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

GLAUCO CARLOS SILVA

Mineração de Regras de Associação Aplicada a Dados da Secretaria Municipal de Saúde de Londrina - PR

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

Prof. Dr. Luis Otavio Campos Álvares
Orientador

Porto Alegre, junho de 2004

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Silva, Glauco Carlos

Mineração de Regras de Associação Aplicada a Dados da Secretaria Municipal de Saúde de Londrina -PR/ Glauco Carlos Silva – Porto Alegre: Programa de Pós-Graduação em Ciência Computação, 2005.

76f.:il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2003. Orientador: Luis Otavio Campos Álvares.

1.Informática.2.Mineração de dados. I. Álvares, Luis Otavio Campos. II. Título.

AGRADECIMENTOS

A minha família e amigos.

Ao Prof. Dr. Luis Otavio Campos Álvares por ter acreditado na proposta, e pelo indispensável auxílio fornecido durante o desenvolvimento do trabalho.

Aos colegas da Autarquia do Serviço Municipal de Saúde de Londrina - PR.

A FACCAR e a UFRGS.

SUMÁRIO

LISTA DE ABREVIATURAS

RDBMS	*Relational Database Management System*;
OLTP	*On Line Transaction Processing*;
SMP	Multi-Processamento Simétrico;
MPP	Processamento Maciçamente Paralelo;
DSS	*Decision Support Systems*;
EIS	*Executive Information System*;
DCBD	Descoberta de Conhecimento em Bases de Dados;
KDD	*Knowledge Discovery in Data Base;*
SQL	*Structured Query Language*;
ER	Entidade Relacionamento;
SUS	Sistema Único de Saúde;
PSF	Programa Saúde da Família;
CBOR	Classificação Brasileira de Ocupações – Reduzida;
CADSUS	Sistema de Cadastramento dos Usuários do SUS;
CLEITOS	Sistema de Controle de Central de Leitos;
SIAB	Sistema de Informações da Atenção Básica.

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Com o grande crescimento dos volumes de dados que as organizações vêm registrando e a diversidade das fontes destes dados, o fato de se aproveitar informações contidas nessas massas de dados se tornou uma necessidade. Surgiu então uma área denominada Descoberta de Conhecimento em Bases de Dados (DCBD).

Tal área utiliza alguns modelos, técnicas e algoritmos que realizam operações de extração de conhecimento útil de grandes volumes de dados. Entre as principais técnicas utilizadas para minerar os dados está a de Regras de Associação.

A técnica de Regras de Associação se propõe a encontrar todas as associações relevantes entre um conjunto de itens aplicados a outros itens, e utiliza alguns algoritmos para realizar seu objetivo.

Este estudo apresenta alguns algoritmos para a aplicação da técnica de Regras de Associação, também, busca abranger um pouco da tecnologia de *Data Warehouse*, muito útil para que o processo de mineração de dados possa ser realizado com maior sucesso.

Neste trabalho são aplicadas técnicas de descoberta de conhecimento na área de saúde, vinculando dados referentes à situação socioeconômica do paciente com os procedimentos que foram realizados nas internações hospitalares a que foi submetido. Devido ao grande número de regras que poderiam se geradas resultantes das inúmeras possibilidades da base de dados, foi construído um protótipo de uma ferramenta para extração de regras de associação, que não só é baseado no suporte e confiança, mas também utiliza os conceitos de *lift* e *improvement* os quais ajudam na diminuição de regras triviais.

Foram realizadas minerações com a base de dados de pacientes da Secretaria Municipal de Saúde de Londrina-PR, para análise da utilidade dos dados minerados.

Palavras-chave: Regras de Associação, Mineração de Dados.

Mining of Association Rules Applied to Londrina´s Health City Department – PR

ABSTRACT

The increasing amount of data that organization have been registering and the diversity of data sources have generate the necessity of extract knowledge from this mass of data. Based on this necessity a new area has emerged which is named Knowledge Discovery in Data Base (KDD).

In this work apply the association rule mining technique in the public health area, linking social economic situation of patients which were attended in our hospitals. Because of large number of rules that can be produce we developed a prototype of a tool for extract association rules, not only based on support and confidence, but using too the measures lift and improvement in order to reduce the number of rules.

