresentada as soluções existentes para a representação dos dados de entrada.

3.5.1 Os BD Transacionais com Informação Temporizada

As soluções existentes para BD de transação que apresentam a informação temporizada utilizam tipos de seqüências temporais de difícil representação porque são muito diversificadas. Um exemplo de seqüência temporal é considerar a data de internação de cada um dos pacientes em um hospital como informação temporal de um BD.

Um método existente para encontrar sub-seqüências é considerar cada conjunto de elementos como um símbolo novo de um alfabeto. Estes símbolos são processados utilizando-se métodos para se encontrar sub-seqüências.

Esta representação é eficiente quando o objetivo é descobrir eventos comuns e correlações temporais significantes. Entretanto, esta representação apresenta desvantagem para a predição e classificação dos dados. Na predição, esta representação torna-se deficiente na escolha das variáveis temporais a utilizar para prever um valor desconhecido. Na classificação, esta representação encontra desvantagem para classificar dados temporais sem um conceito pré estabelecido.

Neste trabalho, a representação para os dados de entrada adotada será a de BD transacionais com informação temporizada. Entretanto, foram desenvolvidas outras formas de representação devido ao problema da alta-dimensionalidade dos dados, principalmente quando se trata de séries temporais com dados contínuos e diretamente manipulados. Contudo, as séries temporais não foram utilizadas no experimento efetuado. Assim, a título de informação, são apresentadas quatro representações para as

séries temporais objetivando reduzir o problema da alta-dimensionalidade existente nestes tipos de dados temporais.

3.5.2 A Representação Baseada na Transformação

A representação baseada na transformação converte a seqüência temporal em uma freqüência menor para melhorar o gerenciamento espacial dos dados contínuos. Assim, utiliza-se um ponto deste novo domínio para representar cada seqüência original.

Uma solução existente utiliza a Transformação Discreta de Fourier (TDF) que modifica uma seqüência temporal para um ponto de freqüência. A TDF representa cada seqüência como um ponto em um espaço multidimensional. A TDF tem a propriedade atrativa da invariância para as mudanças de amplitude dos coeficientes de Fourier. Esta propriedade permite estender o método para encontrar seqüências similares ignorando as alterações existentes [AGR 93a]. Observa-se que a TDF têm sido amplamente empregada no processamento de imagens e som para melhoramento, restauração, codificação e descrição das suas seqüências respectivamente [GON 2000].

A eficiência é uma vantagem da TDF. Entretanto, para projetar uma série temporal n-dimensional para um espaço multi-dimensional, os seus coeficientes necessitam ser utilizados em todas as séries do BD. Isto não torna a utilização da TDF otimizada para todo o BD, ocasionando ineficiência no algoritmo para encontrar qual o coeficiente de maior significância da série. Desta forma, esta característica constitue-se também, em uma desvantagem, pois o algoritmo da TDF necessita percorrer cada coeficiente do BD para obter o coeficiente máximo da série [GUN 2000].

Outra transformação representa cada seqüência como um ponto num espaço n-dimensional, onde n representa o comprimento da seqüência [GAV 2000]. Esta representação pode ser interpretada como uma representação temporal de domínio contínuo descrita na seção 3.2.5. Entretanto, se cada sub-seqüência é convertida para um ponto em um espaço com coordenadas derivadas de cada ponto da seqüência original, esta conversão é uma representação baseada na transformação [CHA 99]. Neste caso, cada ponto é determinado pela diferença entre o ponto e o seu antecedente.

A solução da Transformação Discreta de Wavelet (TDW) transforma cada seqüência temporal em um domínio de freqüência. A TDW é uma transformação linear

que decompõe a seqüência original em diferentes elementos de freqüência sem perda da informação sobre o instante da ocorrência dos elementos da seqüência quando não reduzida a poucos coeficientes. Entretanto, poucos coeficientes da TDW são necessários para representar uma seqüência. Desta forma, a seqüência é representada com poucos coeficientes de Wavelet [CHA 99; GUN 2000].

3.5.3 Os Métodos Baseados na Discretização

Os métodos baseados na discretização são traduções de uma seqüência inicial, elementos de valores reais, para uma seqüência discretizável. Um exemplo deste método é uma linguagem que descreve símbolos de um alfabeto e as relações entre as suas seqüências [AGR 95a], como o Modelo da Gramática da Linguagem (MGL) que descreve a seqüência dos dados temporais de um BD. No MGL, o primeiro passo no processo desta representação é a definição do alfabeto de símbolos e o segundo, é a tradução da seqüência inicial nesta seqüência de símbolos. Esta tradução é feita considerando as transições entre os instantes de ocorrência dos eventos e associando-se um símbolo de um alfabeto a cada transição.

Outro método de discretização existente, utiliza a segmentação de uma seqüência através da proporção da mudança entre um elemento da seqüência e o seu sucessor. Neste método de discretização, todos os elementos consecutivos são representados com esta mesma proporção de mudança formando um segmento único. Depois desta partição, cada segmento é representado por um símbolo e a sua seqüência por um *string* [HUA 99].

Uma solução diferente para converter uma seqüência em uma representação discreta é a utilização de um método de discretização por *clustering*. Neste método, a seqüência origina um conjunto de sub-seqüências. A partir de então, o conjunto de todas as sub-seqüências é clusterizado e um símbolo diferente é associado a cada *cluster* [DAS 98].

A vantagem destes métodos é que a série temporal é particionada de maneira natural conforme os seus valores. A discretização de séries temporais é mais fácil de manipular e processar do que os valores contínuos de uma série temporal que sofre as conseqüências da alta dimensionalidade dos dados. Entretanto, a desvantagem destes métodos está na escolha dos símbolos do alfabeto empregados. Pois os símbolos

utilizados pelos métodos baseados na discretização são escolhidos externamente, ou seja, são impostos pelos usuário ou estabelecidos pelo sistema de MDT.

3.5.4 Os Modelos Genéricos

Os modelos genéricos consistem em obter um modelo de representação para os dados temporais suficientemente genérico para que possa gerar as demais seqüências do BD. Esta solução é mais sofisticada pois utiliza um modelo estatístico ou determinístico para solucionar problemas mais complexos dos dados temporais.

Uma solução existente utiliza o modelo semi-Markov para modelar uma seqüência e encontrar as semelhanças de uma sub-seqüência particular nas demais seqüências do BD. A partir de então, o modelo estima probabilidades da seqüência apresentar sub-seqüências semelhantes [SMY 2000].

Já, a proposta de [MAN 95, MAN 96, MAN 97], identifica ordenações parciais entre símbolos e gera uma seqüência complexa de episódios com eventos básicos seriais ou paralelos.

Outras soluções dizem respeito a inferir gramáticas de seqüências temporais. Geralmente, elas se baseiam na busca por algoritmos discretos que inferem gramáticas complexas, entretanto recentes estudos dizem que as redes neurais para a inferência de gramáticas tem apresentado melhores resultados [GIL 2001].

3.5.5 A Representação Temporal de Domínio Contínuo

A representação temporal de domínio contínuo utiliza os dados com a mínima transformação. Assim, cada dado conserva a sua forma original ou utiliza janelas e aproximação linear para manipular sub-seqüências.

Um exemplo deste tipo de representação para uma seqüência de elementos de valores reais é utilizado, por exemplo, em uma série temporal, na qual os elementos iniciais são ordenados pelos seus instantes de ocorrência sem qualquer pré-processamento [AGR 95b; LIN 98]. Esta alternativa consiste em encontrar a aproximação da função linear que descreve a seqüência inicial. Desta forma, a seqüência inicial é particionada em vários segmentos e cada segmento é representado por uma função linear. Para particionar uma seqüência existem duas possibilidades: definir o número de segmentos, ou descobrir o número e a partir de então, identificar o segmento correspondente [DAS 98; GUR 99]. O objetivo é obter uma representação

capaz de identificar mudanças significativas em uma seqüência. Esta é uma representação forçada e que não acrescenta ganho de manipulação para representações mais genéricas. Uma utilização deste tipo de representação serve para a detecção de pontos de mudança significativa.

Outra solução para esta representação está baseada na idéia de sistemas visuais humanos e utiliza partições no plano de curvas em segmentos lineares. Esta proposta consiste em segmentar uma seqüência por *merging* interativo em dois segmentos similares. A escolha do segmento que será unido é baseado no critério da minimização do erro quadrado [KEO 97]. Uma estensão para este modelo consiste em associar a cada sub-sequência um peso que representa a sua importância na seqüência inteira, assim é possível comparar as seqüências [PAZ 98].

Estes tipos de representações tornam as séries temporais mais manipuláveis, permitem a definição da semântica dos eventos, são facilmente aplicáveis na MDT, além de reduzir o impacto ao ruído e demais irregularidades dos dados. Entretanto, problemas com a diferença de amplitude como, por exemplo, problemas de escalabilidade, e a existência de lacunas ou outras distorções no eixo temporal não foram satisfeitas.

3.5.6 Quadro Resumo das Representações de Dados Temporais

A importância das representações de dados temporais está na habilidade em identificar e representar as sub-seqüências de uma seqüência temporal. A tabela 3 apresenta um resumo das representações de dados temporais estudadas.

TABELA 3 - Resumo das representações de dados temporais.

Representação	**Vantagens**	**Desvantagens**
BD transacionais	É eficiente para representar eventos de dados comuns e correlações temporais.	Falha nas aplicações de predição e classificação.
Transformação	Melhora o gerenciamento espacial da seqüência.	Pode haver perda de informação no instante de ocorrência do elemento.
Discretização	A série temporal discretizada é particionada naturalmente e mais fácil de manipular.	Os símbolos dos alfabetos utilizados são escolhidos externamente.
Genérico	Para representar seqüências genericamente.	Difícil compreensão.
Domínio Contínuo	Torna a série temporal mais manipulável, permite a semântica dos eventos, e reduz o impacto ao ruído.	Existe problemas de escalabilidade e lacunas no eixo temporal.

3.6 A Representação dos Dados de Saída

A representação dos dados de saída apresentam o conhecimento da descoberta efetuada em seqüências e séries temporais.

Na representação dos dados de saída para seqüências temporais, os padrões previsíveis de comportamento são expressos na forma de regras temporais. As regras temporais apresentam um comportamento que está mais suscetível em produzir um outro comportamento ou uma seqüência de comportamentos durante um determinado período de tempo. Adicionalmente, a representação dos dados de saída para seqüências temporais apresenta o percentual de suporte para as seqüências temporais encontradas constituirem-se em um padrão temporal.

Um exemplo de representação dos dados de saída para seqüências temporais na forma de uma regra temporal é ``10% dos pacientes com *diabetes Mellitus* tem a tendência em desenvolver deficiências na visão'' (Retinopatia diabética).

Na representação dos dados de saída para séries temporais, os padrões temporais envolvem a descoberta de seqüências similares. A descoberta de seqüências similares é representada por quantis. Quantis são subconjuntos de dados dentro de um intervalo de valor previsto [IBM 99]. Por exemplo, a faixa de quantis Q[20,30] contém registros cujos valores previstos estão entre os quantis Q(20) e Q(30) do conjunto múltiplo de todos os valores previstos. Assim, a representação dos dados de saída para séries temporais apresentam as seqüências similares encontradas. A figura 3.4 apresenta uma visualização de uma janela de representação de seqüências similares. A partir da figura 3.4, observa-se que os nomes estão agrupados em pares das seqüências de dados que contém as subseqüências similares. O par do início da lista contém os nomes das seqüências de cada par. A terceira coluna apresenta a fração de correspondência entre cada par de seqüências. A fração de correspondência indica o grau no qual as seqüências similares se assemelham.

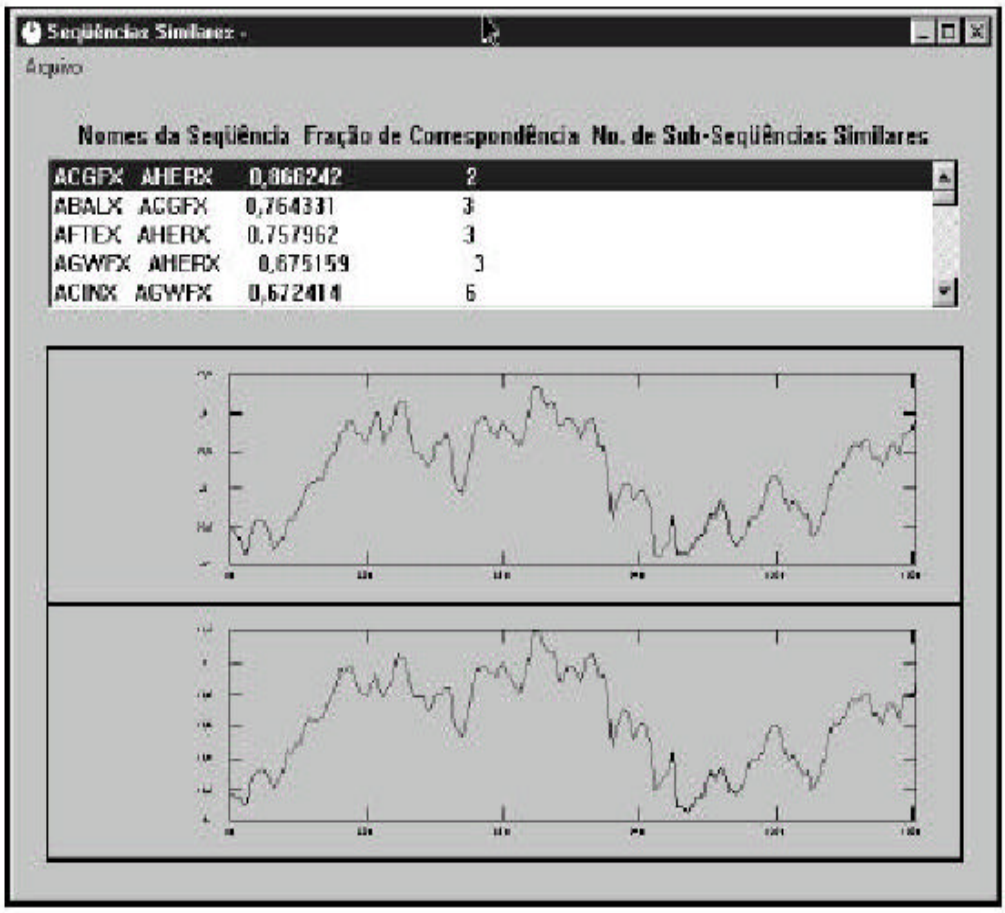

FIGURA 3.4 – Representação de seqüências similares.

Fonte: IBM, 1999. P. 319.

3.7 Considerações Finais

A principal característica da MDT é a presença de um domínio temporal dinâmico onde os dados são atualizados em BD regulares. Assim, é interessante examinar a forma como, tanto os dados, quanto o conhecimento são derivados através das mudanças do tempo. Para descobrir padrões temporais sobre grandes BD, os algoritmos propostos geram uma seqüência dos dados para os registros do BD e procuram, neste conjunto das seqüências por um padrão temporal freqüente. Assim, a descoberta de tendências, ciclos e padrões nos BD e a sua utilização em análises históricas e na previsão de dados futuros é identificável quando se detecta padrões temporais, tendências aparentes, processos que realçam a troca, e a periodicidade nas informações do BD.

4 Uma Metodologia para a MDT

Este capítulo apresenta uma metodologia para a MDT utilizando o software *Intelligent Miner* da IBM © com o objetivo de procurar padrões temporais que existem em grandes BD, mas que estão encobertos pela grande quantidade de dados. Adicionalmente, observa-se que as técnicas de BD e IA utilizadas estão muito vinculadas as necessidades da aplicação em questão. Desta maneira, esta metodologia procura tornar a MDT mais geral e independente da aplicação dentro do processo de DCBD objetivando atender aos diversos domínios existentes.

A figura 4.1 representa o processo para a descoberta temporal em relacionamentos causais de eventos de acordo com as três grandes fases da DCBD propostas por Fayyad que são: pré-processamento, MDT e pós-processamento.

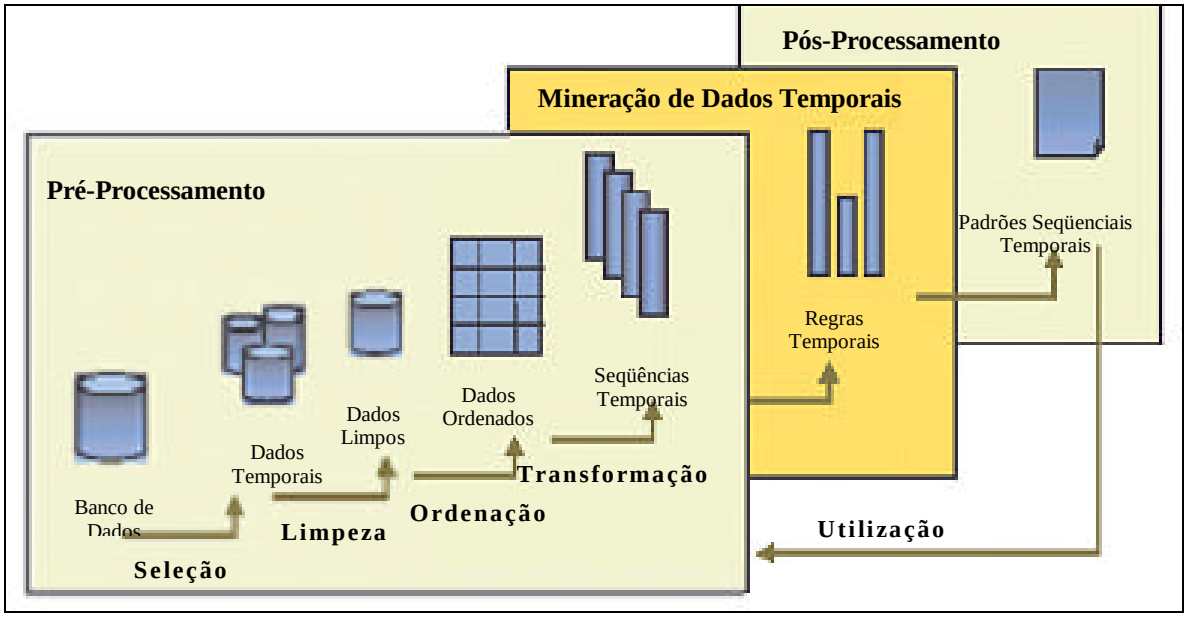

FIGURA 4.1 – Fases do Processo de Descoberta do Conhecimento.

Fonte: FAYYAD, 1997b p. 10 com adaptações.

Este processo inicia com a fase de pré-processamento. Nesta fase, um BD é selecionado, os dados temporais são escolhidos e limpos. Após a seleção e limpeza, os dados temporais são ordenados. A partir disto, os dados ordenados são transformados em seqüências temporais.

A seguir, a fase da MDT procura dentre todas das seqüências temporais obtidas no pré-processamento, aquela seqüência que atenda ao suporte mínimo de apresentar uma freqüencia de comportamento e não esteja contida em qualquer sub-seqüências. O modelo de conhecimento obtido nesta fase é representado na forma de regras temporais.

A fase do pós-processamento, utiliza as regras temporais extraídas na fase da MDT para obter padrões temporais.

A figura 4.2 abaixo, apresenta um conjunto de passos da metodologia para a MDT.

Efetuar o pré-processamento dos dados a partir de um BD
Selecionar os dados temporais
Limpar os dados temporais
Ordenar os dados temporais
Transformar os dados temporais ordenados
em seqüências temporais
Efetuar a MDT sobre as seqüências temporais para obter regras temporais
Efetuar o pós-processamento das regras temporais para utilizar
o conhecimento dos padrões temporais obtidos

FIGURA 4.2 – Passos da metodologia para a MDT.

A seguir, os passos da metodologia para a MDT serão explicados conforme as três grandes fases da DCBD em pré-processamento, MDT e pós-processamento.

4.1 Pré-processamento

A fase do pré-processamento é composta pelas etapas: seleção, limpeza, ordenação e transformação dos dados a partir de um BD.

4.1.1 Seleção

A etapa da seleção consiste em selecionar um conjunto dos dados temporais mais significativos sobre o qual a descoberta será efetuada objetivando a solução de um problema específico. A seleção utiliza linguagens de consulta como, por exemplo, SQL

visando obter como resultado um sub-conjunto dos dados temporais mais significativos do BD.

Na etapa da seleção, para escolher o conjunto dos dados temporais mais significativos é necessário escolher três tipos de atributos que são: principal, temporal e elementar.

O atributo principal é selecionado por apresentar-se em várias ocorrências ao longo do tempo. Por exemplo, o atributo principal selecionado para uma aplicação médica poderia ser os pacientes que sofrem internações hospitalares ao longo do tempo.

O atributo temporal refere-se as diversas ocorrências temporais do atributo principal. Um exemplo de atributo temporal poderia ser as datas das internações hospitalares dos pacientes.

O atributo elementar constitui-se nos elementos contidos em cada ocorrência temporal do atributo principal. Um exemplo de atributo elementar poderia ser os procedimentos médicos que são realizados.

A figura 4.3 apresenta uma representação para o conjunto de dados temporais selecionados em um arquivo. O atributo principal contém duas ocorrências temporais. A primeira ocorrência temporal contém os elementos 1 e 2. A segunda ocorrência temporal contém apenas o elemento 1.

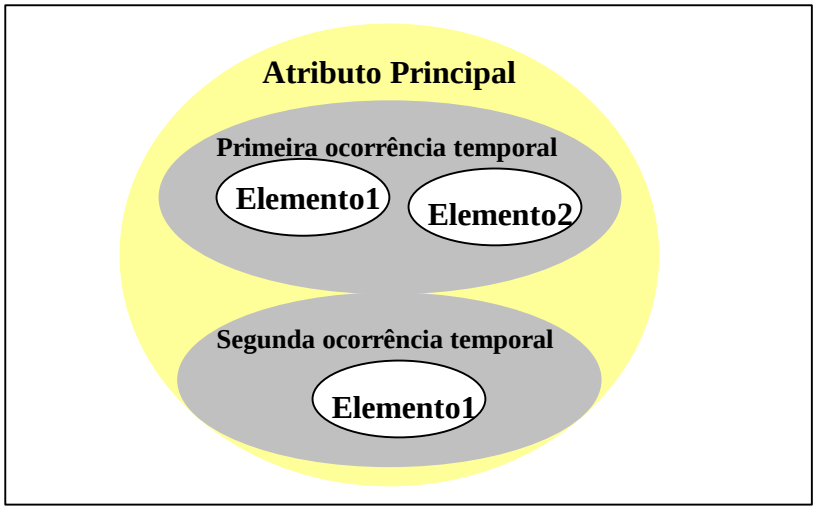

FIGURA 4.3 – Representação dos atributos selecionados em um arquivo.

Na figura 4.3, observa-se que as ocorrências temporais contém atributos diferentes que podem ser representados em listas de atributos tais como, a primeira

ocorrência temporal contém a lista de atributos <Elemento1, Elemento2> e a segunda ocorrência temporal contém apenas o atributo <Elemento1>. Assim, cada atributo que pertença a uma lista de atributos é considerado um elemento desta lista.

A nomenclatura de atributo temporal adotada para representar as ocorrências temporais do conjunto de dados temporais selecionados nos BD, foi utilizada para evitar a ambigüidade com o termo transação em BD.

A denominação de atributo elementar para os dados contidos em cada atributo temporal foi utilizado para generalizar o conceito de item empregado por sistemas da área comercial.

A importância da classificação do conjunto de dados temporais selecionados nos BD em atributos: principal e temporal foi escolhida para representar os atributos freqüentes nos arquivos que apresentavam diversas ocorrências de outros atributos ao longo do tempo.

A seguir, a tabela 4.1 apresenta um arquivo original (D). Nesta tabela, o atributo <sexo> é identificado como o atributo principal; o atributo <data>, como atributo temporal, e o atributo <valor> é identificado como o atributo elementar. Observa-se que o arquivo selecionado pode estar desordenado ou apresentar uma ordenação qualquer como a existente.

TABELA 4.1 – Arquivo original (D).

ORDENAÇÃO	SEXO	DATA	VALOR
5	2	04/01/02	Z
24	1	01/01/01	B
963	1	04/01/02	F
1000	1	02/01/02	C
6897	2	03/01/02	C
6989	2	01/01/02	F
11030	1	01/01/02	A
50974	2	02/01/02	X
78451	2	01/01/02	A
99999	1	03/01/02	D

A figura 4.4 representa o arquivo original da tabela 4.1 com conjuntos de dados temporais.

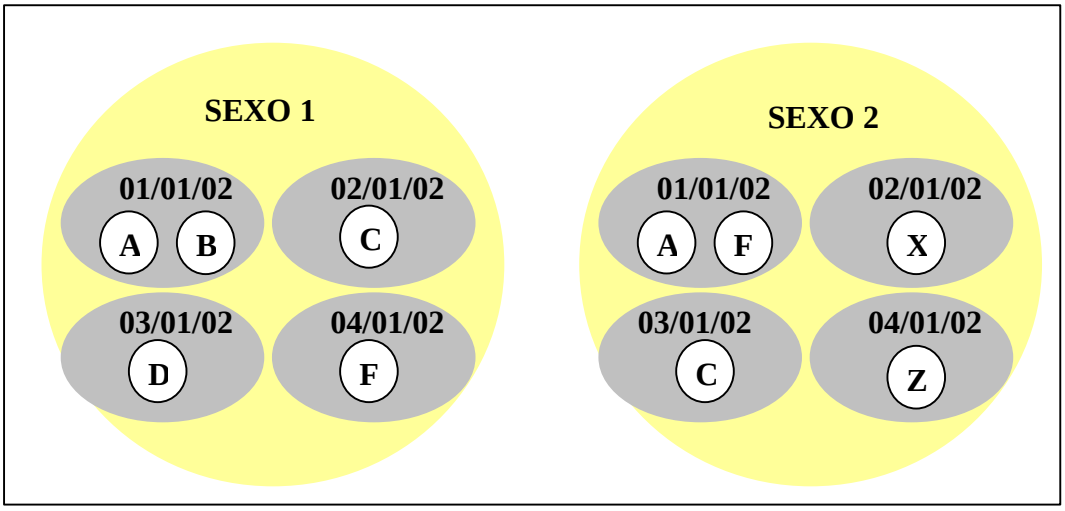

FIGURA 4.4 – Representação dos atributos da tabela 4.1
como conjuntos de dados temporais.

4.1.2 Limpeza

De acordo com [MIL 92], o grande volume dos BD torna o trabalho da MDT tedioso e suscetível a erros. Muitos destes erros são decorrentes de problemas de armazenamento e de inconsistências nos próprios dados. Além dos dados armazenados conterem muitas irregularidades decorrentes das alterações temporais da informação, como, por exemplo, na renomeação dos campos, na semântica especial dos valores de dados e na infidelidade temporal. Nesta situações, os atributos do BD contém informações que não são conhecidas no tempo em que o modelo pretende ser aplicado, porque estas informações sofreram manutenções com o decorrer do tempo [JOH 97].

Assim, verifica-se que é necessário tratar as irregularidades encontradas nos dados temporais selecionados. A etapa da MDT chamada de limpeza trata as irregulariades existentes em BD promovendo a simplificação da estrutura, os testes de validade, tratamento dos valores nulos, entre outros, garantindo a consistência dos dados.

A limpeza dos dados através da exclusão dos registros irregulares resulta em perda no volume dos dados do BD. Observa-se que a perda do volume dos dados no aprendizado de máquina (AM) [LAN 96] é menos significativa do que na MDT devido

a etapa de limpeza dos dados temporais. Isto é explicado, porque as técnicas de AM tratam dos dados de entrada em estruturas simples apropriadas para o processo de aprendizagem, enquanto que na MDT, a redução do volume nos dados temporais durante a etapa de limpeza comprometerem a qualidade do conhecimento a ser descoberto.

A solução mais indicada que deve ser aplicada para evitar a perda de dados de um BD é através da utilização de ferramentas que filtrem os dados para eliminar o ruído existente, pois elas determinam e executam estratégias para tratar as inconsistências existentes nos dados temporais. Existem diversas ferramentas que se propõem a determinar e executar estratégias para tratar dos valores dos atributos inexistentes como, por exemplo, *CopyManager, DataStage, ExtractPowerMart, DecisionBase, DataTransformationService, MetaSuite, SargentSolutionPlataform* e *WarehouseAdministrator* [RAH 2000].

A partir disto, observa-se que a etapa da limpeza é responsável por avaliar e aumentar a qualidade dos dados temporais selecionados preparando-os para a etapa de ordenação a seguir.

4.1.3 Ordenação

Após a seleção dos atributos de dados temporais, é necessário agrupá-los dentro do arquivo. Para isto, efetua-se a ordenação do arquivo pelo atributo principal, seguido do atributo temporal. A partir disto, observa-se que o atributo principal é o que se repete em várias ocorrências do atributo temporal.

Este procedimento, implicitamente, converte o arquivo original em um arquivo de conjuntos de seqüências (L) conforme a tabela 4.2. Adicionalmente, verifica-se que o arquivo fica desordenado em relação ao atributo de ordenação anterior.

TABELA 4.2 - Arquivo de conjuntos de seqüências (L).

Atributo de ordenação anterior	PRINCIPAL	TEMPORAL	ELEMENTAR
24	1	01/01/01	B
11030	1	01/01/02	A
1000	1	02/01/02	C
99999	1	03/01/02	D
963	1	04/01/02	F
6989	2	01/01/02	F
78451	2	01/01/02	A
50974	2	02/01/02	X
6897	2	03/01/02	C
5	2	04/01/02	Z

Após a ordenação dos dados temporais, identificam-se todos os elementos (l) do arquivo (L) conforme a tabela 4.3. Assim, simultaneamente, encontram-se os conjuntos de todos os elementos dos conjuntos expressos por { <l> | l ? L }. Este mapeamento é necessário para a identificação e comparação de l_1 = <B,A,C,D,F> e l_2 = <F,A,X,C,Z>.

TABELA 4.3 - Identificação dos conjuntos de elementos do arquivo.

Conjuntos de elementos	IDENTIFICAÇÃO	SEQÜÊNCIA
1	l_1	<B,**A**,**C**,D,F>
2	l_2	<F,**A**,X,**C**,Z>

Através deste procedimento é possível identificar as seqüências de elementos que são formadas pela lista temporal ordenada de um atributo principal. Assim, a lista temporal do atributo principal 1, contém a seqüência de elementos <B,**A**,**C**,D,F>, e a lista temporal do atributo principal 2, contém a seqüência de elementos <F,**A**,X,**C**,Z>.

Adicionalmente, verifica-se que os elementos A e C se repetem nos dois conjuntos de elementos formando uma seqüência temporal. No primeiro conjunto, os elementos A e C formam uma seqüência temporal contínua de elementos. No segundo conjunto, os elementos A e C formam uma seqüência temporal descontínua de elementos, porque o elemento X apresenta-se entre os elementos A e C.

4.1.4 Transformação

A etapa da transformação consiste em simplificar a estrutura complexa dos dados para obter um formato apropriado às técnicas de MDT. Para isto, as técnicas de

generalização e de transformação são comumente empregadas. O resultado deste processamento deve ser uma base representativa do formato atributo - valor.

Na etapa da transformação, determinam-se os maiores conjuntos de elementos contidos no arquivo das seqüências e os converte para um tipo inteiro. Assim, cada seqüência será transformada em uma nova seqüência derivada da primeira. A partir disto, cada seqüência pode ser transformada em uma representação alternativa discretizada conforme a tabela 4.4. A transformação efetuada sobre o arquivo (L) é representada por L_T. Observa-se, que a transformação leva em consideração o primeiro e último elemento da seqüência. O elemento X da segunda seqüência é desprezado porque ele não se apresenta na primeira seqüência.

TABELA 4.4 - Transformação das seqüências do arquivo.

Conjunto de elementos	**Seqüências**	**L_t**	**L_t presente no conjunto de elementos**
1	<A,C>	<A><C>	1,2
2	<A,X,C>	<A> <C> ;<A,C>	1,2;2

Observa-se que na fase do pré-processamento da MDT, os dados temporais são originários de um BD, a estrutura de dados é mais complexa, a quantidade de informações é enorme e nem todas as informações são relevantes. Desta maneira, é necessário realizar dois pré-processamentos para posterior aplicação de técnicas de MDT: um para encontrar um subconjunto de informações relevantes compostas pela seleção, limpeza e ordenação dos dados temporais; e outro, para a formatação destas informações em uma estrutura apropriada correspondente ao tratamento dos dados temporais [BIG 2000; RAI 2001].

4.2 MDT

Na fase da MDT, conforme o objetivo de descoberta que o usuário possui, deve-se adaptar os dados às necessidades de algum método particular de MDT como, por exemplo, *clustering*, classificação, regras associativas, padrões temporais entre outros.

Neste passo, faz-se a escolha dos algoritmos de MDT. Como existem diversos algoritmos que podem realizar aproximadamente a mesma tarefa, deve-se decidir quais os mais adequados para a atual aplicação, pois seus resultados irão repercutir em um significativo efeito na qualidade dos padrões temporais extraídos.

A dificuldade em se efetuar a MDT está na disparidade entre o problema e o algoritmo que é geralmente causado por conjuntos inadequados de ferramentas de MDT ou tentativas de se adequar o algoritmo a diferentes tipos de problemas para os quais o algoritmo não tenha sido desenvolvido.

A figura 4.3 apresenta a MDT dentro do processo de descoberta temporal. Esta fase é composta por um conjunto de dados de entrada que se constituem nas seqüências temporais obtidas na fase do pré-processamento, e um conjunto de saída que são os padrões seqüenciais (PS) obtidos como resultados desta extração. Adicionalmente, observa-se que as regras temporais são o modelo de representação do conhecimento extraído a partir da MDT.

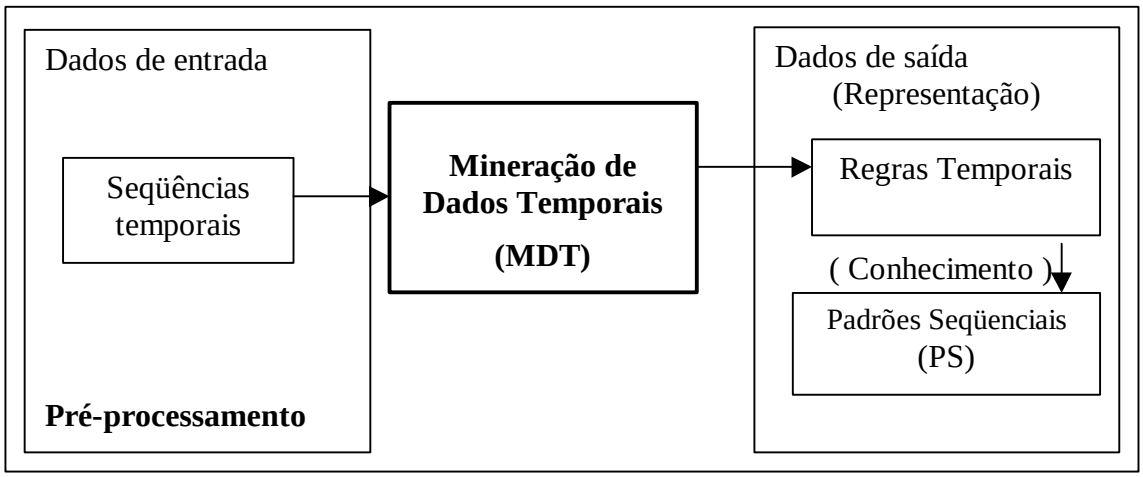

FIGURA 4.5 - A MDT dentro do processo de descoberta temporal completo.

Na fase da MDT, efetuam-se múltiplos passos sobre as seqüências temporais para se encontrar as seqüências candidatas a seqüência máxima que apresente um padrão seqüencial freqüente nos dados através de extensões do algoritmo *Apriori* descritos na seção 2.3.2.

Em regras gerais, o funcionamento do algoritmo *Apriori* se divide em três fases descritas a seguir:

? A primeira fase do algoritmo *Apriori* consiste em percorrer todas as seqüências temporais determinando o suporte mínimo de cada elemento e o número de ocorrências temporais que apresentem o elemento. Ao final deste passo, são obtidos quais elementos são freqüentes e o suporte mínimo exigido para que os dados apresentem um padrão freqüente de comportamento.

? A segunda fase do algoritmo *Apriori* efetua a transformação das seqüências temporais. Esta transformação do BD é feita de modo que cada ocorrência temporal seja substituída pelo conjunto dos elementos mais freqüentes presentes em cada ocorrência temporal.

? A terceira fase do algoritmo *Apriori* obtém-se os padrões temporais presentes nos conjuntos de elementos das seqüências máximas do passo anterior. As seqüências máximas são aquelas que não estão contidas em qualquer outra sub-seqüência.

Verifica-se que a MDT é a fase que mais consome recursos computacionais devido à sua complexidade e tempo necessário para o processamento de grandes volumes de informação. Portanto, é necessário um estudo bem preciso para redução de tempo, recursos gastos e acertos dos erros decorrentes dos passos anteriores de seleção, agrupamento e tratamento dos dados temporais referentes a fase do pré-processamento, pois eles podem acarretar em maiores prejuízos no decorrer deste passo [AGR 96a, AGR 93].

4.3 Pós-processamento

Após a MDT, deve-se interpretar os padrões temporais extraídos. Isto resulta, provavelmente, em retornar a alguma das etapas anteriores para uma próxima iteração. Os métodos utilizados para analisar os padrões temporais dependem da solução do problema.

A interpretação formata os modelos de conhecimento obtidos em representações exploráveis de regras temporais. Esta representação pode ser apresentada ao usuário ou reutilizada pelo sistema. Normalmente quando o especialista interpreta o modelo, uma destas três coisas acontecem: o especialista fica satisfeito com o modelo, porém já conhecia os padrões contidos; o especialista fica satisfeito com o modelo e genuinamente surpreso com alguns padrões; ou o especialista fica insatisfeito com o modelo. No primeiro caso, o modelo é interessante porque comprova, ratifica e extrai os conhecimentos do especialista. No segundo caso, maiores análises são necessárias sobre os padrões que surpreenderam o especialista. No terceiro caso, o problema necessita ser redefinido.

Na fase do pós-processamento, também encontram-se os níveis de suporte para as seqüências temporais dentro do conjunto das maiores seqüências obtidas na fase de MDT. A tabela 4.5 apresenta o resultado da fase do pós-processamento.

TABELA 4.5 – Seqüência máxima extraída do arquivo da tabela 4.2.

Padrão seqüencial temporal com suporte = 100%
(A,C)

O suporte da seqüência temporal de 100% foi obtido pelo número de conjuntos de elementos do arquivo que suportam a seqüência temporal dividido pelo número de conjuntos dos elementos do arquivo. O resultado encontrado é multiplicado por cem para se obter o percentual deste nível.

A representação dos dados seqüenciais, segundo o *Intelligent Miner* da IBM © (IM) , é descrita por informações estatísticas tais como:

? Quantidade de dados diferentes

? Ocorrências temporais encontradas

? Número de elementos possíveis.

A figura 4.6, a seguir, exibe uma representação dos dados de saída do IM obtida ao se efetuar a MDT aplicado aos dados da tabela 4.2. Ela consiste em uma janela de estatísticas que apresenta as informações do BD como parâmetros da execução dos padrões temporais.