Experiments were performed with the "Secretaria Municipal de Saúde de Londrina – PR" database.

Keywords: Associations Rules, Data Mining.

1 INTRODUÇÃO

As últimas décadas acompanharam um aumento dramático na quantidade de informações e/ou dados armazenados em forma eletrônica. Este aumento se deve à facilidade de armazenamento dos dados com uma grande capacidade computacional com baixo custo. Estes dados possuem, além do beneficio direto para os sistemas para os quais foram definidos, a capacidade de se transformar em informações úteis para suporte a decisões.

Para que esta transformação se torne realidade, as pesquisas para a descoberta de conhecimento têm crescido. Uma das razões para investir esforços é afirmada por [PAR89], ao dizer que a informação é abundante e a capacidade de armazenamento é maior do que a capacidade de recuperação, ficando, portanto, clara a necessidade de prover recursos mais poderosos para a recuperação de informações. Esta necessidade se encaixa no conceito de *Data Mining*.

Nos últimos anos o uso de técnicas de *Data Mining* tem se tornado freqüente por vários motivos. O volume de dados disponível hoje é muito grande e o conceito de *Data Mining* se aplica às grandes massas de dados onde seus algoritmos se calibram e tiram conclusões confiáveis. Os dados estão mais bem organizados e padronizados devido à tecnologia de *Data warehouse*, os recursos computacionais vêm crescendo e seus custos baixando[CAR 01].

1.1 Áreas de Aplicação

Técnicas de *Data Mining* têm sido aplicadas com sucesso para a solução de problemas em diversas áreas, como exemplificado a seguir:

Vendas – Identificar padrões de comportamento dos consumidores; encontrar características dos mesmos de acordo com a região demográfica; prever quais consumidores serão atingidos nas campanhas de marketing.

Finanças – Detectar padrões de fraudes no uso dos cartões de crédito; identificar os consumidores que estão tendendo a mudar a companhia do cartão de crédito; descobrir regras de estocagem a partir dos dados do mercado.

Transporte – Determinar a distribuição dos horários entre os vários caminhos e analisar padrões de sobrecarga.

Medicina – Caracterizar o comportamento dos pacientes para prever novas consultas; buscar terapias de sucessos para diferentes doenças.

1.2 Conceitos gerais de *Data Mining*

O termo *Data Mining* é na verdade uma etapa em um processo maior que pode ser chamado de *Knowledge Discovery in Databases* (KDD) ou Descoberta do Conhecimento em Banco de Dados (DCBD).

Data Mining, portanto, é uma etapa do processo de descoberta de novas correlações, padrões e tendências entre as informações contidas em uma grande base de dados. Este processo é dado pela análise dos dados, utilizando técnicas de reconhecimento de padrões e estatísticas.

O termo KDD ou DCBD foi criado em 1989, sustentado por [PAR89], para se referir ao amplo processo de descoberta do conhecimento em bases de dados. O conceito geral do KDD passa por uma seqüência de fases, como mostra a Figura 1.1:

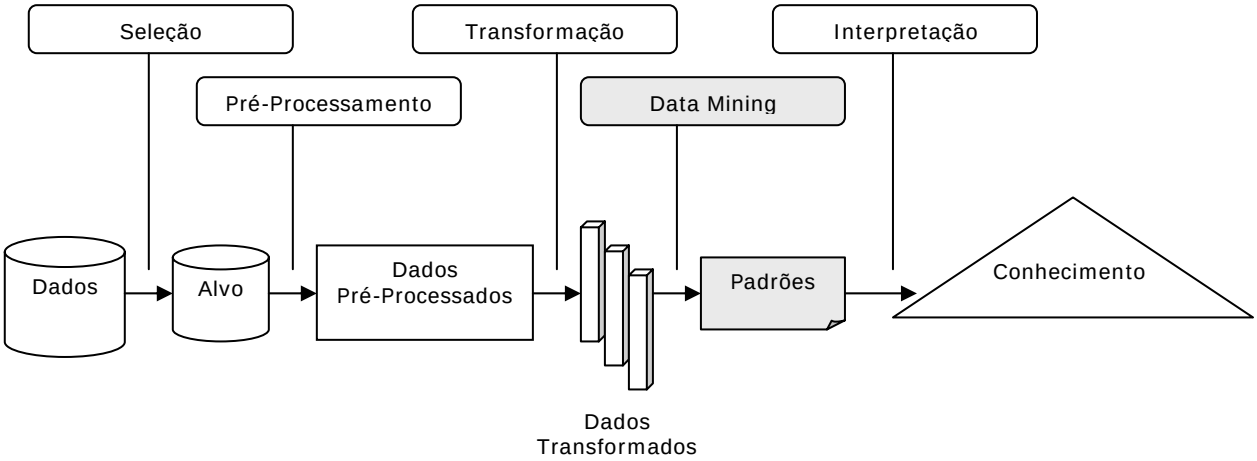


Figura 1.1: Conceito geral do KDD e sua seqüência de fases.

Seleção – Nesta fase se faz a seleção ou segmentação dos dados de origem, seguindo-se critérios básicos.

Pré-processamento – É o momento em que os problemas com dados incompletos, compatibilidade de tipos de dados ou algum tipo de ruído serão tratados.

Transformação – Nesta fase poderão ser aplicadas operações de projeção e redução na qual, o número de variáveis sob consideração, pode ser reduzido.

Data Mining – Nesta fase a extração de padrões de dados é efetuada, esta operação é realizada através e algoritmos, muitas vezes de forma repetitiva.

Interpretação e avaliação – Os padrões identificados pelo sistema são interpretados e avaliados em forma de gráficos e/ou relatórios e assim convertidos em conhecimento.

Devido à importância da fase de *Data Mining*, muitas vezes este termo é tido como sinônimo de todo o processo, mas a qualidade dos resultados depende diretamente da realização correta de todas as fases [FAY96], [BRA96].

1.3 Motivação

O crescimento da população nos grandes centros urbanos faz com que a qualidade dos serviços públicos oferecidos, como, segurança, saúde, transportes, tenha de crescer em igual proporção. Na busca desta qualidade de atendimento é necessária uma análise mais profunda da população que se quer atingir, agilizando a tomada de decisão.

Atualmente as organizações que prestam serviços de utilidade pública, devem apresentar suas contas a algum órgão superior. Tal apresentação de contas normalmente é feita por meio informatizado e tem em conseqüência uma grande quantidade de dados armazenados, gerando assim uma potencial fonte de informações que poderão ser exploradas a fim de melhorar o atendimento da população. Tais organizações podem explorar suas bases de dados com ferramentas de mineração de dados, elevando a qualidade no atendimento.

No caso da área específica de Saúde Pública os dados armazenados possuem o potencial de revelar uma infinidade de informações que podem ser úteis, tanto no atendimento reativo, quanto no atendimento pró-ativo da população.

No atendimento reativo, as Unidades Básicas de Saúde, Pronto Atendimentos e Hospitais, tendo um melhor conhecimento da população a que se deve atender, a informação gera agilidade e qualidade.

No caso dos atendimentos pró-ativos, sendo eles, programa médico de família, PSF (Programa Saúde da Família) e agentes comunitários de saúde, a agilidade é fundamental, filtrando as pessoas que realmente têm um potencial para o acompanhamento das equipes.

1.4 Objetivos

Este trabalho tem a proposta de estudar os fundamentos da tecnologia de *Data Mining* com foco na técnica de Regras de Associação, desenvolver um protótipo com base no algoritmo *Apriori,* utilizando os conceitos de *Lift* e *Improvement* que auxiliam na diminuição de regras triviais. O protótipo construído deverá analisar uma base real de dados de saúde que una características sócio-econômicas com dados de procedimentos realizados em internações hospitalares, as regras geradas deverão ser validadas e discutidas com um especialista.

Este trabalho tem como objetivos específicos:

1) Analisar alguns algoritmos existentes para descoberta de regras de associação em bancos de dados, especialmente algoritmos seqüenciais [AGR93] [AGR94] [BAY99];

2) Obter uma base de dados que contenham informações de pacientes com suas características sócio econômicas e também, dos procedimentos realizados em internações hospitalares;

3) Implementar o protótipo de uma ferramenta de *Data Mining* capaz de minerar a base acima descrita;

4) Realizar a mineração da base com o protótipo;

Validar as regras geradas por um especialista.