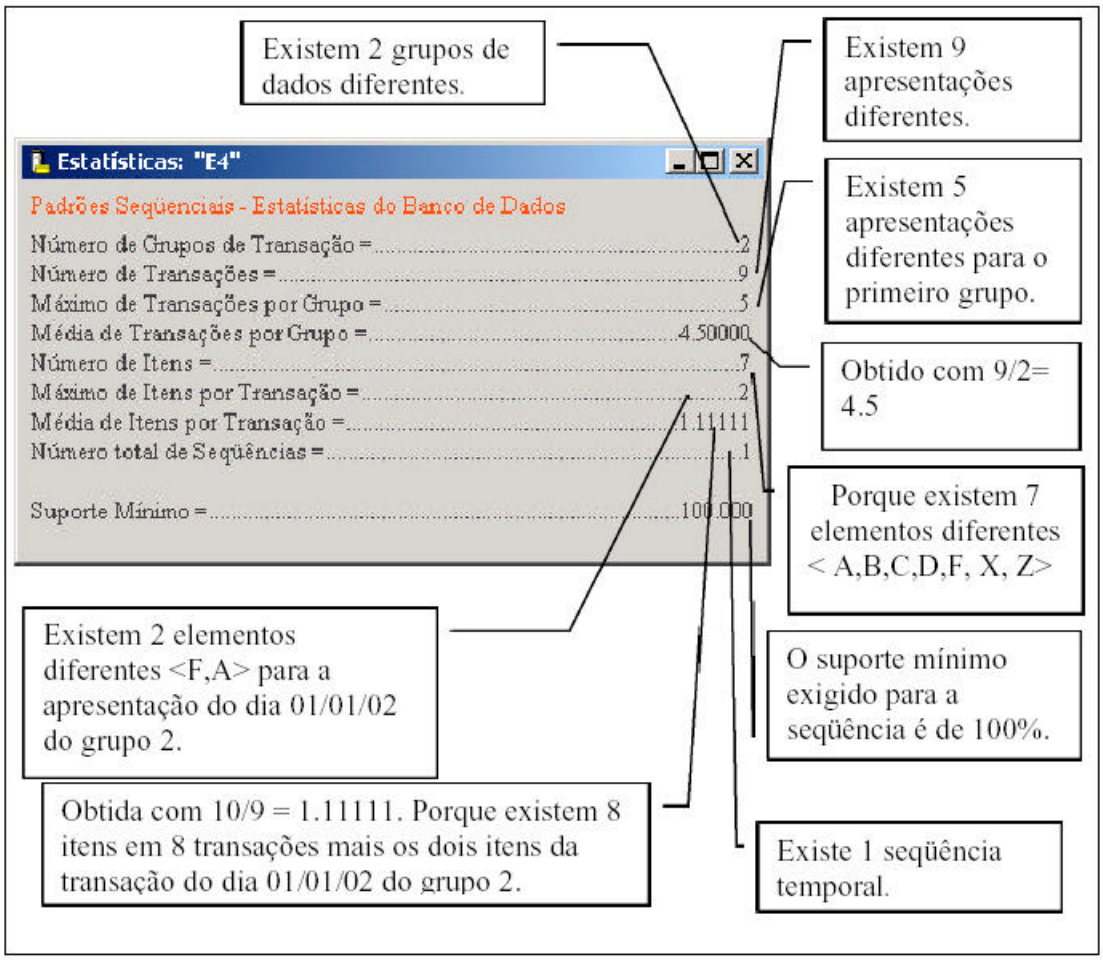

FIGURA 4.6 – Representação para os dados de saída do *Intelligent Miner* do IBM ©.

O primeiro parâmetro indica o número de identificadores da seqüência como, por exemplo, o número de pacientes encontrados no BD.

O segundo parâmetro representa o total de ocorrências temporais encontradas. Este número representa o número de ocorrências temporais diferentes existentes de pacientes e datas de internação.

O terceiro parâmetro consiste do número máximo de ocorrências temporais encontradas no BD.

O quarto parâmetro expressa a média de ocorrências temporais por grupo de pacientes obtida pelo resultado da divisão do número de pacientes pelo total de ocorrências temporais apresentadas.

O quinto parâmetro exibe a quantidade de elementos diferentes apresentados nas seqüências como, por exemplo, os diferentes procedimentos médicos realizados.

O sexto parâmetro indica o número máximo de elementos diferentes por ocorrência temporal.

O sétimo parâmetro expressa o número médio de elementos por ocorrência temporal.

O oitavo parâmetro apresenta o número total de seqüências temporais extraídas do BD.

4.4 Validação dos Resultados

Após avaliar os padrões temporais obtidos, passa-se à consolidação do conhecimento descoberto, incorporando-o ao sistema para ações futuras, ou pode-se simplesmente documentá-lo e fornecê-lo a quem interessar. Isto inclui a verificação e resolução de conflitos potenciais existentes em conhecimentos prévios como, por exemplo, a quebra de antigas crenças e concepções arraigadas aos negócios. Nesta etapa, procede-se a validação dos resultados ou conhecimento adquiridos.

A validação pode ser efetuada pela equipe do projeto que compara o desempenho do sistema com relação ao conhecimento da equipe de especialistas.

A dificuldade encontrada na interpretação e validação dos dados está na infidelidade temporal mencionada na fase de pré-processamento dos dados que mascaram os resultados e análises desta fase. Outra dificuldade existente de incorporação dos dados está na descoberta de padrões temporais úteis e válidos às necessidades das organizações.

Adicionalmente no pós-processamento, pode-se efetuar alguma estratégia envolvendo ações preventivas, corretivas ou decisórias entre outras sobre a organização em questão. A partir da descoberta de padrões temporais sobre o BD, pode-se desenvolver várias maneiras de se obter novas ações estratégicas sobre a organização e encontrar outros tipos de padrões temporais como, por exemplo, os descritos na seção 3.4.

4.5 Considerações Finais

Após a fase incorporação dos dados da MDT, quando todos as outras fases se completam, isto não significa que a atividade de descoberta de conhecimento acabou.

Dentro dos próximos anos se presenciará a explosão de novas informações, com mais e mais arquivos contendo mais e mais dados, isto repetidamente criará novas oportunidades para as organizações. O significado da DCBD não está em apenas uma atividade que é implementada e depois ignorada; para o permanente sucesso de uma organização, esta deve ser continuamente alimentada e alertada das possíveis novas fontes de informação e tecnologia disponíveis.

5 Experimentos

Este capítulo descreve os experimentos realizados sobre os dados da Secretaria Estadual da Saúde do Rio Grande do Sul (SES). Estes experimentos seguem a metodologia para a MDT apresentada no capítulo 4 e a aplica sobre dois experimentos de dados reais.

O primeiro experimento refere-se ao problema que procura descobrir padrões seqüencias de comportamento nas Autorizações de Internação Hospitalares (AIH) que determinem uma seqüência de comportamentos ao longo do tempo.

As AIH são fornecidas pelo Ministério da Saúde mensalmente à SES de acordo com o quantitativo estipulado para o Estado que é proporcional a sua população residente. Os números de AIH têm uma validade de quatro meses, isto permite uma certa compensação temporal, naqueles meses em que a sazonalidade da ocorrência de doenças influencia fortemente no número de internações hospitalares.

O segundo experimento refere-se ao problema que procura descobrir padrões seqüenciais de comportamento apresentado pelas fichas individuais de notificação e investigação, declarações e boletins de atendimento de doenças de agravos notificáveis como, por exemplo, leptospirose, malária, toxoplasmose entre outras. Estas informações fazem parte do Sistema Nacional de Agravos Notificáveis (SINAN) que coleta e processa os dados sobre agravos de notificação de doenças em todo o território nacional.

Os dados do BD da SES estão organizados em vários arquivos que são: ocorrências de doenças de agravos notificáveis, movimento das AIH, procedimentos especiais, atos, movimento dos hospitais, valores da AIH, controle de AIH e demais tabelas de cadastros adicionais. A seguir, são apresentados os arquivos que compõem o BD.

O arquivo de ocorrências de doenças de agravos notificáveis contém registros referentes ao ano de 1999. Estes registros contém informações a respeito do atendimento dos pacientes como, por exemplo, município de ocorrência, data de atendimento, identificação dos pacientes, especificação das doenças apresentadas, entre outras.

Os arquivos de movimento das AIH contém registros mensais do ano de 2000 referentes a cada internação hospitalar efetuada nos municípios do RS. Estes arquivos apresentam informações básicas de identificação do paciente e de caracterização da internação. Alguns exemplos das informações de pacientes presentes nestes arquivos são: registro do paciente no cadastro nacional de pessoas físicas, nome do paciente, data de nascimento, sexo, residência entre outras. Alguns exemplos de informações de caracterização da internação são: número da AIH, hospital, especialidade, procedimentos solicitado e realizado, diagnósticos, datas de internação e alta, motivo da cobrança e outras.

Os arquivos de procedimentos especiais possuem um arquivo para cada mês do ano e contém registros correspondentes a cada procedimento especial autorizado de AIH nos municípios do RS.

Os arquivos de atos contém registros de cada ato profissional efetuado mensalmente nos municípios do RS.

Os arquivos de movimento dos hospitais apresentam registros com informações de lançamentos como, por exemplo, pagamentos e descontos de cada hospital em cada procedimento efetuado.

Os arquivos de valores da AIH contém registros referentes aos faturamentos de cobrança de AIH para cada mês do ano de 2000.

O arquivo de controle apresenta AIH que foram bloqueadas pela SES por serem de baixa permanência, reapresentadas, ressarcidas pelo Estado ou apresentarem problemas de repetição por homonímia ou alguns procedimentos médicos realizados como, por exemplo, cuidados prolongados, cirurgias múltiplas, septicemia, politraumatizados, transplantes, acidente vaso-cerebral (AVC) agudo ou doença pulmonar obstrutiva crônica (DBPOC).

Os arquivos de tabelas correspondem a cadastros adicionais de atos, de identificação de doenças e de municípios.

5.1 Ferramenta Utilizada

Os experimentos foram realizadas sobre a ferramenta *Intelligent Miner* da IBM © (IM), porque ela apresenta uma tecnologia de extração do conhecimento bastante completa para efetuar a DCBD. Além desta ferramenta apresentar funções de pré-processamento, transformação e extração do conhecimento, ela também permite efetuar a MD considerando os aspectos temporais dos dados. A seguir, algumas das características da ferramenta IM são descritas.

A ferramenta IM apresenta várias funções tais como: classificação, agrupamento, associação, padrão seqüencial, previsão e seqüência similar.

Adicionalmente, as funções estatíticas presentes na ferramenta IM permitem a exploração e análise dos dados a fim de auxiliar na seleção dos dados relevantes de entrada como, por exemplo, análise de fator, curva univariada e regressão linear.

A partir da especificação dos dados de entrada, eles podem ser transformados através de funções de pré-processamento como, por exemplo, agregar valores e registros, articular campos em registros, calcular valores, codificar valores ausentes e inválidos, discretização, filtragem e ligação entre outras funções em grandes BD com ferramentas de visualização que permitem exibir e interpretar os resultados da extração de conhecimento.

Os tipos de conhecimento que a ferramenta IM descobre são árvores de decisão binárias, matrizes de classificação neural, agrupamentos, regras de associação, regras temporais, quantis e seqüências similares.

A figura 5.1 apresenta os componentes do cliente e do servidor do IM onde verifica-se que: o acesso aos dados pode ser realizado através de arquivos planos, tabelas ou *views* de BD; os dados de entrada podem ser arquivos planos ou tabelas de BD; a biblioteca de processamento provém o acesso às funções do BD; as bases de pesquisa apresentam um conjunto de metodologias de extração de conhecimento disponíveis na ferramenta; os resultados da extração de conhecimento, as API[11] de resultados e as ferramentas de exportação permitem visualizar os dados extraídos que podem ser exportados para um processamento posterior ou serem utilizados com as ferramentas de visualização.

[11] API - *Application Programming Interface.*

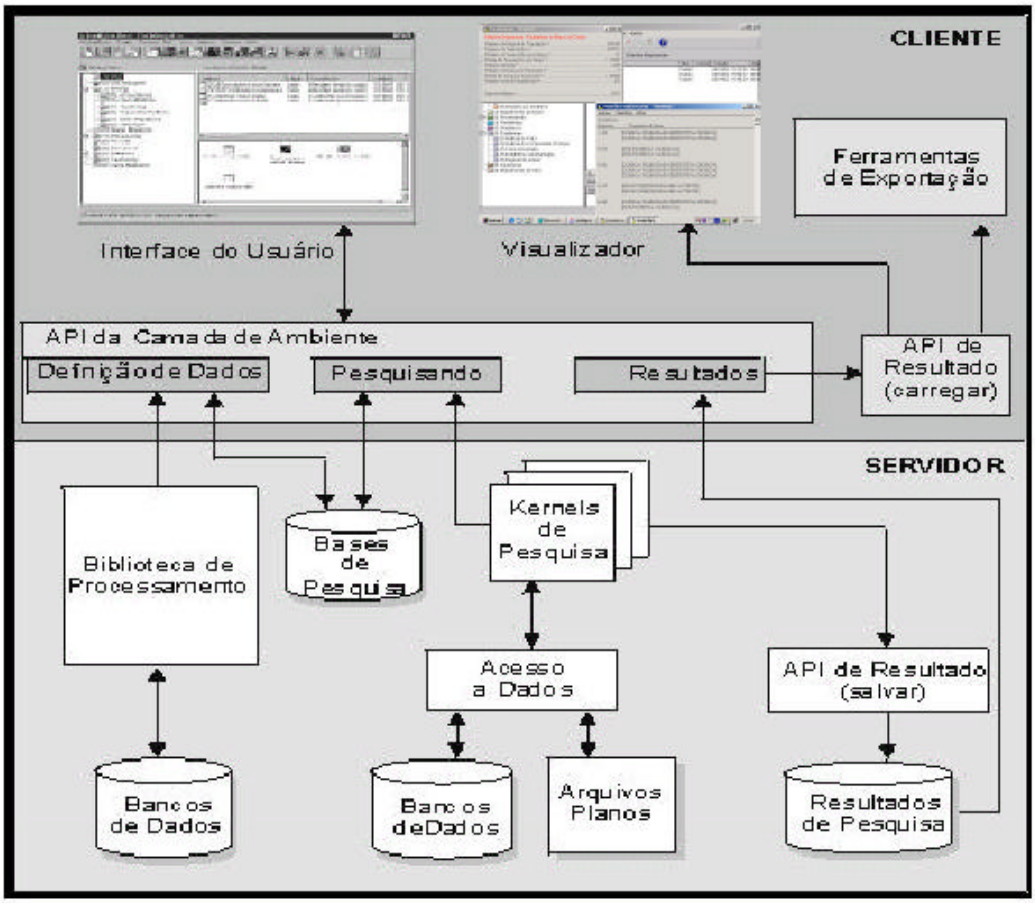

FIGURA 5.1 - Arquitetura do *Intelligent Miner* da IBM ©.

Fonte: IBM, 1999 p. 5 com adaptações.

Devido a MDT ser uma tarefa automática e totalmente efetuada sobre a ferramenta IM, o seu entendimento é melhor observado através da sua aplicação. A partir disto, é apresentado um exemplo da aplicação de MDT com as seguintes finalidades: explicar como é obtido o suporte das seqüências temporais e elucidar a identificação das regras temporais.

5.2 Exemplo Didático

A seguir, é apresentado um pequeno exemplo didático contendo vinte (20) registros para melhorar o entendimento da descoberta de padrões seqüenciais que foi realizada sobre os dados reais da SES.

O suporte das seqüências temporais é obtido pelo número de pacientes que atendem ao padrão temporal dividido pelo número total de pacientes.

Uma regra temporal é detectada a partir de uma seqüência temporal de dados de entrada.

A etapa da seleção dos dados compreendeu a escolha dos campos: pacientes, internação e diagnóstico. Na etapa da limpeza, foram excluídos os registros nulos e inconsistentes. Na ordenação, os pacientes foram agrupados obedecendo as ocorrências de internação efetuadas.

A tabela 5.1 apresenta os vinte (20) registros dos dados de entrada do exemplo com a análise dos seus dados a seguir:

TABELA 5.1 – Arquivo do exemplo didático.

Grupos	Paciente	Internação	Máx. de internações	Diagnóstico	AIH
1	ABEDONI FC08/15/31	01/10/00	1	J449	2155753501
1	ABEGAHY RC03/02/35	07/24/00	1	E232	2299036950
ST 1	ABEGAIL DSP08/09/61	06/21/00	2	F609	2296111488
	ABEGAIL DSP08/09/61	07/18/00		F102	2296808822
1	ABEGAIL MD10/19/94	09/14/00	1	A059	2299067265
1	ABEGAIR BRDS02/13/52	09/14/00	1	J449	2298475905
1	ABEGAIR DAL11/06/46	05/25/00	1	G458	2295566658
ST ST 1	ABEGAIR DSN09/17/35	06/23/00	3	J449	2296993974
ST	ABEGAIR DSN09/17/35	09/06/00		J459	2297737662
ST	ABEGAIR DSN09/17/35	09/28/00		K229	2299135256
1	ABEGAIR DSN09/17/37	04/15/00	1	J449	2294833013
1	ABEGAIR RM10/04/53	05/07/00	1	I200	2294652679
1	ABEGAIR TDM12/08/37	10/05/00	1	K829	2299045496
ST 1	ABEGAY DSS05/13/30	02/14/00	3	J449	2155799745
ST	ABEGAY DSS05/13/30	04/06/00		J449	2294924698
	ABEGAY DSS05/13/30	05/03/00		J449	2295662391
1	ABEGAY NDS04/13/1914	05/04/00	1	G458	2295706688
1	ABEGHAIR DM04/19/1928	06/23/00	1	I499	2295729172
1	ABEL V12/27/31	08/04/00	1	M725	2297057400
1	ABEL ADC01/06/1922	10/20/00	1	K869	2299678909
? 15			20	13	

Os resultados obtidos a partir da MDT efetuada sobre o arquivo da tabela 5.1 foram:

? o número de pacientes diferentes existentes são quinze (15) conforme verificado na coluna grupos;

? o número de internações existentes são vinte (20), porque todas as internações apresentadas são diferentes para os pacientes investigados;

? o número máximo de internações por paciente correspondem a três (03) conforme apresentado nos dois quadros internos a tabela;

? a média de internações por paciente é obtida pelo cálculo do somatório do máximo de internações por grupo que são vinte (20), dividido pelo número de pacientes encontrados que correspondem a quinze (15). Isto resulta em 1,33;

? o número de diagnósticos principais diferentes são treze (13);

? o número total de seqüências temporais (ST) encontradas são sete (07).

De acordo com a tabela 5.2, observa-se que as ST procuram por padrões temporais contíguos e descontíguos de diagnósticos principais entre as internações dos pacientes. Os exemplos de ST contíguas são apresentadas nos conjuntos de elementos 1, 2, 3, 5, 6 e 7. Um exemplo de ST descontínua está representada no conjunto de elementos 4.

TABELA 5.2 - Transformação das seqüências do arquivo de exemplo didático da tabela 5.1.

Conjunto de elementos	**Seqüências**	**D_t**	**D_t presente no conjunto de elementos**
1	<F609,F102>	<F609><F102>	1
2	<J449,J449>	<J449> <J449>	2,5
3	<J449,J459>	<J449> <J459>	3,7
4	<J449,K229>	<J449> <K229>	4
5	<J449,J449,J449>	<J449> <J449> <J449>	5
6	<J459,K229>	<J459> <K229>	6,7
7	<J449,J449,K229>	<J449> <J459> <K229>	7

? o suporte mínimo de 5% corresponde ao percentual mínimo exigido para uma ST apresentar um padrão temporal.

A seguir, o padrão temporal descoberto é representado através das regras temporais extraídas com a MDT. A figura 5.2 apresenta as regras temporais (RT) descobertas a partir do arquivo da tabela 5.1.

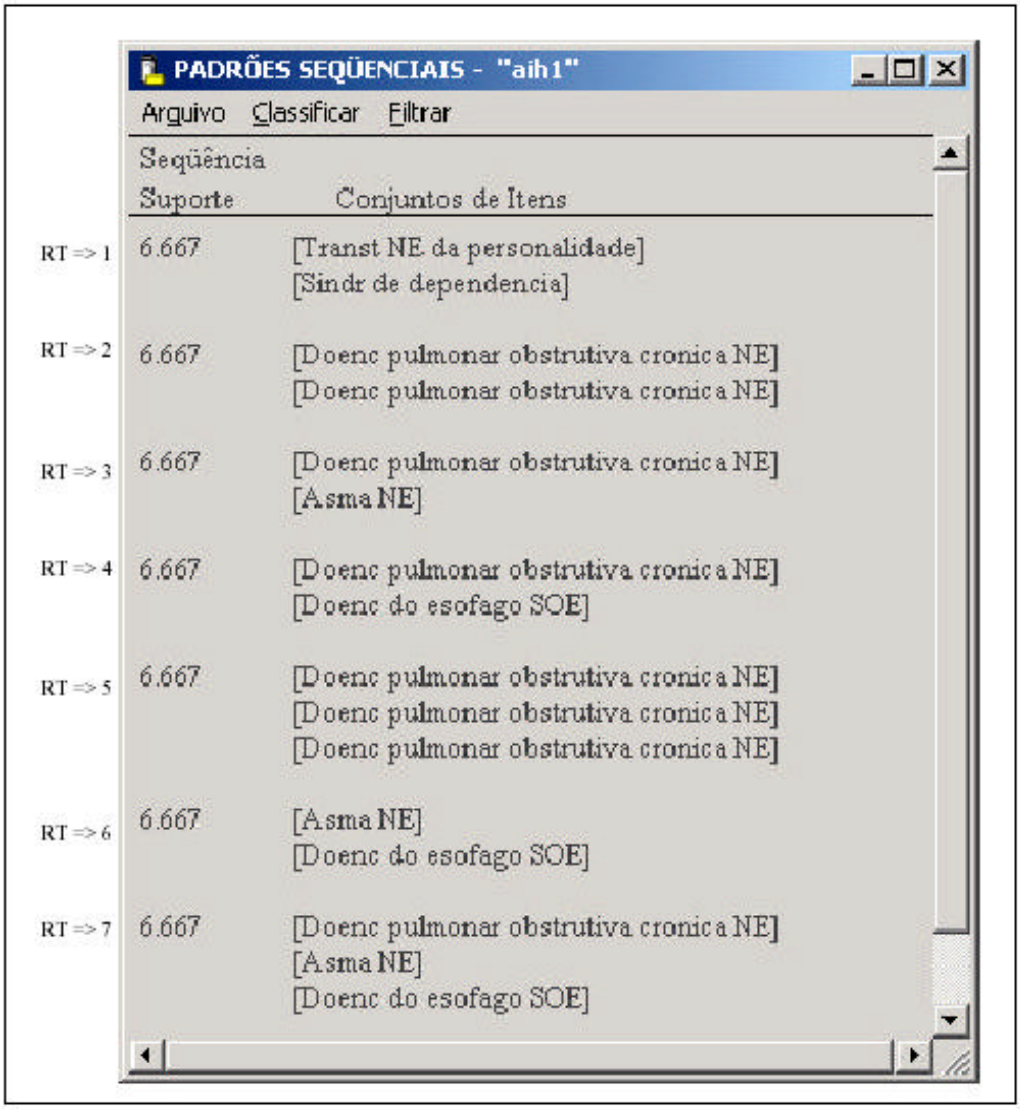

FIGURA 5.2 – Regras temporais obtidas a partir do exemplo didático da tabela 5.1.

A RT 1 indica que o diagnóstico principal de Transtornos Não Específicos (NE) da personalidade é seguido pelo diagnóstico principal de Síndrome de dependência com 6,667% de suporte sobre os quinze (15) pacientes investigados. Este suporte referente-se a presença de um caso com este padrão seqüencial.

A RT 2 indica que o diagnóstico principal de Doença Pulmonar Obstrutiva Crônica Não Especificada (NE) é seguido pelo mesmo diagnóstico principal com 6,667% de suporte. Este suporte referente-se a presença de um caso com este padrão seqüencial.

A RT 3 indica que o diagnóstico principal de Doença Pulmonar Obstrutiva Crônica NE é seguido pelo diagnóstico principal de Asma NE com 6,667% de suporte. Este suporte referente-se a presença de um caso com este padrão seqüencial.

A RT 4 indica que o diagnóstico principal de Doença Pulmonar Obstrutiva Crônica NE é seguido pelo o diagnóstico principal de Doença do Esôfago Sem Origem Específica (SOE) com 6,667% de suporte. Este suporte referente-se a presença de um caso com este padrão seqüencial.

A RT 5 indica que o diagnóstico principal de Doença Pulmonar Obstrutiva Crônica NE é seguido pelo mesmo o diagnóstico principal com 6,667% de suporte. Este suporte referente-se a presença de um caso com este padrão seqüencial.

A RT 6 indica que o o diagnóstico principal de Asma NE é seguido pelo o diagnóstico principal de Doença do Esôfago Sem Origem Específica (SOE) com 6,667% de suporte. Este suporte referente-se a presença de um caso com este padrão seqüencial.

A RT 7 indica que o diagnóstico principal de Doença Pulmonar Obstrutiva Crônica NE é seguido pelo diagnóstico principal de Asma NE e, após, pelo diagnóstico principal de Doença do Esôfago Sem Origem Específica (SOE) com 6,667% de suporte. Este suporte referente-se a presença de um caso com este padrão seqüencial.

A seguir, os experimentos são detalhados.

5.3 Experimento com as AIH

O primeiro experimento refere-se a procura por padrões seqüenciais que determinem uma seqüência de comportamentos temporais nas AIH.

5.3.1 Pré-processamento

A fase do pré-processamento compreendeu as etapas de seleção, limpeza, ordenação e transformação dos dados descritos, a seguir:

5.3.1.1 Seleção

A etapa da seleção constituiu-se na escolha dos arquivos de movimento das AIH do ano de 2000. A escolha dos arquivos de movimento das AIH foi feita porque estes arquivos contém registros correspondentes a caracterização das internações hospitalares como a identificação do paciente internado, a ocorrência temporal e o procedimento médico realizado que permitem comprovar tradicionais comportamentos nos tratamentos das AIH, detectar situações anômalas e acompanhar a evolução das doenças.

A partir da escolha do arquivo de movimento das AIH, foi efetuada a junção dos arquivos mensais em um único arquivo correspondente ao movimento das AIH do ano de 2000.

Após a seleção dos arquivos de dados, foi efetuada a escolha dos atributos: principal, temporal e elementar do arquivo de movimento das AIH de 2000.

Os pacientes foram escolhidos como o atributo principal porque eles constituem-se em um conjunto de registros que compreendem diversas ocorrências de internações.

As internações foram selecionadas como a ocorrência temporal por apresentar vários procedimentos médicos realizados.

Os procedimentos médicos foram selecionados como atributo elementar.

Os diagnósticos principais foram escolhidos como atributo elementar para uma segunda MDT sobre os dados das AIH.

5.3.1.2 Limpeza

A etapa da limpeza objetivou aumentar a qualidade dos dados selecionados. A princípio, a identificação para pacientes utilizada foi o cadastro nacional de pessoas físicas. Entretanto, observou-se que o arquivo de movimento das AIH apresentava muitos registros que não dispunham desta informação ou continham a mesma informação para mais de um paciente conforme a figura 5.3 a seguir:

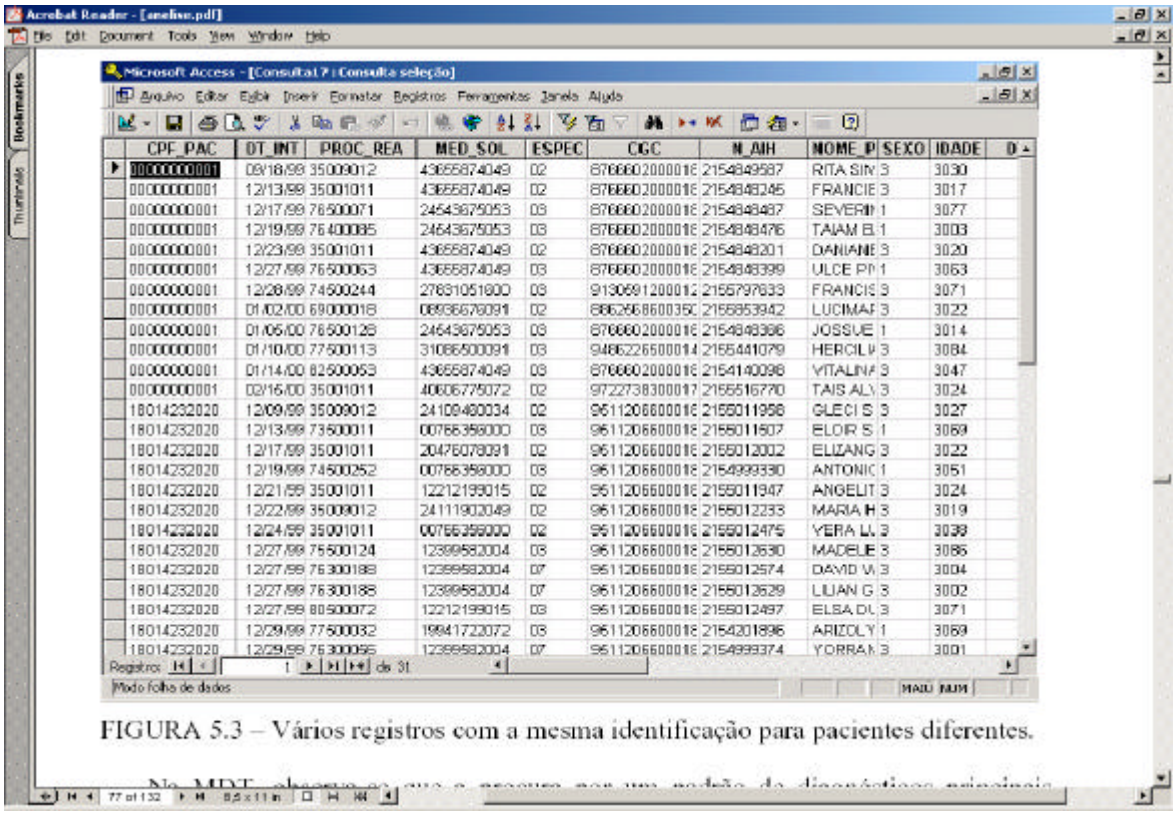

FIGURA 5.3 – Vários registros com a mesma identificação para pacientes diferentes.

Na MDT, observa-se que a procura por um padrão de diagnósticos principais objetiva seguir uma seqüência efetuada por um paciente e, após isto, verificar a repetição desta seqüência em outros pacientes. A partir disto, a escolha do cadastro nacional de pessoas físicas como atributo principal torna a MDT impraticável, visto que, diferentes pacientes foram cadastrados sob a mesma codificação, o que impossibilita acompanhar o tratamento de um paciente. Alguns exemplos que erroneamente poder-se-ia obter seriam: pacientes do sexo masculino apresentando diagnósticos de parto, pacientes sofrendo seguidos procedimentos realizados de retiradas de órgãos vitais, e por conseguinte, pacientes que sofreram óbito continuariam a existir, entre outros.

A solução adotada constituiu-se da exclusão dos registros que apresentavam estas irregularidades. Entretanto, observou-se que o volume do BD resultante foi significativamente reduzido, pois o BD da SES que apresentava inicialmente 574.352 registros, após a limpeza de registros que apresentavam identificação de pacientes inconsistentes, o número de registros do BD diminuiu para 9.968 registros. Esta significativa redução do volume dos dados de entrada refletiu na qualidade das regras temporais extraídas.

Desta maneira, o problema da grande perda de informação devido a escolha impraticável do atributo principal foi tratada com a mudança da seleção do atributo principal para o nome do paciente. Este campo foi concatenado com a data de nascimento do paciente para evitar problemas de homonímia. Assim, através da escolha do nome do paciente como atributo principal é possível resolver o problema de acompanhamento do tratamento dos pacientes. Todavia, pacientes idênticos com variação na digitação do nome constituir-se-ão em pacientes diferentes.

Adicionalmente, observou-se que existem AIH com o mesmo número, em diferentes períodos, devido as reinternações e transferências de pacientes de outros hospitais que geram duplas ou triplas de registros de um mesmo paciente. Também, verificou-se que em alguns casos como, por exemplo, psiquiátricos e fora de possibilidade terapêutica, após 180 diárias, e no tratamento em reabilitação, após 45 diárias, que é emitida uma nova numeração de AIH. Além disto, uma única AIH é dividida em várias outras AIH de modo fraudulento. Assim, foram excluídos do arquivo de movimento das AIH os registros correspondentes as AIH bloqueadas conforme com o arquivo de controle de AIH.

O arquivo de movimento de AIH foi associado ao arquivo de controle de AIH com equivalência dos registros que apresentavam a mesma numeração de AIH, hospital e procedimento realizado. As AIH bloqueadas corresponderam aqueles registros com *status* de permanece bloqueada, libera aih com codigo novo e sem resposta do auditor. Os registros com *status* de libera aih com mesmo codigo foram mantidos no arquivo de movimento de AIH.

O arquivo resultante apresentou 568.833 registros.

5.3.1.3 Ordenação

Após a limpeza dos atributos pacientes, internações e procedimentos realizados ou diagnósticos principais, foi efetuada a ordenação das seqüências existentes. Para tanto, o arquivo foi ordenado pelos pacientes seguido das suas internações.

5.3.1.4 Transformação

Nesta etapa, os dados ordenados são transformados em seqüências temporais. Esta etapa foi efetuada no *Intelligent Miner* da IBM ©, devido ao grande volume de dados a ser transformados e por esta ser uma tarefa totalmente automatizada nesta ferramenta.

5.3.2 MDT

A fase da MDT foi efetuada na ferramenta *Intelligent Miner* da IBM ©. Nesta fase, foram utilizados os mapeamentos de nomes para atribuir nomes mais signicativos aos valores do atributo elementar. Para tanto, os arquivos de descrição dos procedimentos realizados e o Cadastro Internacional de Doenças (CID) foram utilizados para tornar mais representativo o conhecimento extraído a partir dos procedimentos realizados e dos diagnósticos principais utilizados.

O tempo decorrido durante a fase de MDT para processar os 568.833 registros foi de cinqüenta e cinco (55) segundos.

5.3.3 Pós-Processamento

Após a MDT, foram extraídos os resultados das descobertas efetuadas sobre os dasos reais. O conhecimento obtido foi representado na forma de regras temporais.

5.3.3.1 Resultados

Na primeira MDT, os resultados procuraram apresentar o conhecimento descoberto em relação aos procedimentos realizados efetuados nas internações de pacientes no período de um ano.

A figura 5.4 apresenta os dados de saída dos procedimentos realizados.

O número de pacientes diferentes encontrados no BD foi 459.189.

O número de internações efetuadas foi 550.878.

O número máximo de internações por paciente foi 25.

A média de internações por grupo de pacientes foi aproximadamente 1,2.

O número de procedimentos realizados diferentes foi 1.668.

O número máximo de procedimentos realizados por internação foi 3.

O número total de seqüências temporais extraídas foi 258.

O suporte mínimo utilizado foi de 0,014%.

FIGURA 5.4 – Dados de saída dos procedimentos realizados.

Foram descobertas várias regras temporais para o problema do procedimento realizado conforme o anexo I. Todavia, a seguir, são apresentadas três regras temporais obtidas devido a sua importâcia de interpretação, como:

1. A primeira regra temporal extraída apresenta que o procedimento realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo mesmo procedimento realizado com suporte de 1, 103 % entre os 459.189 pacientes encontrados no BD. Isto equivale a 5.064 casos investigados.

A interpretação da regra temporal 1 está diretamente relacionada ao caráter crônico da patologia de DBPOC que faz com que um paciente sofra seguidas internações do mesmo procedimento realizado. Adicionalmente, observa-se que esta regra temporal apresenta o comportamento do procedimento realizado mais freqüente no Estado de acordo com os auditores da SES.

2. Uma regra temporal apresenta o procedimento realizado de Colecistite Aguda (inflamação na vesícula) seguido do procedimento realizado de Colecistectomia (retirada da vesícula) com suporte de 0,088 % entre os 459.189 pacientes do BD da SES. Isto equivale a 404 casos investigados.

A interpretação para a regra temporal 2 consiste em um padrão temporal de comportamento para o tratamento da doença de Colecistite Aguda. Ela constitui-se em

um indicativo para os gestores da SES. Uma doença como esta, assim que diagnosticada, deve ser encaminhada a um procedimento de cirurgia eletiva de Colecistectomia, pois a seguida internação só acarreta sofrimento desnecessário ao paciente e ônus ao sistema.

3. Uma regra temporal apresenta o procedimento realizado de Colecistite Aguda seguido do mesmo procedimento realizado com suporte de 0,031 % entre os 459.189 pacientes do BD da SES. Isto equivale a 142 casos investigados.

Na interpretação para a regra temporal 3, observa-se que esta regra temporal tem um comportamento diferente do padrão temporal já encontrado na regra temporal 2 e deve ser analisada pelos especialistas da SES.

Na segunda MDT efetuada sobre estes dados, os resultados obtidos procuram descobrir seqüências de comportamentos nos diagnósticos principais de pacientes no período de um ano. A figura 5.5 apresenta os dados de saída dos diagnósticos principais.

O número de pacientes diferentes encontrados no BD foi 459.189.

O número de internações efetuadas foi 550.878.

O número máximo de internações por paciente foi 25.

A média de internações por grupo de pacientes foi aproximadamente 1,2.

O número de diagnósticos principais diferentes foi 3.285

O número máximo de diagnósticos principais por internação foi 3.

O número total de seqüências temporais extraídas foi 240.

O suporte mínimo utilizado foi de 0,013%.

FIGURA 5.5 – Dados de saída dos diagnósticos principais.

Foram descobertas várias regras temporais para o problema do diagnóstico principal conforme o anexo I. A seguir, são apresentadas três regras temporais obtidas devido a sua importâcia de interpretação, como:

1. Uma regra temporal apresenta que o diagnóstico principal de Parto por Cesariana NE é seguido pelo diagnóstico principal de Infecção da Incisão

Cirúrgica com suporte de 0, 021 % entre os 459.189 pacientes encontrados. Isto equivale a 96 casos investigados.

A interpretação da regra temporal 1 comprova a ocorrência de internação por infecções e complicações relacionadas ao parto.

2. Uma regra temporal com o diagnóstico principal de Sessão de quimioterapia p/ neoplasias seguido do mesmo diagnóstico com suporte de 0,018 entre os 459.189 pacientes encontrados no BD. Isto equivale a 82 casos investigados.

A interpretação da regra temporal 2 refere-se ao tratamento continuado do câncer.

3. Uma regra temporal com o diagnóstico principal de Insuficiência Cardíaca NE seguido do diagnóstico de Edema Pulmonar NE com suporte de 0,018 % entre os 459.189 pacientes encontrados no BD. Isto equivale a 82 casos investigados.

5.3.3.2 Validação dos Resultados

A seguir, são apresentadas a análise segundo os especialistas da SES a partir do conhecimento extraído.

De acordo com os especialistas da SES, mais de 70 % das internações efetuadas dizem respeito ao baixo nível sócio-econômico dos pacientes, falta de políticas de prevenção de doenças, insuficiência de atenção básica, deficiência no acompanhamento dos pacientes no posto de saúde e carência de saúde básica da população.

A tabela 5.3 apresenta a validação dos padrões temporais extraídos e classifica este conhecimento como já conhecido ou desconhecido. As observações são relativas a utilização de procedimentos realizados ou diagnósticos principais referentes ao mesmo grupo de patologias. Isto significa, que a regra temporal descoberta não agrega conhecimento algum a base de conhecimento, visto que, o procedimento realizado ou o diagnóstico principal utilizado pode ser substituído por qualquer outro pertencente a mesma patologia.