1.5 Metodologia

Foi realizada uma revisão bibliográfica sobre o problema da descoberta de regras de associação e selecionados alguns dos algoritmos mais utilizados para descoberta destas regras em bancos de dados.

Para uso do algoritmo foi utilizada uma base de dados, com a união das informações do sistema do Cartão Nacional de Saúde, com as do sistema de Central de Leitos.

Foi implementado um protótipo de uma ferramenta de mineração baseada no algoritmo *Apriori*, utilizando Borland Delphi e o banco de dados Microsoft Access.

Foram realizadas minerações sobre a base de dados visando a validação das regras geradas por um especialista.

O trabalho deu-se nas seguintes fases:

Revisão bibliográfica

Estudo das técnicas referentes ao domínio do problema

Análise dos Algoritmos

Analisar alguns algoritmos existentes para descoberta de regras de associação em bancos de dados, especialmente algoritmos seqüenciais [AGR93] [AGR94] [BAY99].

Obtenção da Base de Dados

Obter nos sistema disponíveis na Autarquia do Serviço Municipal de Saúde a base de dados para uso no projeto.

Implementação do Protótipo

Implementação do protótipo de uma ferramenta de *Data Mining* capaz de minerar a base obtida.

Mineração e Validação das Regras Obtidas

Validação das regras através de experimentos.

1.6 Organização do Trabalho

Este trabalho encontra-se dividido em seis capítulos, sendo que, no capítulo 2 são explanados conceitos gerais sobre a tecnologia de *Data Mining*, no capítulo 3 é descrita com mais detalhes a técnica de Regra de Associação e os algoritmos para implementação da técnica de Regra de Associação são descritos no capítulo 4. O capítulo 5 descreve os passos para a obtenção da base de dados e, também, a construção do protótipo e sua arquitetura. O capítulo 6 descreve os passos para se realizar minerações aplicadas à base de dados real obtida, bem como seus resultados.

2 TECNOLOGIA DE DATA MINING

As metas básicas da tecnologia de *Data Mining* são baseadas nos modelos de prognóstico e descrição. O prognóstico faz uso de variáveis existentes na base de dados para predizer o desconhecido ou valores futuros. A descrição se baseia na procura de padrões contidos em uma base de dados e os apresenta ao usuário para interpretação.

2.1 Modelos de Data Mining[MEN 98], [FAY 96]

2.1.1 Modelo de Prognóstico

O modelo de prognóstico resume-se na avaliação do valor futuro de algum índice, baseando-se em dados do comportamento passado deste índice, ou seja, prognóstico envolve o uso de atributos do banco de dados para predizer o valor futuro e desconhecido de outra variável. O prognóstico pode incluir tarefas como determinar se o índice de uma bolsa de valores subirá ou descerá em uma determinada situação, quanto o valor de uma dada ação da bolsa variará no próximo pregão, qual será a população de uma certa cidade daqui a dez anos, entre outras.

2.1.2 Modelo de Descrição

O modelo de descrição difere, em sua ênfase, do sistema que descobre informação importante automaticamente escondido nos dados. Os dados são peneirados à procura de padrões de ocorrências freqüentes, tendências e generalizações sobre os dados sem intervenção ou direção do usuário. As ferramentas de mineração visam revelar um número grande de fatos sobre os dados no menor tempo possível.

Um exemplo de tal modelo, é um banco de dados financeiro que é explorado para descobrir os muitos grupos de clientes para atingir em uma campanha dirigida de marketing. Os dados são procurados sem hipóteses pré-definidas, diferente do sistema que agrupa os clientes de acordo com as características comuns encontradas.

Cada um dos modelos, seja ele o prognóstico ou a descrição, utiliza uma gama de algoritmos, que por sua vez são incorporados por várias técnicas de *Data Mining*.

2.2 Técnicas de *Data Mining*

Existem várias técnicas para realizar a operação de *Data Mining*, cada qual usada para resolver problemas específicos ou encontrar um objetivo. Em geral as técnicas seguem os dois modelos descritos anteriormente.