A seguir, a tabela 5.3 apresenta a validação dos padrões temporais extraídos em relação ao comportamento dos procedimentos realizados nas AIH de 2000.

TABELA 5.3 – Validação dos padrões seqüenciais extraídos do comportamento dos procedimentos realizados pelos pacientes nas AIH de 2000 conforme a SES.

	Comportamento dos procedimentos realizados pelos pacientes nas AIH de 2000	**Validação do conhecimento**		
		Conhecido	**Desconhecido**	**Obs.**
1	Um procedimento realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo mesmo procedimento até nove (09) internações em 82 casos analisados.	X		X
2	Um procedimento realizado de Insuficiência Cardíaca é seguido pelo procedimento de Broncopneumonia em 192 casos investigados.	X		X
3	Um procedimento realizado de Insuficiência Cardíaca é seguido pelo procedimento de Acidente Vaso-Cerebral (AVC) Agudo em 280 casos.		X	
4	Um procedimento realizado de Acidente Vaso-Cerebral (AVC) Agudo é seguido pelo procedimento de Insuficiência Cardíaca em 275 casos.		X	
5	Um procedimento realizado de Insuficiência Cardíaca é seguido pelo procedimento Doença Pulmonar Obstrutiva Crônica (DBPOC) em 932 casos.	X		X
6	Um procedimento realizado de Crise Asmática é seguido pelo procedimento de Broncopneumonia em Lactente em 192 casos analisados.	X		
7	Um procedimento realizado de Acidente Vaso-Cerebral (AVC) Agudo é seguido pelo mesmo procedimento em 821 casos.	X		X
8	Um procedimento realizado de Diabetes Sacarino é seguido pelo mesmo procedimento até três (03) internações em 197 casos investigados.	X		X
9	Um procedimento realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo procedimento de Insuficiência Respiratória Aguda em 720 casos.	X		X

A seguir, a tabela 5.4 apresenta a validação dos padrões seqüenciais extraídos em relação ao comportamento dos diagnósticos principais dos pacientes nas AIH de 2000.

TABELA 5.4 – Validação dos padrões seqüenciais extraídos do comportamento dos diagnósticos principais dos pacientes nas AIH de 2000 conforme a SES.

	Comportamento dos diagnósticos principais dos pacientes nas AIH de 2000	**Validação do conhecimento**		
		Conhecido	**Desconhecido**	**Obs.**
1	Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo mesmo diagnóstico até nove (09) internações.	X		X
2	Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo mesmo diagnóstico até seis (06) internações.		X	
3	Um diagnóstico principal de Afecção Respiratória do recém-nascido é seguido pelo mesmo diagnóstico.	X		X
4	Um diagnóstico principal de Broncopneumonia NE é seguido pelo mesmo diagnóstico.	X		X
5	Um diagnóstico principal de Asma NE é seguido pelo mesmo diagnóstico.	X		X
6	Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo diagnóstico de Insuficiência Cardíaca NE.	X		
7	Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo diagnóstico de Doença Pulmonar Obstrutiva Crônica (DBPOC).	X		X
8	Um diagnóstico principal de Outras Infecções intestinais específicas é seguido pelo diagnóstico de Broncopneumonia NE.		X	
9	Um diagnóstico principal de Afecções respiratórias do recém-nascido é seguido pelo diagnóstico de Asma NE.	X		X
10	Um diagnóstico principal de Outros acidentes isquêmicos cerebrais é seguido pelo diagnóstico de Doença Pulmonar Obstrutiva Crônica (DBPOC).		X	

5.4 Experimento com as Doenças de Agravos Notificáveis

O segundo experimento refere-se a procura por padrões seqüenciais que determinem uma seqüência de comportamentos temporais nas doenças de agravos notificáveis.

5.4.1 Pré-processamento

A fase do pré-processamento compreendeu as etapas de seleção, limpeza, ordenação e transformação dos dados descritos, a seguir:

5.4.1.1 Seleção

A etapa da seleção de dados constituiu-se na escolha do arquivo de ocorrências de doenças de agravos notificáveis do ano de 1999. A escolha deste arquivo foi feita devido as informações de identificação do paciente, ocorrência temporal e caracterização da doença de agravo notificável que permitem fazer o acompanhamento temporal das doenças investigadas.

Os pacientes foram escolhidos como o atributo principal porque eles constituem–se em um conjunto de registros que compreendem diversas ocorrências de atendimento. As datas dos atendimentos foram selecionadas como a ocorrência temporal e as doenças foram escolhidas como atributo elementar.

5.4.1.2 Limpeza

Na etapa da limpeza, foram excluídos registros sem identicação de paciente. A BD com 21.262 registros foi reduzida para 18.150 registros.

5.4.1.3 Ordenação

Após a limpeza dos atributos pacientes, datas de atendimento e doenças diagnosticadas, foi efetuada a ordenação das seqüências existentes. Para tanto, o arquivo foi ordenado por pacientes seguido das datas de atendimento.

5.4.1.4 Transformação

A etapa da transformação das seqüencias temporais encontradas foi totalmente efetuada na ferramenta *Intelligent Miner* da IBM ©.

5.4.2 MDT

A fase da MDT foi executada na ferramenta *Intelligent Miner* da IBM ©. Esta fase consituiu do mapeamento de nome dos valores das doenças de agravos notificáveis conforme o CID para tornar o conhecimento extraído mais significativo.

O tempo decorrido para processar os 18.150 registros foi de três (03) segundos.

5.4.3 Pós-Processamento

A seguir, são apresentados os resultados da MDT que procuraram descobrir seqüências de comportamentos sobre as doenças de agravos notificáveis de pacientes no período de um ano.

A figura 5.6 apresenta os dados de saída das doenças de agravos notificáveis.

O número de pacientes diferentes encontrados no BD foi 17.677.

O número de atendimentos efetuados foi 17.916.

O número máximo de atendimentos por paciente foi 3.

A média de atendimentos por grupo de pacientes foi aproximadamente 1,0.

O número de doenças diferentes foi 42.

O número máximo de doenças por atendimento foi 3.

O número total de seqüências temporais extraídas foi 88.

O suporte mínimo utilizado foi de 0,003%.

FIGURA 5.6 – Dados de saída das doenças de agravos notificáveis.

O total de regras temporais de agravos notificáveis descobertas encontram-se no anexo I. A seguir, são apresentadas quatro regras temporais obtidas devido ao elevado número de casos observados.

Uma regra temporal onde a doença de agravo notificável Hepatite por vírus é seguida pela mesma doença com suporte de 0.317 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 56 casos.

Uma regra temporal onde a doença de agravo notificável de Leptospirose é seguida pela doença Hepatite por vírus com 0,119% de suporte entre os 17.677 pacientes existentes no BD. Isto equivale a 21 casos investigados.

Uma regra temporal em que a doença de agravo notificável Meningite de causa não especificada é seguida pela doença Varicela com suporte de 0.028 % entre os 17.677 pacientes existentes no BD. Isto equivale a 7 casos investigados.

A doença de agravo notificável Meningite de causa não especificada é seguida pela doença Outros exantemas por vírus com suporte de 0.028 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 4 casos.

5.5 Considerações Finais

Neste capítulo, procedeu-se a adaptação da técnica dos padrões seqüenciais para a aplicação do movimento das AIH e das doenças de agravos notificáveis onde obteve-se a extração implícita, não trivial, nova e potencialmente utilizável da informação temporal dos dados.

Os padrões temporais extraídos através da MDT procuraram atender aos problemas dos dois experimentos apresentados. O primeiro experimento, procurou encontrar padrões seqüenciais de comportamento sobre os procedimentos realizados e os diagnósticos principais efetuados pelos pacientes nas AIH de 2000. O segundo experimento, objetivou encontrar padrões seqüenciais de comportamento sobre as doenças de agravos notificáveis.

Analisando-se os dados disponíveis, os algoritmos de MDT puderam encontrar padrões passíveis de serem utilizados para predizer quais procedimentos realizados ou diagnósticos principais oferecem seqüência nas AIH e quais devem ser considerados na autorização de uma internação hospitalar. Adicionalmente, os padrões seqüenciais extraídos dos dados do SINAN proporcionaram a descoberta do conhecimento a respeito do comportamentos das doenças de agravos notificáveis.

A partir, dos padrões seqüenciais extraídos verificou-se que eles representam um conhecimento valioso acerca do BD, apresentando o mundo real armazenado em seus dados.

6 Conclusões

Hoje em dia, observa-se que o crescimento do volume e do número de BD existentes excedem a capacidade humana de analisar seus dados com simples consultas ou pesquisas e, desta forma, não podem analisar o seu conteúdo quanto a algum conhecimento implícito importante existentes nos BD. A mineração de dados temporais permite que os computadores executem o trabalho de explorar as imensas quantidades de dados armazenados procurando por padrões significativos.

O presente trabalho apresentou sucintamente as etapas do processo de DCBD, enfatizando a etapa de mineração de dados, visto que a abordagem central do trabalho foi a mineração de dados temporais. Neste contexto, foram considerados os aspectos temporais dos dados em um BD, as diferentes tarefas existentes, as principais técnicas empregadas e os algoritmos utilizados na mineração de dados. A partir disto, foi apresentada uma metodologia para efetuar-se a mineração de dados temporais, que foi validada com dados reais provindos da Secretaria Estadual da Saúde do Rio Grande do Sul.

O resultado esperado ao final deste trabalho foi verificado ao utilizar as técnicas de MD considerando os aspectos temporais.

A MDT sempre tem como um de seus resultados, novas perguntas. Para possibilitar tal iteração, o tomador de decisão precisa ter acesso aos recursos da exploração de dados.

Entretanto, acredita-se que os resultados alcançados podem ser vistos como um resultado preliminar da tarefa de DCBD realizada, e constituir-se como parte de um processo de DCBD em contínuo aprimoramento e evolução.

6.1 Contribuições

Dentre as principais contribuições deste trabalho pode-se citar:

Um estudo aprofundado sobre os aspectos temporais no processo de DCBD, enfatizando as dificuldades nas etapas inicias deste processo.

A proposição de uma metodologia para a mineração de dados temporais que engloba todo o processo de DCBD.

Uma validação em dados reais que continham imprecisões e contradições.

6.2 Perspectivas Futuras

O seguimento deste trabalho de DCBD deve realizar novas MDT em trabalhos posteriores sobre esta mesma base pré-processada objetivando a obtenção de regras temporais com maiores níveis de suporte para as seqüências.

Ao longo deste trabalho, pôde-se perceber que as técnicas de DCBD realmente são excelentes para descobrir conhecimento novo e oculto em grandes BD. Durante a etapa de pré-processamento no processo de DCBD se confirmou que ela realmente demanda o maior tempo de um trabalho de DCBD. Acrescenta-se que um tópico que deve ser alvo de estudos futuros dentro da DCBD constitui-se em estudar e pesquisar um método de validação de regras e conhecimento descoberto, no pós-processamento, pois esta etapa é de fundamental importância e muito pouco foi encontrado, sobre a mesma, na bibliografia.

Este trabalho não se justifica apenas pela aplicação de técnicas de aprendizado para análise de dados sobre o aspecto temporal, mas a organização necessita aprender automaticamente por si própria. O maior problema que fascina a sociedade da informação é a superabundância de dados. No futuro, todas as organizações deverão encontrar o seu próprio caminho para julgar e descobrir conhecimento, e a atividade da extração do conhecimento prova ser uma parte crucialmente importante neste processo.

Anexo 1 Regras Temporais

Os resultados obtidos com as MDT efetuadas foram:

1 Regras Temporais de Procedimentos Realizados

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo mesmo procedimento com 1,103 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Cardíaca é seguido pelo mesmo procedimento com 0,659 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo mesmo procedimento em três internações consecutivas com 0,421 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Broncopneumonia em Lactente é seguido pelo mesmo procedimento com 0,259 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo procedimento de Insuficiência Cardíaca com 0,245 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo mesmo procedimento em quatro internações consecutivas com 0,215 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Cardíaca é seguido pelo procedimento Doença Pulmonar Obstrutiva Crônica (DPBOC) com 0,203 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Cardíaca é seguido pelo mesmo procedimento em três internações com 0,200 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Acidente Vaso-Cerebral (AVC) Agudo é seguido pelo mesmo procedimento com 0,179 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Diabetes Sacarino é seguido pelo mesmo procedimento com 0,173 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo procedimento de Insuficiência Respiratória Aguda com 0,157 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Crise Asmática é seguido pelo mesmo procedimento com 0,144 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Intercorrências Clínicas de Paciente Oncológico é seguido pelo mesmo procedimento com 0,131 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Respiratória Aguda é seguido pelo procedimento de Doença Pulmonar Obstrutiva Crônica (DPBOC)com 0,124 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo mesmo procedimento em cinco internações com 0,120 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Respiratória Aguda é seguido pelo mesmo procedimento com 0,115 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Broncopneumonia é seguido pelo procedimento de Doença Pulmonar Obstrutiva Crônica (DPBOC) com 0,107 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Broncopneumonia é seguido pelo mesmo procedimento com 0,107 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Crise Asmática é seguido pelo mesmo procedimento com 0,107 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Broncopneumonia é seguido pelo mesmo procedimento com 0,101 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo procedimento de Broncopneumonia com 0,100 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo procedimento de Crise Asmática com 0,096 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Renal Crônica Acidose Metabólica é seguido pelo mesmo procedimento de com 0,092% de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Enfise Pulmonar é seguido pelo procedimento de Doença Pulmonar Obstrutiva Crônica (DPBOC) com 0,091 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Trabalho de Parto Prematuro é seguido pelo procedimento de Parto Normal com 0,090 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar 0,089 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Colecistite Aguda é seguido pelo procedimento de Colecistectomia com 0,088 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo procedimento de Crise Asmática com 0,088 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Pielonefrite é seguido pelo mesmo procedimento com 0,081 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Pielonefrite é seguido pelo procedimento de Parto Normal com 0,081 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Entero Infecções em Lactente é seguido pelo procedimento de Broncopneumonia em Lactente com 0,107 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Enfisema Pulmonar é seguido pelo mesmo procedimento de com 0,076 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo mesmo procedimento em seis internações com 0,072 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Cardíaca é seguido pelo mesmo procedimento em quatro internações com 0,072 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Tratamento em Psiquiatria em Hospital Psiquiátrico é seguido pelo mesmo procedimento com 0,070 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Hemorragias Digestivas é seguido pelo mesmo procedimento com 0,068 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Coronariana Aguda é seguido pelo mesmo procedimento com 0,066 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo procedimento e depois por Insuficiência Cardíaca de com 0,065 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Cirrose Hepática é seguido pelo mesmo procedimento com 0,063 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Entero Infecções em Lactente é seguido pelo mesmo procedimento com 0,062 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Cardíaca é seguido pelo procedimento de Acidente Vaso-Cerebral (AVC) Agudo com 0,061 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Acidente Vaso-Cerebral (AVC) Agudo é seguido pelo procedimento de Insuficiência Cardíaca com 0,060 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Trabalho de Parto Prematuro é seguido pelo procedimento de Cesariana com 0,060 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Crise Hipertensiva é seguido pelo mesmo procedimento de com 0,059 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Broncopneumonia em Lactente é seguido pelo procedimento de Crise Asmática com 0,058 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo mesmo procedimento e depois do procedimento de Insuficiência Respiratória Aguda com 0,056 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Cardíaca é seguido pelo procedimento de Insuficiêcia Respiratória Aguda com 0,056 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo procedimento de Insuficiência Cardíaca por três internações com 0,055 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Acidente Vaso-Cerebral (AVC) Agudo é seguido pelo procedimento de Doença Pulmonar Obstrutiva Crônica (DBPOC) com 0,054 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Tratamento em Psiquiatria em Hospital Geral é seguido pelo mesmo procedimento com 0,053 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Diabetes Sacarino é seguido pelo procedimento de Insuficiência Cardíaca com 0,052 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Trabalho de Parto Prematuro é seguido pelo procedimento de Parto Normal com Atendimento do Recém Nascido na Sala de Parto com 0,052 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de é seguido pelo procedimento de Insuficiência Cardíaca depois do procedimento de Doença Pulmonar Obstrutiva Crônica (DBPOC) com 0,052 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Pneumonia em Lactente é seguido pelo procedimento de Entero Infecções em Lactente com 0,051 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Cardíaca é seguido pelo procedimento de de Doença Pulmonar Obstrutiva Crônica (DBPOC) em duas internações com 0,049 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Coronariana Aguda é seguido pelo procedimento de Insuficiência Cardíaca com 0,049 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Cardíaca é seguido pelo procedimento de Insuficiência Coronariana Aguda com 0,048 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Broncopneumonia é seguido pelo procedimento de Insuficiência Cardíaca com 0,047 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo procedimento de Insuficiência Respiratória Aguda e depoias um procedimento de Doença Pulmonar Obstrutiva Crônica (DPBOC) com 0,046 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo mesmo procedimento em sete internações com 0,046 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Doença Pulmonar Obstrutiva Crônica (DPBOC) é seguido pelo procedimento de AVC Agudo com 0,045 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Respiratória Aguda é seguido pelo procedimento de Insuficiência Cardíaca com 0,045 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Broncopneumonia é seguido pelo procedimento de Crise Asmática com 0,045 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Respiratória Aguda é seguido pelo procedimento de Insuficiêncai Cardíaca com 0,045 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Broncopneumonia é seguido pelo procedimento de Crise Asmática com 0,045 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Desnutrição (Clínica Médica) seguido pelo mesmo procedimento com 0,044 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Respiratória Aguda é seguido pelo procedimento Doença Pulmonar Obstrutiva Crônica (DPBOC) consecutiva com 0,044 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Diabetes Sacarino é seguido pelo mesmo procedimento em mais duas internações com 0,043 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Pneumonia do Lactente é seguido pelo mesmo procedimento com 0,043 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Cardíaca é seguido pelo mesmo procedimento e depois por DBPOC com 0,043 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Broncopneumonia em Lactente é seguido pelo procedimento de Pneumonia do Lactente com 0,043 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Broncopneumonia em Lactente é seguido pelo procedimento em três internações com 0,043 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Cardíaca é seguido pelo procedimento de Diabetes Sacarino com 0,042 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Respiratória Aguda é seguido pelo mesmo procedimento com 0,042 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Crise Asmática é seguido pelo procedimento de Broncopneumonia em Lactente com 0,042 % de suporte desta regra temporal atender aos 459.189 pacientes.

Um procedimento médico realizado de Insuficiência Cardíaca é seguido pelo procedimento de Broncopneumonia com 0,042 % de suporte desta regra temporal atender aos 459.189 pacientes.

2 Regras Temporais de Diagnósticos Principais

Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo mesmo diagnóstico com 1,040 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 4.775 casos.

Um diagnóstico principal de Insuficiência Cardíaca Não Especificada (NE) é seguido pelo mesmo diagnóstico com 0,613 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 2.814 casos.

Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo mesmo diagnóstico em três internações com 0,392 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 1.800 casos.

Um diagnóstico principal de Afeccção Respiratória do recém-nascido é seguido pelo mesmo diagnóstico com 0,269 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 1.235 casos.

Um diagnóstico principal de Broncopneumonia NE é seguido pelo mesmo diagnóstico com 0,265 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 1.216 casos.

Um diagnóstico principal de Asma NE é seguido pelo mesmo diagnóstico com 0,236 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 1.083 casos.

Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo diagnóstico de Insuficiência Cardíaca NE com 0,220 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 1.010 casos.

Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo mesmo diagnóstico em quatro internações com 0,197 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 904 casos.

Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo mesmo diagnóstico em três internações com 0,185 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 849 casos.

Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo diagnóstico de Doença Pulmonar Obstrutiva Crônica (DBPOC) com 0,183 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 840 casos.

Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo diagnóstico de Insuficiência respiratória aguda com 0,149 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 684 casos.

Um diagnóstico principal de Outros Acidentes Isquêmicos Cerebrais é seguido pelo mesmo diagnóstico com 0,144 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 661 casos.

Um diagnóstico principal de Com complicações NE é seguido pelo mesmo diagnóstico com 0,131 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 601 casos.

Um diagnóstico principal de Outras Infecções Intestinais Específicas é seguido pelo mesmo diagnóstico com 0,125 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 573 casos.

Um diagnóstico principal de Infecção Respiratória Aguda é seguido pelo mesmo diagnóstico com 0,119 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 546 casos.

Um diagnóstico principal de Infecção Respiratória Aguda é seguido pelo diagnóstico de DBPOC com 0,118 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 541 casos.

Um diagnóstico principal de DPBOC é seguido pelo mesmo diagnóstico em cinco (05) internações com 0,112 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 514 casos.

Um diagnóstico principal de Broncopneumonia NE é seguido pelo diagnóstico de DBPOC com 0,103 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 472 casos.

Um diagnóstico principal de Parto pré-termo é seguido pelo diagnóstico de Parto Único Espontâneo NE com 0,101 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 463 casos.

Um diagnóstico principal de DBPOC é seguido pelo diagnóstico de Broncopneumonia NE com 0,099 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 454 casos.

Um diagnóstico principal de Outras Infeccções Intestinais Específicas é seguido pelo diagnóstico de Afecção Respiratória do Recém-Nascido com 0,090 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 413 casos.

Um diagnóstico principal de Asma NE é seguido pelo diagnóstico de Doença Pulmonar Obstrutiva Crônica (DBPOC) com 0,087 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 399 casos.

Um diagnóstico principal de Enfisema NE é seguido pelo diagnóstico de Doença Pulmonar Obstrutiva Crônica (DBPOC) com 0,086 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 394 casos.

Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo diagnóstico de Enfisema NE com 0,085 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 390 casos.

Um diagnóstico principal de Pielonefrite obstrutiva crônica é seguido pelo diagnóstico de Parto único espontâneo NE com 0,082 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 376 casos.

Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo diagnóstico de Asma NE com 0,082 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 376 casos.

Um diagnóstico principal de Enfisema NE é seguido pelo mesmo diagnóstico com 0,073 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 335 casos.

Um diagnóstico principal de Broncopneumonia NE é seguido pelo diagnóstico de Asma NE com 0,072 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 330 casos.

Um diagnóstico principal de Parto pré-termo é seguido pelo diagnóstico de Parto p/ cesariana NE com 0,069 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 316 casos.

Um diagnóstico principal de DBPOC é seguido pelo mesmo diagnóstico em seis (06) internações com 0,069% de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 316 casos.

Um diagnóstico principal de Colecistite Aguda é seguido pelo diagnóstico de Doença da Vesícula Biliar SOE com 0,069 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 316 casos.

Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo mesmo diagnóstico em quatro (04) internações com 0,068 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 312 casos.

Um diagnóstico principal de Pielonefrite Obstrutiva Crônica é seguido pelo mesmo diagnóstico com 0,065 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 298 casos.

Um diagnóstico principal de Asma NE é seguido pelo diagnóstico de Broncopneumonia NE com 0,063 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 289 casos.

Um diagnóstico principal de Outras Infecções Intestinais Específicas é seguido pelo diagnóstico de Broncopneumonia NE com 0,063 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 289 casos.

Um diagnóstico principal de Hemorragia gastrointestinal SOE é seguido pelo mesmo diagnóstico com 0,062 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 284 casos.

Um diagnóstico principal de DBPOC é seguido pelo mesmo diagnóstico e, após por Insuficiência Cardíaca NE com 0,060 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 275 casos.

Um diagnóstico principal de Outras Localizações Mal Definidas é seguido pelo mesmo diagnóstico com 0,059 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 270 casos.

Um diagnóstico principal de Afeccção Respiratória do recém-nascido é seguido pelo diagnóstico de Outras Infecções Intestinais Específicas com 0,059 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 270 casos.

Um diagnóstico principal de Afeccção Respiratória do recém-nascido é seguido pelo diagnóstico de Asma NE com 0,055 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 252 casos.

Um diagnóstico principal de Angina Instável é seguido pelo mesmo diagnóstico com 0,054 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 247 casos.

Um diagnóstico principal de DBPOC é seguido pelo mesmo diagnóstico e, após por Insuficiência Respiratória Aguda com 0,054 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 247 casos.

Um diagnóstico principal de Transtornos glomerulares doenças endócrinas é seguido pelo mesmo diagnóstico com 0,053 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 243 casos.

Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo diagnóstico de Insuficiência Respiratória Aguda com 0,051 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 234 casos.

Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo diagnóstico de Outros Acidentes Isquêmicos Cerebrais Transitórios com 0,050 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 229 casos.

Um diagnóstico principal de Asma NE é seguido pelo mesmo diagnóstico em três (03) internações com 0,049 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 225 casos.

Um diagnóstico principal de Cirrose biliar SOE é seguido pelo mesmo diagnóstico com 0,049 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 225 casos.

Um diagnóstico principal de DBPOC é seguido por dois diagnósticos de Insuficiência Cardíaca NE com 0,048 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 220 casos.

Um diagnóstico principal de DBPOC c/ Infecção Respiratória Aguda é seguido pelo mesmo diagnóstico com 0,048 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 220 casos.

Um diagnóstico principal de Outros Acidentes Isquêmicos Cerebrais Transitórios é seguido pelo diagnóstico de Insuficiênica Cardíaca NE com 0,048 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 220 casos.

Um diagnóstico principal de Hipertensão essencial é seguido pelo mesmo diagnóstico com 0,046 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 211 casos.

Um diagnóstico principal de Afeccção Respiratória do recém-nascido é seguido pelo mesmo diagnóstico em três (03) internações com 0,046 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 211 casos.

Um diagnóstico principal de DBPOC é seguido pelo diagnóstico de Insuficiência Cardíaca NE e, após por DPBOC com 0,045% de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 206 casos.

Um diagnóstico principal de DPBOC é seguido pelo diagnóstico de Insuficiência Respiratória Aguda e, após por DPBOC com 0,045 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 206 casos.

Um diagnóstico principal de Broncopneumonia NE é seguido pelo diagnóstico de Insuficiência cardíaca NE com 0,045 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 206 casos.

Um diagnóstico principal de Insuficiência cardíaca NE é seguido por duas (02) internações de DBPOC com 0,042 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 192 casos.

Um diagnóstico principal de Outros Acidentes Isquêmicos Cerebrais Transitórios é seguido pelo diagnóstico de DBPOC com 0,042% de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 192 casos.

Um diagnóstico principal de DBPOC é seguido pelo mesmo diagnóstico em seis (06) internações com 0,042 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 192 casos.

Um diagnóstico principal de Insuficiência Respiratória Aguda é seguido pelo diagnóstico de Insuficiência Cardíaca NE com 0,042 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 192 casos.

Um diagnóstico principal de Insuficiência Cardìaca NE é seguido pelo diagnóstico de Broncopneumonia NE com 0,040 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 183 casos.

Um diagnóstico principal de Asma NE é seguido pelo diagnóstico de Afecção Respiratória do recém-nascido com 0,040 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 183 casos.

Um diagnóstico principal de Pielonefrite obstrutiva crônica é seguido pelo diagnóstico de Parto p/ cesariana NE com 0,040 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 183 casos.

Um diagnóstico principal de Epilepsia NE é seguido pelo mesmo diagnóstico com 0,040 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 183 casos.

Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo mesmo diagnóstico e, após por DBPOC com 0,038 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 174 casos.

Um diagnóstico principal de Com complicações NE é seguido pelo diagnóstico de Insuficiência Cardíaca NE com 0,038 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 174 casos.

Um diagnóstico principal de DBPOC é seguido pelo diagnóstico de Outros Acidentes Isquêmicos Cerebrais Transitórios com 0,037 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 169 casos.

Um diagnóstico principal de Desnutrição Protêico-calória NE é seguido pelo mesmo diagnóstico com 0,037 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 169 casos.

Um diagnóstico principal de Insuficiência respiratória do recém-nascido é seguido pelo mesmo diagnóstico com 0,036 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 165 casos.

Um diagnóstico principal de Pneumonia NE é seguido pelo diagnóstico de DBPOC com 0,036 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 165 casos.

Um diagnóstico principal de Colecistite Aguda é seguido pelo mesmo diagnóstico com 0,036 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 165 casos.

Um diagnóstico principal de Neoplasias Malignas s/ especificação de localidade é seguido pelo mesmo diagnóstico com 0,036 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 165 casos.

Um diagnóstico principal de Pneumonia NE é seguido pelo mesmo diagnóstico com 0,034 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 156 casos.

Um diagnóstico principal de Broncopneumonia NE é seguido pelo mesmo diagnóstico em três (03) internações com 0,034 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 156 casos.

Um diagnóstico principal de Pré-eclâmpsia grave é seguido pelo diagnóstico de Parto p/ cesariana NE com 0,034 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 156 casos.

Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo mesmo diagnóstico em cinco (05) internações com 0,033 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 151 casos.

Um diagnóstico principal de Com complicações NE é seguido pelo mesmo diagnóstico em três (03) internações com 0,033 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 151 casos.

Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo diagnóstico de DBPOC e, após por Insuficiência Cardíaca NE com 0,032 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 146 casos.

Um diagnóstico principal de Insuficiência renal crônica NE é seguido pelo mesmo diagnóstico com 0,032 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 146 casos.

Um diagnóstico principal de Insuficiênica cardíaca congestiva é seguido pelo mesmo diagnóstico com 0,031 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 142 casos.

Um diagnóstico principal de Broncopneumonia NE é seguido pelo diagnóstico de Insuficiência respiratória Aguda com 0,031% de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 142 casos.

Um diagnóstico principal de Insuficiência respiratória do recém-nascido é seguido pelo diagnóstico de Afecção respiratória do recém-nascido com 0,031 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 142 casos.

Um diagnóstico principal de Insuficiência cardíaca NE é seguido pelo diagnóstico Com complicações NE com 0,031 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 142 casos.

Um diagnóstico principal de Dor lombar baixa é seguido pelo mesmo diagnóstico com 0,031 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 142 casos.

Um diagnóstico principal de Enfisema NE é seguido por dois (02) diagnósticos de DBPOC com 0,030 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 137 casos.

Um diagnóstico principal de DBPOC é seguido pelo mesmo diagnóstico e, após por Enfisema NE com 0,030 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 137 casos.

Um diagnóstico principal de DBPOC é seguido por dois (02) diagnósticos de Insuficiência Aguda com 0,030 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 137 casos.

Um diagnóstico principal de Asma NE é seguido por dois (02) diagnósticos de DBPOC com 0,030 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 137 casos.

Um diagnóstico principal de DBPOC é seguido pelo diagnóstico de Asma NE e, após de DBPOC com 0,030 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 137 casos.

Um diagnóstico principal de Insuficiência respiratória aguda é seguido pelo mesmo diagnóstico em três (03) internações com 0,030 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 137 casos.

Um diagnóstico principal de DBPOC é seguido pelo diagnóstico de Enfisema NE e, após por DBPOC com 0,030 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 137 casos.

Um diagnóstico principal de Pneumonia Bacteriana NE é seguido pelo mesmo diagnóstico com 0,030 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 137 casos.

Um diagnóstico principal de Pré-eclâmpsia grave é seguido pelo diagnóstico de Parto Único Espontâneo NE com 0,029 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 133 casos.

Um diagnóstico principal de Afeccção respiratória do recém-nascido é seguido pelo diagnóstico de Insuficiência respiratória do recém-nascido com 0,029 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 133 casos.

Um diagnóstico principal de Falso trabalho de parto NE é seguido pelo diagnóstico de Parto ùnico espontâneo NE com 0,029 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 133 casos.

Um diagnóstico principal de DBPOC é seguido pelo diagnóstico de Pneumonia Bacteriana NE com 0,029 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 133 casos.

Um diagnóstico principal de Hipertensão Essencial é seguido pelo diagnóstico de Insuficiência Cardíaca NE com 0,029 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 133 casos.

Um diagnóstico principal de DBPOC é seguido pelo mesmo diagnóstico e, após por Asma NE com 0,029 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 133 casos.

3 Regras Temporais de Doenças de Agravos Notificáveis

A doença de agravo notificável Hepatite por vírus é seguida pela mesma doença com suporte de 0.317 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 56 casos.

A doença de agravo notificável Varicela é seguida pela mesma doença com suporte de 0.125 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 22 casos.

A doença de agravo notificável Leptospirose é seguida pela doença Hepatite por Vírus com suporte de 0.119 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 21 casos.

A doença de agravo notificável Hepatite por vírus é seguida pela doença Leptospirose com suporte de 0.090 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 15 casos.

A doença de agravo notificável Meningite de causa não especificada é seguida pela mesma doença com suporte de 0.073 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 12 casos.

A doença de agravo notificável Leptospirose é seguida pela mesma doença com suporte de 0.068 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 12 casos.

A doença de agravo notificável Meningite de causa não especificada é seguida pela doença Hepatite por vírus com suporte de 0.040 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 7 casos.

A doença de agravo notificável Toxoplasmose é seguida pela mesma doença com suporte de 0.040 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 7 casos.

A doença de agravo notificável Dengue é seguida pela doença Malária com suporte de 0.034 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 6 casos.

A doença de agravo notificável Meningite de causa não especificada é seguida pela doença Outros exantemas por vírus com suporte de 0.028 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 4 casos.

A doença de agravo notificável Meningite de causa não especificada é seguida pela doença Varicela com suporte de 0.028 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 4 casos.

A doença de agravo notificável Dengue é seguida pela doença Hepatite por vírus com suporte de 0.023 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 4 casos.

A doença de agravo notificável Outros exantemas por vírus é seguida pela doença Varicela com suporte de 0.023 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 4 casos.

A doença de agravo notificável Meningite de causa não especificada é seguida pela doença Leptospirose com suporte de 0.023 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 4 casos.

A doença de agravo notificável Hepatite por vírus é seguida pela doença Varicela com suporte de 0.023 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 4 casos.

A doença de agravo notificável Malária é seguida pela doença Leptospirose com suporte de 0.017 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 3 casos.

A doença de agravo notificável de Paroditite Epidêmica é seguida pela doença Varicela com suporte de 0.017 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 3 casos.

A doença de agravo notificável Outros exantemas por vírus é seguida pela mesma doença com suporte de 0.017% entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 3 casos.

A doença de agravo notificável Dengue é seguida pela doença Leptospirose com suporte de 0.011 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 2 casos.

A doença de agravo notificável Malária é seguida pela doença Malária em com suporte de 0.011 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 2 casos.

A doença de agravo notificável Leptospirose é seguida pela doença Meningite de causa não especificada em com suporte de 0.011 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 2 casos.

A doença de agravo notificável Hepatite por vírus é seguida pela doença Meningite de causa não especificada em com suporte de 0.011 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 2 casos.

A doença de agravo notificável Hepatite por Vírus é seguida pela mesma doença com suporte de 0.006 % até três vêzes os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 1 caso.

A doença de agravo notificável Varicela é seguida pela doença Meningite de causa não especificada em com suporte de 0.006 % entre os 17.677 pacientes existentes no BD. Isto equivale a aproximadamente 1 caso.

Anexo 2 Validações de Padrões Temporais

1 Validação dos Padrões Seqüenciais Extraídos dos Procedimentos Realizados

TABELA 5.5 – Validação dos padrões seqüenciais extraídos dos procedimentos realizados de acordo com a SES.

	Comportamento dos procedimentos realizados pelos pacientes nas AIH de 2000	Validação do conhecimento		
		Conhecido	Desconhecido	Obs.[12]
1	Um procedimento realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo mesmo procedimento até nove (09) internações em 82 casos analisados.	X		X
2	Um procedimento realizado de Insuficiência Cardíaca é seguido pelo procedimento de Broncopneumonia em 192 casos investigados.	X		X
3	Um procedimento realizado de Insuficiência Cardíaca é seguido pelo procedimento de Acidente Vaso-Cerebral (AVC) Agudo em 280 casos.		X	
4	Um procedimento realizado de Acidente Vaso-Cerebral (AVC) Agudo é seguido pelo procedimento de Insuficiência Cardíaca em 275 casos.		X	
5	Um procedimento realizado de Insuficiência Cardíaca é seguido pelo procedimento Doença Pulmonar Obstrutiva Crônica (DBPOC) em 932 casos.	X		X
6	Um procedimento realizado de Crise Asmática é seguido pelo procedimento de Broncopneumonia em Lactente em 192 casos analisados.	X		
7	Um procedimento realizado de Acidente Vaso-Cerebral (AVC) Agudo é seguido pelo mesmo procedimento em 821 casos.	X		X
8	Um procedimento realizado de Diabetes Sacarino é seguido pelo mesmo procedimento até três (03) internações em 197 casos investigados.	X		X
9	Um procedimento realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo procedimento de Insuficiência Respiratória Aguda em 720 casos.	X		X

[12] Obs. são observações relativas a utilização de procedimentos realizados ou diagnósticos principais referentes a um mesmo grupo de patologias.