2.2.1 Regras de Associação

Esta técnica se propõe a encontrar todas as associações relevantes entre um conjunto de itens aplicados a outros itens. Esta técnica se enquadra no modelo de descrição.

Como as Regras de Associação é um dos objetos de estudo deste trabalho veremos mais sobre esta técnica e alguns dos algoritmos mais utilizados para sua implementação nos capítulo 3 e 4.

2.2.2 Classificação

Esta técnica se enquadra no modelo de prognóstico e se propõe a gerar o perfil de diferentes grupos previamente definidos, em que, após ser determinado um modelo de classificação, o algoritmo é capaz de predizer a classe na qual novos casos serão enquadrados.

Entre os algoritmos mais utilizados na implementação da técnica de classificação destacam –se:

- **Árvore de Decisão**: utiliza a idéia de segmentar recursivamente o conjunto até encontrar uma partição que represente casos pertencentes à mesma classe.

- **Redes Neurais:** fornecem um método prático para funções de aprendizado, sendo que sua principal característica é a robustez com que lida com os erros no conjunto de treinamento da própria rede. Em geral as redes neurais apresentam um tempo maior de conclusão que as árvores de decisão, apesar deste tempo variar dependendo do número de casos e da definição de parâmetros para o algoritmo.

2.2.3 Padrões seqüenciais

Esta técnica se enquadra no modelo de descrição e se propõe a identificar padrões, ou tendências de seqüência, aplicados a uma regra mínima definida pelo usuário. A entrada é tipicamente uma seqüência de transações temporais.

A descoberta de padrões seqüenciais tem sido motivada por aplicações na indústria e no varejo.

2.2.4 *Clustering* ou Agrupamento

Esta técnica se enquadra no modelo de descrição e se propõe a segmentar a base de dados em subconjuntos, ou seja, agrupar objetos físicos ou abstratos em classes de objetos similares de forma a identificar agrupamentos que descrevam os dados. Estas categorias podem ser mutuamente exclusivas e exaustivas ou consistir de representações mais aprimoradas como hierarquias ou categorias sobrepostas.

A análise dos agrupamentos ajuda a construir partições significativas de grandes conjuntos de objetos. Um bom exemplo de aplicação desta técnica está na descoberta de subpopulações de clientes em um banco de dados comercial.

2.3 Data Warehouse e Data Mining

O potencial do *Data Mining* pode ser aumentado se os dados apropriados forem colecionados e armazenados em um *Data Warehouse*. Um Data Warehouse é um sistema de administração de banco de dados relacional (RDBMS), especificamente projetado para satisfazer as necessidades de sistemas de processamento de transações. Pode ser definido como repositório de dados centralizado, que pode ser examinado para benefício do negócio.

Data Warehousing é uma técnica poderosa que torna possível extrair dados operacionais arquivados e superar inconsistências entre formatos de dados de legados diferentes. Como também integram dados ao longo de um empreendimento, indiferentemente da localização, formato, ou exigências de comunicação, será possível incorporar informação adicional ou especialista. É o vinculo lógico entre o que os gerentes vêem no apoio de suas decisões de aplicações de EIS e as atividades operacionais da companhia, em outras palavras, o *Data Warehouse* provê dados que já são transformados e resumidos e por esta razão fazem isto em um ambiente apropriado para maior eficiência de aplicações de DSS e de EIS [INM 02].

O Repositório de Dados e a Mineração de Dados:

Existe uma relação de cumplicidade entre as atividades de *Data mining* e *Data warehouse* – a fundação da arquitetura dos sistemas de suporte à decisão. Os *Data warehouse* organizam os dados para um efetivo processo de *Data mining*. O *Data mining* pode ser feito onde não exista nenhum *Data warehouse*, mas este, aumenta as chances do sucesso do *Data mining*.

Os *Data warehouse* organizam os dados de acordo com sua natureza, as quais incluem:

- Dados integrados;
- Dados detalhados e resumidos;
- Dados históricos;
- Metadados.

Cada um desses elementos melhora o processo de *Data mining* e os prospectos de sucesso:

- **Dados Integrados** permitem ao minerador visualizar de forma rápida e fácil os dados. Sem dados integrados, o minerador gastaria uma grande quantia de tempo limpando e condicionando os dados antes do processo de *Data mining*, só então poderia iniciar o processo de mineração de uma maneira efetiva. Chaves teriam de ser reconstituídas, valores codificados reconciliados, estruturas de dados padronizadas, entre outros, para que o minerador não tivesse que trabalhar no processo de *Data mining* com dados crus. Os *Data warehouse* são integrados e têm todas essas tarefas (e muitas outras) feitas, portanto o minerador pode se concentrar na *Data mining* ao invés de limpar e integrar os dados;

- **Dados detalhados** e resumidos são ambos incluídos nos repositórios de dados. Dados detalhados são necessários quando o minerador deseja examiná-los na sua forma mais granular. Níveis muito baixos de detalhes importantes de modelos que não podem ser discernidos, a não ser pela cuidadosa análise dos detalhes dos dados. Justamente por isso, dados resumidos asseguram que se uma análise prévia já foi feita, o minerador não tem que repetir o trabalho que alguém já fez antes dele começar o processo de exploração. Dados resumidos asseguram que o minerador pode utilizar-se do trabalho de outros, em vez de fazer todo o processo desde o início.
- **Dados Históricos** são importantes para a operação de *data mining*, porque grandes quantidades de dados estão implicitamente guardadas lá. Um minerador que tem de trabalhar somente com informações atuais pode nunca detectar tendências e padrões de comportamento ao longo do tempo. Informações históricas são cruciais para entender o condicionamento dos negócios;
- **Meta-dados** servem como um mapa para o minerador, o qual os usa para descrever, não o conteúdo, mas o contexto da informação. Quando a informação está sendo examinada, com o passar do tempo, o contexto torna-se mais importante que o conteúdo. Em outras palavras, torna-se muito difícil para o minerador trabalhar com conteúdo de dados crus, quando não existe explanação a respeito do significado dos mesmos.

Conseqüentemente, os repositórios de dados organizam o estágio para o sucesso e a eficácia na exploração do mundo dos dados, e o minerador que trabalha nessas fundações de repositórios de dados aproveita o sucesso que vem com a exploração destes como recurso.

3 REGRAS DE ASSOCIAÇÃO

Devido aos avanços tecnológicos que possibilitaram a coleta de dados referentes às vendas onde são registradas as transações e os itens comprados, o interesse em extrair informações úteis aumentou. Com base nessa necessidade surgiu a técnica de regras de associação introduzida por [AGR 93], cujo objetivo seria encontrar relacionamentos ou padrões freqüentes entre conjuntos de dados

O algoritmo de associação tem diversas aplicações: supermercados, planejamento de inventário, organização de gôndolas, planos de vendas, entre outros. Um exemplo de aplicação de regras de associação seria em uma base de dados de vendas, associando regras entre os grupos de itens vendidos nas transações comerciais.

Quando determinados padrões de comportamento, como associação de produtos durante um processo de compras, por exemplo, começam a se repetir com freqüência, as ferramentas *Data Mining* indicam a presença de oportunidades e "*insights*" em relação àquele público consumidor.

Um exemplo clássico de *Data Mining* foi desenvolvido pela Wal-Mart. A empresa descobriu que o perfil do consumidor de cervejas era semelhante ao de fraldas. Eram homens casados, entre 25 e 30 anos, que compravam fraldas e/ou cervejas às sextas-feiras à tarde no caminho do trabalho para a casa. Com base na verificação destas hipóteses, a Wal-Mart optou por uma otimização das atividades junto às gôndolas nos pontos de vendas, colocando as fraldas ao lado das cervejas.