	Comportamento dos procedimentos realizados pelos pacientes nas AIH de 2000	Validação do conhecimento		
		Conhecido	Desconhecido	Obs.
10	Um procedimento realizado de Crise Asmática é seguido pelo mesmo procedimento em 661 casos.	X		
11	Um procedimento realizado de Intercorrências Clínicas de Paciente Oncológico é seguido pelo mesmo procedimento em 601 casos.	X		
12	Um procedimento realizado de Insuficiência Respiratória Aguda é seguido pelo procedimento de Doença Pulmonar Obstrutiva Crônica (DBPOC) em 569 casos analisados.	X		X
13	Um procedimento realizado de Insuficiência Respiratória Aguda é seguido pelo mesmo procedimento em 192 casos.	X		
14	Um procedimento realizado de Broncopneumonia é seguido pelo procedimento de Doença Pulmonar Obstrutiva Crônica (DBPOC) em 491 casos.	X		X
15	Um procedimento realizado de Broncopneumonia é seguido pelo mesmo procedimento em 491 casos.	X		
16	Um procedimento realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo procedimento de Broncopneumonia em 459 casos.	X		X
17	Um procedimento realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo procedimento de Crise Asmática em 404 casos.	X		X
18	Um procedimento realizado de Insuficiência Renal Crônica Acidose Metabólica é seguido pelo mesmo procedimento em 422 casos.	X		
19	Um procedimento realizado de Enfisema Pulmonar é seguido pelo procedimento de Doença Pulmonar Obstrutiva Crônica (DBPOC) em 417 casos.	X		X
20	Um procedimento realizado de Trabalho de Parto Prematuro é seguido pelo procedimento de Parto Normal em 413 casos.	X		

	Comportamento dos procedimentos realizados pelos pacientes nas AIH de 2000	Validação do conhecimento		
		Conhecido	Desconhecido	Obs.
21	Um procedimento realizado de Colecistite Aguda é seguido pelo procedimento de Colecistectomia em 404 casos.	X		
22	Um procedimento realizado de Colecistite Aguda é seguido pelo mesmo procedimento em 142 casos.	X		
23	Um procedimento realizado de Pielonefrite é seguido pelo procedimento de Parto Normal em 371 casos.	X		
24	Um procedimento realizado de Entero Infecções em Lactente é seguido pelo procedimento de Broncopneumonia em Lactente em 353 casos.	X		X
25	Um procedimento realizado de Enfisema Pulmonar é seguido pelo mesmo procedimento em 348 casos.	X		
26	Um procedimento realizado de Tratamento em Psiquiatria em Hospital Psiquiátrico é seguido pelo mesmo procedimento em 321 casos.	X		
27	Um procedimento realizado de Hemorragias Digestivas é seguido pelo mesmo procedimento em 312 casos.	X		
28	Um procedimento realizado de Insuficiência Coronariana Aguda é seguido pelo mesmo procedimento em 303 casos.	X		
29	Um procedimento realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo procedimento e depois por Insuficiência Cardíaca em 298 casos.	X		
30	Um procedimento realizado de Cirrose Hepática é seguido pelo mesmo procedimento em 289 casos.	X		
31	Um procedimento realizado de Entero Infecções em Lactente é seguido pelo mesmo procedimento em 284 casos.	X		X
32	Um procedimento realizado de Broncopneumonia em Lactente é seguido pelo mesmo procedimento em 1.189 casos.	X		

	Comportamento dos procedimentos realizados pelos pacientes nas AIH de 2000	Validação do conhecimento		
		Conhecido	Desconhecido	Obs.
33	Um procedimento realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo procedimento de Insuficiência Cardíaca em 1.125 casos.	X		
34	Um procedimento realizado de Trabalho de Parto Prematuro é seguido pelo procedimento de Cesariana em 275 casos.	X		
35	Um procedimento realizado de Crise Hipertensiva é seguido pelo mesmo procedimento em 270 casos.	X		
36	Um procedimento realizado de Broncopneumonia em Lactente é seguido pelo procedimento de Crise Asmática em 266 casos.	X		X
37	Um procedimento realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo mesmo procedimento, e depois do procedimento de Insuficiência Respiratória Aguda em 261 casos.	X		
38	Um procedimento realizado de Insuficiência Cardíaca é seguido pelo procedimento de Insuficiência Respiratória Aguda em 257 casos.	X		X
39	Um procedimento realizado de Acidente Vaso-Cerebral (AVC) Agudo é seguido pelo procedimento de Doença Pulmonar Obstrutiva Crônica (DBPOC) em 247 casos.		X	
40	Um procedimento realizado de Tratamento em Psiquiatria em Hospital Geral é seguido pelo mesmo procedimento em 243 casos.	X		
41	Um procedimento realizado de Diabetes Sacarino é seguido pelo procedimento de Insuficiência Cardíaca em 238 casos.	X		
42	Um procedimento realizado de Trabalho de Parto Prematuro é seguido pelo procedimento de Parto Normal com Atendimento do Recém Nascido na Sala de Parto em 238 casos.	X		
43	Um procedimento realizado de Pneumonia em Lactente é seguido pelo procedimento de Entero Infecções em Lactente em 234 casos.	X		

	Comportamento dos procedimentos realizados pelos pacientes nas AIH de 2000	Validação do conhecimento		
		Conhecido	Desconhecido	Obs.
44	Um procedimento realizado de Insuficiência Coronariana Aguda é seguido pelo procedimento de Insuficiência Cardíaca em 225 casos.	X		
45	Um procedimento realizado de Broncopneumonia é seguido pelo procedimento de Insuficiência Cardíaca em 215 casos.	X		
46	Um procedimento realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo procedimento de Insuficiência Respiratória Aguda e depois por um procedimento de Doença Pulmonar Obstrutiva Crônica (DBPOC) em 211 casos.	X		
47	Um procedimento realizado de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo procedimento de AVC Agudo em 206 casos.		X	
48	Um procedimento realizado de Insuficiência Respiratória Aguda é seguido pelo procedimento de Insuficiência Cardíaca em 206 casos.	X		
49	Um procedimento realizado de Broncopneumonia é seguido pelo procedimento de Crise Asmática em 206 casos.	X		X
50	Um procedimento realizado de Insuficiência Respiratória Aguda é seguido pelo procedimento de Insuficiência Cardíaca em 206 casos.	X		
51	Um procedimento realizado de Desnutrição (Clínica Médica) seguido pelo mesmo procedimento em 202 casos.	X		
52	Um procedimento realizado de Insuficiência Respiratória Aguda é seguido pelo procedimento Doença Pulmonar Obstrutiva Crônica (DBPOC) em 202 casos.	X		X
53	Um procedimento realizado de Pneumonia do Lactente é seguido pelo mesmo procedimento em três internações em 197 casos.	X		

2 Validação dos Padrões Seqüenciais Extraídos dos Diagnósticos Principais

TABELA 5.6 – Validação dos padrões seqüenciais extraídos dos diagnósticos principais de acordo com a SES.

	Comportamento dos diagnósticos principais dos pacientes nas AIH de 2000	Validação do conhecimento		
		Conhecido	Desconhecido	Obs.
1	Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo mesmo diagnóstico até nove (09) internações em 78 casos.	X		X
2	Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo mesmo diagnóstico até seis (06) internações em 73 casos.		X	
3	Um diagnóstico principal de Afeccção Respiratória do recém-nascido é seguido pelo mesmo diagnóstico em 1.235 casos.	X		X
4	Um diagnóstico principal de Broncopneumonia NE é seguido pelo mesmo diagnóstico em 1.216 casos.	X		X
5	Um diagnóstico principal de Asma NE é seguido pelo mesmo diagnóstico em 1.083 casos.	X		X
6	Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo diagnóstico de Insuficiência Cardíaca NE em 1.010 casos.	X		
7	Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo diagnóstico de Doença Pulmonar Obstrutiva Crônica (DBPOC) em 840 casos.	X		X
8	Um diagnóstico principal de Outras Infecções intestinais específicas é seguido pelo diagnóstico de Broncopneumonia NE em 289 casos.		X	
9	Um diagnóstico principal de Afecções respiratórias do recém-nascido é seguido pelo diagnóstico de Asma NE em 252 casos.	X		X
10	Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo diagnóstico de Angina Instável em 192 casos .	X		

	Comportamento dos diagnósticos principais dos pacientes nas AIH de 2000	Validação do conhecimento		
		Conhecido	Desconhecido	Obs.
11	Um diagnóstico principal de Outros acidentes isquêmicos cerebrais é seguido pelo diagnóstico de Doença Pulmonar Obstrutiva Crônica (DBPOC) em 110 casos.		X	
12	Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo diagnóstico de Insuficiência Respiratória Aguda em 684 casos.	X		
13	Um diagnóstico principal de Outros Acidentes Isquêmicos Cerebrais é seguido pelo mesmo diagnóstico em 661 casos.		X	
14	Um diagnóstico principal de Asma NE é seguido pelo diagnóstico de Doença Pulmonar Obstrutiva Crônica (DBPOC) em 399 casos.	X		X
15	Um diagnóstico principal de Enfisema NE é seguido pelo diagnóstico de Doença Pulmonar Obstrutiva Crônica (DBPOC) em 394 casos.	X		X
16	Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo diagnóstico de Enfisema NE em 390 casos.	X		X
17	Um diagnóstico principal de Pielonefrite obstrutiva crônica é seguido pelo diagnóstico de parto único espontâneo em 376 casos.	X		
18	Um diagnóstico principal de Doença Pulmonar Obstrutiva Crônica (DBPOC) é seguido pelo diagnóstico de Asma NE em 376 casos.	X		X
19	Um diagnóstico principal de Enfisema NE é seguido pelo mesmo diagnóstico em 335 casos.	X		
20	Um diagnóstico principal de Broncopneumonia NE é seguido pelo diagnóstico de Asma NE em 330 casos.	X		X
21	Um diagnóstico principal de Parto pré-termo é seguido pelo diagnóstico de Parto p/ cesariana NE em 316 casos.	X		

Anexo 3 Observações de Padrões Temporais

TABELA 5.7 - Nova validação do comportamento extraído dos diagnósticos principais dos pacientes nas AIH de 2000 conforme a SES.

	Comportamento nos diagnósticos principais apresentado pelos pacientes nas AIH de 2000	Comentários/Observações
1	Um diagnóstico principal de Afeccção Respiratória do recém-nascido é seguido pelo diagnóstico de Outras Infecções Intestinais Específicas com 0,059 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 270 casos.	Este conhecimento é desinteressante porque estatisticamente insignificante, mas com antibiótico e terapia, um dos para efeitos é diarréia.
2	Um diagnóstico principal de Afeccção Respiratória do recém-nascido é seguido pelo diagnóstico de Asma NE com 0,055 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 252 casos.	Este conhecimento é desinteressante porque a princípio é raro o diagnóstico de Asma NE para recém nascido.
3	Um diagnóstico principal de Angina Instável é seguido pelo mesmo diagnóstico com 0,054 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 247 casos.	Este conhecimento é desinteressante porque significa que paciente com angina instável está sendo bem tratado.
4	Um diagnóstico principal de Transtornos glomerulares doenças endócrinas é seguido pelo mesmo diagnóstico com 0,053 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 243 casos.	Este conhecimento é desinteressante porque transtornos glomeolares é uma patologia, não associada a doença endócrina.
5	Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo diagnóstico de Outros Acidentes Isquêmicos Cerebrais Transitórios com 0,050 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 229 casos.	Este conhecimento é desinteressante porque pode ser complicação de Insuficiência Cardíaca o AVC Isquêmico.
6	Um diagnóstico principal de DBPOC c/ Infecção Respiratória Aguda é seguido pelo mesmo diagnóstico com 0,048 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 220 casos.	Este conhecimento é desinteressante porque o paciente está bem controlado.
7	Um diagnóstico principal de Insuficiência Respiratória Aguda é seguido pelo diagnóstico de Insuficiência Cardíaca NE com 0,042 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 192 casos.	Este conhecimento é desinteressante porque pode acontecer, mas estatisticamente insignificante.
8	Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo diagnóstico de Broncopneumonia NE com 0,040 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 183 casos.	Este conhecimento é interessante porque representa mal acompanhamento à Insuficiência Cardíaca.

Comportamento nos diagnósticos principais apresentado pelos pacientes nas AIH de 2000		Comentários/Observações
9	Um diagnóstico principal de DBPOC é seguido pelo diagnóstico de Outros Acidentes Isquêmicos Cerebrais Transitórios com 0,037 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 169 casos.	Este conhecimento é desinteressante porque são patologias distintas e não estão relacionadas.
10	Um diagnóstico principal de Colecistite Aguda é seguido pelo mesmo diagnóstico com 0,036 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 165 casos.	Este conhecimento é interessante porque paciente deveria ter sido operado antes de novas crises de colecistite.
11	Um diagnóstico principal de Insuficiência Cardíaca NE é seguido pelo mesmo diagnóstico em cinco (05) internações com 0,033 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 151 casos.	Este conhecimento é interessante porque significa que o paciente está sendo mal acompanhado.
12	Um diagnóstico principal de Broncopneumonia NE é seguido pelo diagnóstico de Insuficiência Respiratória Aguda com 0,031% de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 142 casos.	Este conhecimento é interessante porque o quadro de Broncopneumonia pode evoluir para Insuficiência Respiratória Aguda.
13	Um diagnóstico principal de Dor lombar baixa é seguido pelo mesmo diagnóstico com 0,031 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 142 casos.	Este conhecimento é desinteressante porque a patologia é desinteressante. Sem potencial de gravidade.
14	Um diagnóstico principal de Asma NE é seguido pelo mesmo diagnóstico com 0,236 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 1.083 casos.	Este conhecimento é interessante porque existe um pequeno número de crises que demandaram internações.
15	Um diagnóstico principal de Septicemia NE é seguido pelo mesmo diagnóstico com 0,026 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 119 casos.	Este conhecimento é interessante porque não é habitual ter mais de uma Septicemia ao ano.
16	Um diagnóstico principal de Anemia nutricional NE é seguido pelo mesmo diagnóstico com 0,025 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 114 casos.	Este conhecimento é interessante porque refere-se a um mal acompanhamento nutricional do paciente. Rede ambulatorial ineficaz.

Comportamento nos diagnósticos principais apresentado pelos pacientes nas AIH de 2000		Comentários/Observações
17	Um diagnóstico principal de DBPOC é seguido pelo diagnóstico de Hemorragia gastrointestinal SOE com 0,025 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 114 casos.	Este conhecimento é interessante porque significa que paciente com DBPOC necessita fazer prevenção hemorrágica digestiva.
18	Um diagnóstico principal de Transt NE do aparelho urinário é seguido pelo mesmo diagnóstico com 0,025 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 114 casos.	Este conhecimento é desinteressante porque é um diagnóstico genérico.
19	Um diagnóstico principal de Bronquiolite aguda dev outr. microo. é seguido pelo diagnóstico de Afecção respiratória do recém-nascido com 0,025% de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 114 casos.	Este conhecimento é interessante pela fragilidade do paciente, tem alta letalidade.
20	Um diagnóstico principal de DBPOC é seguido pelo diagnóstico de Cardiopatia pulmonar NE com 0,025 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 114 casos.	Este conhecimento é desinteressante porque Cardiopatia pulmonar NE é um diagnóstico genérico.
21	Um diagnóstico principal de Parto único espontâneo é seguido pelo diagnóstico de Esterilização com 0,021 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 96 casos.	Este conhecimento é interessante desde que devidamente justificado.
22	Um diagnóstico principal de Parto único Espontâneo é seguido pelo diagnóstico de Infecção Puerperal com 0,019 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 87 casos.	Este conhecimento é interessante para controle de Infecção Hospitalares. Baixo índice de infecção pós-parto.
23	Um diagnóstico principal de Polineuropatia NE é seguido pelo mesmo diagnóstico com 0,017 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 78 casos.	Este conhecimento é interessante porque não se consegue fazer diagnóstico definitivo e nem tratamento definitivo.
24	Um diagnóstico principal de Septicemia NE é seguido pelo diagnóstico de Broncopneumonia NE com 0,015 % de suporte sobre os 459.189 pacientes. Isto equivale a aproximadamente 68 casos.	Este conhecimento é interessante porque esta não é a seqüência lógica.
25	Um diagnóstico principal de Doença do Pâncreas SOE é seguido pelo diagnóstico de Doença da Vesícula Biliar SOE com 0,014 % de suporte sobre os 459.189 pacientes. Isto equivale a 64 casos.	Este conhecimento é desinteressante porque SOE corresponde as iniciais de Sem Origem Específica, assim sendo, um conhecimento muito genérico.

Referências

[ADR 97] ADRIAANS, P.; ZANTINGE, D. **Data Mining**. Harlow: Addison-Wesley, 1997.

[AGR 93a] AGRAWAL, R. et al. Efficient Similarity Search in Sequences Databases. In: FOUNDATIONS OF DATA ORGANIZATION AND ALGORITHMS - FODO, 1993, Chicago, USA. **Proceedings...** Chicago: [s.n.], 1993.

[AGR 94a] AGRAWAL, R. et al. Fast Algoritms for Mining Association Rules. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, VLDB, 20., 1994, Santiago, Chile. **Proceedings...** Hove: Morgan Kaufmann, 1994. p. 490-501.

[AGR 94b] AGRAWAL, R.; SRIKANT, R. **Mining Sequential Patterns.** San Jose, CA: [s.n.], 1994. (IBM Research Report RJ9910).

[AGR 95b] AGRAWAL, R. ; SRIKANT, R. Mining Sequential Patterns. In: INTERNATIONAL CONFERENCE ON DATA ENGINEERING - ICDE, 1995, Taipei, Taiwan. **Proceedings...** Taiwan: [s.n.], 1995.

[AGR 96a] AGRAWAL, R.; SRIKANT, R. Mining Sequential Patterns: Generalizations and Performance Improvements. In: INTERNATIONAL CONFERENCE ON EXTENDING DATABASE TECHNOLOGY - EDBT, 5., 1996, Avignon, France. **Proceedings...** Berlin: Springer - Verlag, 1996.

[ALL 83] ALLEN, J. Maintaining Knowledge About Temporal Intervals. **Communications of the ACM**, New York, v. 26, n. 11, 1983.

[ALV 99] ALVARES, Luis Otavio Campos. **Ferramentas de Inteligência Artificial**. Porto Alegre: Instituto de Informática da UFRGS, 1999. Apostila da disciplina.

[ARN 96] ARNING, A.; AGRAWAL, R. ; RAGHAVAN, P. A Linear Method for Deviation Detection in Large Databases. In: Conference on Knowledge Discovery in Databases and Data Mining, 2., 1996, Portland, Oregon. **Proceedings...** Hove: Portland: [s.n.], 1996.

[BIG 2000] BIGOLIN, Nara M. Data Mining: Conceitos e Técnicas. In: ESCOLA DE INFORMÁTICA DA SBC SUL, 2000. **Anais...** [S.l.]: SBC, 2000.

[BLA 96] BLAKELEY, José A. Thoughts on Directions in Database Research. **ACM Computing Surveys**, New York, v. 28, 1996. Disponível em: <http://www.acm.org/pubs/citations/journals/surveyes/1996-28-4es/a77-blakeley>. Acesso em: 25 ago. 1997.

[BRE 84] BREIMAN, L. et al. **Classification and regression tree.** NY: Chapman and Hall, 1984.

[CHA 99] CHAN, K. et al. Efficient Times Series Matching by Wavelet. **ICDE** Sidney, Australia, p. 126-133, 1999.

[DAS 98] DAS, G. et al. Rule Discovery from Time Series. **KDD**, New York, USA, p. 16-22, 1998.

[EDE 94] EDELWEISS, N.; PALAZZO, J.M.O. **Modelagem de Aspectos Temporais de Sistemas de Informações**. Recife: UFPE, 1994. 163p. Trabalho apresentado na Escola de Computação, 9., 1994.

[EDE 98] EDELWEISS, N. **Banco de Dados Temporais:** Teoria e Prática. In: JORNADA DE ATUALIZAÇÃO EM INFORMÁTICA, 17., 1998, Belo Horizonte. **Anais...** Belo Horizonte: SBC, 1998.

[ENG 2001] ENGEL, Paulo Martins. **Sistemas de Informação Inteligentes**. Porto Alegre: Instituto de Informática da UFRGS, 2001. Apresentações da disciplina. Conceitos Básicos. Técnicas de classificação. Árvores de Decisão. Redes Neurais Artificais.

[FAY 93] FAYYAD, U.M. et al. Automated Analysis of a Large-Scale Sky Survery; The SKICAT System. In: WORKSHOP ON KNOWLEDGE DISCOVERY IN DATABASES - KDD, 1993. **Proceedings...** Washington D.C.: [s.n.], 1993.

[FAY 97a] FAYYAD, U.M. et al. **Advances in Knowledge Discovery and Data Mining**. Cambridge, MA: AAAI/MIT, 1997.

[FAY 97b] FAYYAD, U.M. et al. **Data Mining and Knowledge Discovery An International Journal**. Redmond, WA: Kluwer Academic Publishers, 1997.

[FEL 97] FELDENS, Miguel Artur. **Engenharia da Descoberta de Conhecimento em Bases de Dados.** 1997. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[FIS 87] FISHER, D. Knowledge Acquisition Via Incremental Conceptual Clustering. **Machine Learning**, [S.l.], v. 2, p. 139-172, 1987. Disponível em: <http://www.ics.uci.edu/AI/ML/Machine-Learning.html>. Acesso em: 22 maio 2001.

[FRA 91] FRAWLEY, W.J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. Knowledge Discovery In Databases: An Overview. In: PIATETSKY-SHAPIRO, G. ; FRAWLEY, W. (Ed.). **Knowledge Discovery In Databases**. Cambridge, MA.: AAAI Press/MIT Press, 1991. p. 1-30.

[FRE 96] FREITAS, A. ; LAVINGTON, S. Using SQL primitives and parallel DB servers to speed up knowledge discovery in large relational databases. In: EUROPEAN MEETING ON CYBERNETICS AND SYSTEMS RESEARCH, 13., 1996. **Cybernetics and Systems**. [S.l:s.n.], 1996. p.955-960.

[GAV 2000] GAVRILOV, M. et al. Mining the Stock Market: With Measure is Best? **KDD**, Boston, USA , 2000.

[GIL 2001] GILES, C. et al. Noise Time Series Prediction using Recurrent Neural Networks and Grammatical Inference. **Machine Learning,** [S. l.], v. 44, p. 161-184, 2001.

[GON 2000] GONZALEZ, R. **Processamento de Imagens Digitais**. São Paulo: Edgard Blücher, 2000.

[GUN 2000] GUNOPULOS, D. et al. **Time Series Similarity Measures**. Riverside: University of California, 2000.