Um modelo matemático foi formalizado em [AGR 93] para referenciar problemas de extração de regras de associação

Onde:

- I é um conjunto de itens de venda descritos por $I = \{ i_1, i_2 ..., i_m \}$;
- T representa uma transação de venda tal que $T \subseteq I$;
- TID representa o chave que identifica a transação;
- D representa um conjunto de transações de vendas;
- X e Y conjuntos de itens de venda contidos em uma transação $X \subseteq T$ e $Y \subseteq T$;
- Uma regra de associação é uma implicação de $X \Rightarrow Y$ onde $X \subset I$, $Y \subset I$, e $X \cap Y = \phi$;

- s é o Suporte de uma determinada regra X ⇒ Y é dado através do total de transações contido no subconjunto de transações que contém X ∪ Y sobre o conjunto de transações D. O suporte é descrito pela seguinte fórmula:

Suporte = Nº de registros da tabela que contêm o conjunto / Nº total de registros da tabela

- c é a Confiança de uma determinada regra X ⇒ Y é dada através do total de registros do subconjunto X ∪ Y sobre o total de registros do subconjunto X. A confiança é descrita pela seguinte fórmula:

Confiança = Número de registros da tabela que contêm todos os itens da regra / Nº de registros da tabela que contêm o antecedente da regra.

O Suporte e a Confiança das regras são de suma importância para o processo, sendo que somente as regras com um alto grau de suporte e confiança serão usadas.

As regras com alta confiança e forte suporte são referenciadas como regras fortes [AGR 93], [SHA 91]. A tarefa de extração de regras de associação é essencial para descobrir regras de associação fortes em bases de dados grandes.

3.1 Representação dos Dados

É na fase de pré-processamento que os dados a serem minerados são fornecidos. Estes dados são basicamente compostos pelo identificador da transação e seus itens. A forma em que os dados fornecidos serão apresentados deve ser considerada e para tanto existem duas alternativas:

3.1.1 Formato Horizontal [ZAC 97]

Tabela 3.1: Formato Horizontal.

TID	A	B	C
1	1		
2		1	
3			1
4	1	1	1

Fonte: ZAC 97.

Em geral este é o formato utilizado onde os dados são apresentados em uma lista de transações identificadas pela chave TID.

Este formato apenas permite o cálculo do suporte com a passagem completa por toda a base de dados. Portanto, para facilitar o processo, o cálculo do suporte é feito para todos os itens apenas em uma passagem.

3.1.2 Formato Vertical [ZAC 97a]

Tabela 3.2: Formato Vertical.

TID	A	B	C
1	1		
2		1	
3			1
4	1	1	1

Fonte: ZAC 97a.

Menos comum que o formato Horizontal, o formato Vertical, também chamado de armazenamento decomposto, caracteriza-se por representar os dados em lista de itens, onde cada item possui suas transações.

Neste caso, como cada lista de itens é independente, não existe a necessidade da passagem por todo o banco de dados para se calcular o suporte, bastando verificar o número de transações que compõe cada lista de itens.

A desvantagem deste formato está no fato de que a lista de transações não fornece nenhuma informação de associação entre os outros itens, e como conseqüência, para obter tal associação, o consumo de processamento será muito alto [HER 95].

3.2 Regras de Associação Quantitativas

O Processo de *Data Mining* por Regras de Associação em tabelas de domínio não binário é chamado de *Data Mining* por regras de Associação Quantitativas [SRI 96].

O domínio não binário é em geral aplicado em situações reais nas quais os atributos podem ser quantitativos, como no caso de "idade", "nº de dependentes", ou categorizados, como no caso de "sexo" e "tipo sanguíneo".

Tabela 3.3: Regras de Associação Quantitativas (Origem).

TID	Idade	Nº de dependentes	Sexo
1	18	0	M
2	45	3	M
3	20	1	F
4	25	1	F
5	65	2	F

Fonte: SRI, 96.

A Tabela 3.3 exibe um banco de dados que contém dados de atributos quantitativos como Idade e nº de dependentes, e também dados de atributos categorizados, representados por "Sexo".

Para se aplicar o processo de mineração de dados binários em um domínio quantitativo, ou categorizado, bastaria mapear este domínio quantitativo e/ou categorizado para o domínio binário, onde cada atributo e seus valores gerariam novas colunas. O atributo "Sexo", por exemplo, geraria outras duas colunas, (Sexo = M) e

(Sexo = F). Esta solução se empregaria aos atributos categorizados e aos atributos quantitativos cujo domínio de valores não fosse muito vasto.

Nos casos em que o domínio de valores for muito vasto, a solução seria a divisão em intervalos, formando faixas de valores. O atributo Idade, neste caso, poderia gerar as colunas (Idade = 0..19), (Idade = 20..59) e (Idade = 60..+).