[GUR 99] GURALNIK, V. et al. Event Detection from Series Data. **KDD**, San Diego, p. 33-34, 1999.

[HAN 2000] HAN, J. et al. Mining Frequent Itemsets Using Support Constraints. In: VLDB CONFERENCE, 26., 2000, Egypt. **Proceedings...** Cairo: [s.n.] , 2000.

[HAN 2001] HAN, J. et al. **Data Mining:** Concepts and Techniques. San Francisco, CA : Morgan Kaufmann , 2001.

[HUA 99] HUANG, Y. et al. Adaptative Querying Processing for Time Series Data. **KDD**, San Diego, p. 282-286, 1999.

[HUG 2000] HUGHES, J. et al. **Data Mining:** Looking Beyond the Tip of the Iceberg. Notherland Ireland: Informatics University of Ulster, 2000.

[IBM 99] **IBM Corpotation. Utilizando o Intelligent Miner for Data**. Versão 6. Release 1. Edição S517-6338-00. © Copyright International Business Machines Corporation 1996,1999.

[JOH 97] JOHN, G. H. **Enhancements to the Data Mining Process**. Stanford, EUA: Stanford University, 1997.

[KEO 97] KEOGH, E. et al. A Probabilistic Approach to Fast Pattern Matching in Time Series Databases. **KDD**, New Port Beach, p. 24-30, 1997.

[KOU 2001] KOUNDOURAKIS, G.; SARAEE, M.; THEODOULIDIS, B. **Data Mining in Temporal Databases. Information Management Group**. Manchester, United Kingdom: Department of Computation, UMIST, 2001.

[LAN 96] LANGLEY, P. **Elements of Machine Learning**. [S. l.]: Morgan Kaufman, 1996.

[LAN 98] LANG, K. et al. Results of the Abbandingo One DFA Learning Competition and a new Evidence-Driven State Merging Algorithm. **ICGI**, Ames, p. 79-89, 1998.

[LIN 98] LIN, L. et al. Querying Continuos Time Sequences. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, VLDB, 1998, New York. **Proceedings...** New York: [s.n.], 1998.

[MAN 95] MANNILA, H. et al. Discovery Frequent Episodes in Sequences. **KDD**, Montreal, p. 210-215, 1995.

[MAN 96] MANNILA, H. et al. Discovery Generalized Episodes Using Minimal Ocurrences. **KDD**, Portland, p. 146-151, 1996.

[MAN 97] MANNILA, H. et al. **Discovery Frequent Episodes in Event Sequences**. Finland: Department of Computer Science, University of Helsinki, 1997.

[MIL 92] MILLER, G. The Data Reduction Expert System. In: ASTRONOMY FROM LARGE DATABASES, 2., 1992, Haguenau, France. **Proceedings ...** France: [s. n.], 1992.

[OGI 99] OGIHARA, M. et al. Mining Features for Sequence Classification. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 5., 1999. **Proceedings...** [S.l.: s.n.], 1999.

[OLI 2001] OLIVEIRA, Arlindo. **Temporal Data Mining:** an overview. **IST/INESC-ID**. Lisbon: Lisbon Technical University, 2001.

[ÖZD 98] ÖZDEN,B. et al. Cyclic Association Rules. **ICDE,** Orlando, USA, p. 412-421, 1998.

[ÖZS 96] ÖZSU, M. Tamer. Future of Database Systems: Changing Applications and Technological Developments. **ACM Computing Surveys**. [S.l.], v. 28 , Dec. 1996. Disponível em: <http://www.acm.org/pubs/citations/journals/surveys/1996-28-4es/a85-ozsu>. Acesso em: 13 ago. 1997.

[PAZ 98] PAZZANI, M. et al. An enhanced representation of time series wich allows fast and accurate classification, clustering and relevance feedback. **KDD**, Madison, p. 239-241, 1998.

[PEI 2000] PEI, Jian ; HAN, Jiawei. Mining Frequent Patterns by Pattern-Growth: Methodology and Implications. **ACM SIGKDD Explorations**. Dec. 2000.

[PLA 2001] PLASTINO, Alexandre. Regras de Associação e Algoritmos de Mineração de Dados. In: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS – SBBD, 16., 2001, Rio de Janeiro. **Anais...** Rio de Janeiro: SBBD, 2001.

[POL 99] POLVINELLI, R. **Time Series Data Mining:** Identifying Temporal Patterns for Characterization and Prediction of Time Series Events. 1999. A Dissertation for the Degree of Doctor of Philosophy. Milwaukee, Wisconsin.

[PRA 98] PRADO, Hércules Antônio do. **Abordagens Híbridas para Mineração de Dados.** 1998. Exame de Qualificação (Doutorado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre.

[QUI 93] QUINLAN, J. **C4.5:** Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.

[RAH 2000] RAHM, E. et al. **Data Cleaning:** Problems and Current Approaches. Germany: University of Leipzig, 2000.

[RAI 99] RAINSFORD, Chris P.; RODDICK, John F. Adding Temporal Semantics to Association Rules. In: EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE AS KNOWLEDGE DISCOVERY IN DATABASES - PKDD, 3., 1999, Prague. **Proceeding...** [S.l.: s.n.], 1999.

[RAI 2001] RAINSFORD, Chris P.; RODDICK, John F. Database Issues in Knowledge Discovery and Data Mining. In: AUSTRALIAN JOURNAL OF INFORMATION SYSTEMS, 6., 2001, [S.l.]. **Proceedings...** [S.l.:s.n.], 2001.

[RAM 98] RAMASWANY, S. et al. Discovery if Interesting Patterns in Association Rules. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, VLDB, New York. **Proceedings...** New York: [s.n.] , 1998, p. 368-379.

[SAR 98] SARAWAGI, S.; AGRAWAL, R. ; MEGIDDO, N. Discovery-driven exploration of OLAP data cubes. In: INTERNATIONAL CONFERENCE ON EXTENDING DATABASE TECHNOLOGY - EDBT, 6., 1998, Valencia, Spain, **Proceedings...** Valencia: [s. n.], 1998. (IBM Research Report RJ 10102 (91918)).

[SAV 95] SAVASERE, A. et. al. An Efficient Algorithm for Mining Association Rules in Large Databases. Technical Report nro. GIT-CC-95-04. In: VLDB CONFERENCE, 21., 1995, Zurich, Switzerland, **Proceedings...** Zurich: [s. n.], 1995.

[SIL 96] SILBERSCHATZ, Avi; ZDONIK, Stan et al. Strategic Directions in Database Systems - Breaking Out of the Box. **ACM Computing Surveys**, [S. l.], v. 28, n. 4, p. 764-778, Dec. 1996.

[SIP 2001a] SIPILIOPOULOU, Myra. **Managing Interesting Rules in Sequence Mining**. Berlin: Institut für Wirtschaftsinformatik, Humboldt-Universität zu Berlin, 2001. Disponível em: <http://www.wiwi.hu-berlin.de/~myra>. Acesso em: 10 out. 2001.

[SMY 2000] SMYTH, P. Deformable Markov Model Templates For Time Series Pattern Matching, **KDD**, Boston, p. 81-90, 2000.

[STO 96] STONEBRAKER, Michael. Object-Relational DBMS - The Next Wave. **Informix**, [S. l.], 1996. Disponível em: <http://www.informix/corpinfo/zines/whitppr-s/iluswp/wave.htm>. Acesso em: 12 set. 1997.

[STO 99] STOLFO, S.; LEE,W. Data Mining Approaches for Intrusion Detection. In: DARPA, NY, 1999. **Proceeedings...** NY: Computer Science Department, Columbia University, 1999.

[THE 98] THEODOULIDIS, B et al. **Data Mining in Temporal Databases. Information Management Group**. Manchester, United Kingdom: Department of Computation, UMIST, 1998.

[WIT 2000] WITTEN, I. et al. **Data Mining**: Practical Machine Learning Tools and Techniques with Java implementations. San Francisco, CA: Morgan Kaufmann, 2000.

[WRI 2001] WRIGHT, Peggy. Knowledge Discovery In Databases: Tools and Techniques. **ACM Crossroads Student Magazine**, [S. l.], The ACM's First Electronic Publication, 2001. Disponível em: <http://www.acm.org/crossroads/xrds5-2/kdd.html#8>. Acesso em: 10 out. 2001.

[ZHO 99] ZHONG, N. ; ZHOU, L. Knowledge Discovery and Data Mining - Research and Practical Experiences. In: THE PACIFIC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 1999. **Proceedings...** [S.l.:s.n.], 1999.

[ZIG 85] ZIGHED, D.A.; RAKOTOMALALA, R. **The SIPINA method.** SIPINA for Windows. Educational version. Lyon: Laboratory ERIC, University of Lyon 2, 1995.

[ZYT 91] ZYTKOW, J.M. et al. Interactive Mining of Regularities Databases. In: IJCAI WORKSHOP ON KNOWLEDGE DISCOVERY IN DATABASES, 1991, Menlo Park, CA. **Proceedings...** Menlo Park, CA: AAAI Press/The MIT Press, 1991, p. 32-55.

Obras Consultadas

[AGR 93b] AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. DataBase mining: A performance perspective. **IEEE Transactions on Knowledge and Engineering**, [S. l.], p. 914-925, Dec. 1993.

[AGR 95a] AGRAWAL, R. et al. Fast Similarity Search in the Presence of Noise, Scaling and Translation in Time Series Databases. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, VLDB, Santiago, Chile. **Proceedings...** [S.l.: s.n.], 1995.

[AGR 96b] AGRAWAL, R. et al. Fast Discovery of Association Rules. In: ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING, 1996. **Proceedings...** [S.l.]: AAAI/MIT Press, 1996.

[ALU 94] ALUR, R. ; DILL, D. A theory of timed automata. Theoretical Computer Science, In: IEEE TRANSACTIONS ON KNOWLEDGE AND ENGINEERING, 1994, New York. **Proceedings...** NY: [s.n.], 1994.

[BEC 98] BECKENKAMP, F. G.; FELDENS, M.A.; PREE, W. Optimizations of the Combinatorial Neural Model. In: BRAZILIAN SIMPOSIUM ON NEURAL NETWORKS, 1998. **Proceedings...** [S.l.:s.n.], 1998.

[BER 96] BERNDT, D.et al. Finding **Patterns in Time Series:** a Dynamic Programming Approach. [S.l.]: AAAI Press, 1996.

[BET 96] BETTINI, C.; WANG, X.; JAJODIA, S.; LIN, J.-L. Discovering Temporal Relationships with Multiple Granularities in Time Sequences. In: IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 1996. **Proceedings...** [S.l.]: George Mason University,1996.Disponível em:<http://isse.gmu.edu/~csis/tdb/tdminin/tdmining_intro.html#AIS90>. Acesso em: 26 jun. 2001.

[BUE 98] BUECHNER, A.; ANAND, S. **Decision Support Using Data Mining.** NY: Financial Times Pitman Publishing, 1998.

[CAB 2000] CABENA. P. et al. **Discovering Data Mining From Concept to Implementation.** Upper Saddle River,New Jersey: IBM Prentice Hall PTR, 2000.

[CAL 96] CALDAS, Ruiter Braga. Aspectos Formais da Computação. In: IEC., 1996, Manaus, Amazonas. **Anais...** Manaus: Universidade do Amazonas Instituto de Ciências Exatas, 1996.

[CAS 99] CASIMIR, H. et al. **Knowledge Representation Forms for Data Mining Methodologies as Applied in Thoracic Surgery**. Slovenia: Faculty in Electrical Engineeing, University if Lijubljana, 1999.

[CHA 97] CHAUDHURI, S.; DAYAL, U. An overview of data warehousing and olap technology. **SIGMOD Record**, [S.l.], p. 65-74, 1997.

[DAS 97] DAS, G. et al. Finding Temporal Time Series. **PKDD**, Throundheim, Noruega , p. 88-100, 1997.

[DEC 91] DECHTER, R.; MEIRI, I.; PEARL, J. Temporal constraint networks. **Artificial Intelligence**, [S.l.], p. 61-95, 1991.

[ELD 96] ELDER, J.; KRIEGEL, H.; XU, X. A statiscal perspective on knowledge discovery in databases. In: FAYYAD, Usama M. et al. **Advantances in Knowledge Discovery and Data Mining.** NY: AAAI/MIT, 1996. p. 83-115.

[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.

[FAL 94] FALOUTSOS, C. et al. Fast Subsequence Matching in Time Series Databases. In: ACM SIGMOD, 1994, Mineapolis. **Proceedings...** Mineapolis: [s.n.], 1994.

[FEL 99] FELDENS, Miguel Artur; MORAES, R. L. M. Inteligência Artificial e Inteligência do Negócio: a evolução das tecnologias para suportar decisões bem informadas. In: OFICINA DE INTELIGÊNCIA ARTIFICIAL, 3., 1999, Pelotas. **Proceedings...** Pelotas: UCPel, 1999. Disponível em: <http://i.am/mfeldens>. Acesso em: 01 out. 2000.

[GAM 2001a] GAMA, João. Probabilistic Linear Tree Ltree. In: LIACC, 2001, Porto. **Proceedings...** Porto: FEP – Universidade do Porto, 2001. Disponível em: <http://www.up.pt/liacc/ML>. Acesso em: 11 nov. 2001.

[GOL 98] GOLENDZINER, Lia G.; SANTOS, Clésio S. Uma abordagem multi-nível para suporte aversões em bancos de dados orientados a objetos. **Revista de Informática Teórica e Aplicada**, Porto Alegre, v. 5, n. 1, p. 67-83, jul. 1998.

[GRA 97] GRAY, J. et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub totals. In: DATA MINING AND KNOWLEDGE DISCOVERY, 1997, Netherlands. **Proceedings...** Netherlands: Kluwer Academic Publishers, 1997.

[HAN 97] HAN, J. Olap mining: An integration of olap with data mining. In: IFIP CONFERENCE ON DATA SEMANTICS, 1997, Leysin. **Proceedings...** Leysin: Chapman & Hall, 1997.

[HOL 94a] HOLSHEIMER, M.; KERSTEN, M. **Architectural Support for Data Mining.** [S. l.: s.n.], 1994. (Report CS-R9429, CWI). Disponível em: <ftp://ftp.cwi.nl/pub/CWireports/AA/CS-R9429>. Acesso em: 30 nov. 2000.

[HOL 94b] HOLSHEIMER, M.; SIEBES, A. **Data mining:** the search for knowledge in databases. Amsterdam, The Netherlands: Centrum voor Wiskunde en Informatica, 1994. p. 78. Technical report.

[INM 96] INMON, W. H. The data warehouse and data mining. **Communications of the ACM**, Pine Cone Systems, Castle Rock. v. 39, n. 11, p. 49-50, Nov. 1996.

[JEN 94] JENSEN, C.S. et al. A Consensus Glossary of Temporal Database Concepts. **ACM SIGMOD Record**, [S.l.], v. 23, n.1, p. 52-64, Mar. 1994.

[JOR 97] JORGE, Alípio. Logical Data Mining. An account on logical approaches to data mining with emphasys on the fields of inductive logic programming and deductive databases. In: LIACC, 1997, Porto. **Proceedings...** Porto: Fac. Economia da Universidade do Porto, 1997. Disponível em: <http://www.ncc.up.pt/~amjorge>. Acesso em: 05 dez. 2001.

[KEO 99] KEOGH, E. et al. Scaling up Dynamic Tome Warpong to Massive Databases. In: PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES, 1999, [S.l.]. **Proceedings...** [S.l.:s.n.], 1999.

[KET 97] KETTERLIN, A. Clustering Sequences of Complex Objects. **KDD**, New Port Beach , 1997.

[KIE 2001] KIETZ, Joerg-Uwe. Data Preparation, Preprocessing and Reasoning for Real-World Data Mining Applications. In: KDD-SISYPHUS, 2001, Zurich. **Proceedings...** Zurich: [s.n.], 2001.

[KIM 96] KIMBALL, Ralph. **Data Warehouse Toolkit**. New York: John Wiley & Sons, 1996.

[LEN 77] LENAT,D.B. On automatic scientific theory formation: A case study using the AM program. In: MACHINE INTELLIGENCE, 9., 1997, NY. **Proceedings...** New York: Halsted Press, 1977.

[LIP 87] LIPPMANN, R. An introduction to computing with neural nets. **IEEE ASSP Magazine**, [S. l.], p 4-22, Apr. 1987. Overview Neural Nets.

[LUC 99] LUCAS, Anelise. **Um estudo das direções em Sistemas de Banco de Dados**. Trabalho de Modelos de Banco de Dados, disciplina ministrada pelos professores Clésio Saraiva dos Santos e Nina Edelweiss, Universidade Federal do Rio Grande do Sul, Porto Alegre, 1999.

[MAT 96] MATHEUS, C. J.; PIATETSKY-SPHAPIRO, G; MCNEILL, D. Selecting and Reporting What is Interesting: The KEFIR Application to Healthcare Data. In: FAYAD, Usama M. et al. **Advances in Knowledge Discovery and Data Mining.** Menlo Park, CA: AAAI/MIT Pres, 1996. p. 401-419.

[PAR 93] PARSAYE, K.; CHIGNELL, M. Data quality control with smart databases. **AI Expert**, [S.l.], v. 8, n. 5, p. 22-27, May 1993.

[PIA 9?] PIATETSKY-SHAPIRO, G.; BEDDOWS, M. **Knowledge Discovery Mine - Data Mining And Knowledge Discovery Resources**. [S.l.:s.n. 199?].

[PIA 2000] PIATETSKY-SHAPIRO, G. Knowledge Discovery in Databases: 10 years after. **SIGKDD Explorations**, NY, v. 1, n. 2, Feb. 2000.

[QUI 86] QUINLAN, J. Introduction of decision trees. **Machine Learning,** [S.l.], v.1, n. p. 81-106, 1986.

[RAI 2000] RAINSFORD, Chris P.; RODDICK, John F. **Adding Temporal Semantics to Association Rules.** Australia: [s.n.], 2000.

[ROD 2000] RODRIGUES FILHO, Ilson Wilmar. **Processamento de Linguagem Natural**. [S. l.]: UNOESC - Universidade para o Desenvolvimento do Oeste de SC, 2000. Disponível em: <http://www.inf.ufsc.br/~ilson/slides.ppt>. Acesso em: 05 fev. 2002.

[ROJ 2001] RODDICK, John F. et al. **Beyond Schema versoning:** A Flexible Model for Spatio-Temporal Schema Selection. The Netherlands: Kluwer Academic, 2001.

[ROJ 93] RODDICK, John F. et al. A taxonomy for schema versoning based on the relational and entity relationship models. In: INTERNATIONAL CONFERENCE ON ENTITY-RELATIONSHIP APPROACH, 1993, Dallas, Texas. **Proceedings...** Texas: Springer-Verlag, 1993. p.143-154.

[ROJ 99] RODDICK, John F. **A Model for Temporal Inductive Inferece and Schema Evolution in Relational Database Systems.** Bundoora, Austrália: School of Computer Science and Computer Engineering Faculty of Science and Technology, La Trobe University, 1999.

[SIP 2001b] SIPILIOPOULOU, Myra; RODDICK, John. **Higher order mining:** modelling and mining the results of knowledge discovery. Adelaide, Australia: University of South Australia, 2001.

[WAN 94] WANG, J. Tsong-Li et al. Combinatorial pattern discovery for scientific data: Some preliminary results. In: SIGMOD CONFERENCE**.** 1994, [S. l.]. **Proceedings...** [S.l.:s.n.], 1994.

[WAN 96] WANG, K. TAN, J. Incremental discovery of sequential patterns. In: WORKSHOP ON RESEARCH ISSUES ON DATA MINING AND KNOWLEDGE DISCOVERY, 1996, Montreal, Canada. **Proceedings...** Montreal: [s.n.], 1996.

[XIA 97] XIA, Betty. **Similarity Search in Time Series Data Sets.** Tese (Doutorado) , Simon Fraser University, [S.l.].