Tabela 3.4: Regras de Associação Quantitativas.

TID	Idade 0..19	Idade 20..59	Idade 60..+	Dependentes 0..2	Dependentes 3..+	Sexo M	Sexo F
1	1	0	0	1	0	1	0
2	0	1	0	0	1	1	0
3	0	1	0	1	0	0	1
4	0	1	0	1	0	0	1
5	0	0	1	1	0	0	1

Fonte: Resultado mapeamento Tabela 3.3.

A Tabela 3.4 exibe o resultado do mapeamento da Tabela 3.3, para o domínio binário, em que os valores dos atributos quantitativos foram divididos em faixas de valores, e o atributo categorizado representado por seus valores possíveis.

Os algoritmos para gerar Regras de Associação podem ser divididos entre algoritmos seqüenciais e paralelos. A regra geral para os algoritmos seqüenciais é que na sua maioria assume que os conjuntos de itens estão em ordem lexicográfica (baseada no nome do item). Esta ordem proporciona uma maneira lógica dos grupos de itens serem contados e gerados. Por outro lado, os algoritmos Paralelos se preocupam em como paralelizar a tarefa de encontrar os grandes conjuntos de itens.

3.3 *Lift*

Em [BAY 99] um novo limite foi proposto para trabalhar junto com o suporte mínimo (minsup) e a confiança mínima (minconf), este limite é chamado de *Lift* e se dá pela seguinte formula:

Lift = Confiança da regra / Suporte do conseqüente da regra.

3.4 Improvement

Em [BAY 99] um novo limite foi proposto para diminuir o número de regras geradas, utilizando o princípio de que uma regra mais simples é uma regra melhor, desde que a regra mais complexa ou mais longa tenha confiança menor ou igual do que a regra mais simples ou menor.

O *Improvement* é a diferença mínima entre as confianças das regras e suas sub-regras, em conceitos gerais *Improvement* se dá pela seguinte fórmula:

$$Imp(A \rightarrow C) = \min(\forall A' \subset A, conf(A \rightarrow C) - conf(A' \rightarrow C))$$

O conceito de *Improvement* esta melhor detalhado na seção 4.6.

4 ALGORITMOS DE REGRAS DE ASSOCIAÇÃO

4.1 Algoritmo AIS

O algoritmo AIS foi o primeiro desenvolvido para gerar todos os grandes grupos de itens em uma base de dados de transações [AGR 93]. Este algoritmo foca na extensão das funcionalidades da base de dados, para processar consultas de suporte a decisão Seu foco está em descobrir regras quantitativas e é limitado a apenas um item subseqüente. Esta técnica é formalizada da seguinte maneira.

$X \Rightarrow I_j \mid \alpha$, onde X é o conjunto de itens, I_j é um único item dentro do domínio e α é a confiança da regra.

O AIS executa vários passos dentro da base de dados inteira, durante cada passo ele verifica todas as transações e no primeiro passo verifica o suporte para cada item, determinando o quanto estes são freqüentes na base de dados. Para cada passo, o grande conjunto de itens é estendido para gerar os conjunto de itens candidatos. Depois de mapear as transações são determinados os conjuntos de itens comuns entre o grande conjunto de passos anterior e os itens da transação atual, estes conjuntos de itens comuns são estendidos com outros itens da transação para gerar o novo conjunto de itens candidatos. O grande conjunto de itens *l* é estendido com apenas os itens dentro de uma transação grande e que estão em uma ordem lexicográfica de itens ao final de qualquer item dentro do conjunto de itens *l* .

Para executar esta tarefa de forma eficiente são usadas ferramentas de estimativa e técnicas de poda. Tais recursos determinam os conjuntos candidatos, omitindo os conjuntos de itens desnecessários aos mesmos, assim o suporte de cada candidato é determinado. Os candidatos que possuem suporte maior ou igual ao suporte mínimo dado anteriormente, são selecionados dentro dos grandes conjuntos de itens. Para o próximo passo os grandes conjuntos de itens são estendidos para gerar os grande conjuntos candidatos. O processo termina quando não há mais grandes conjuntos de itens a serem determinados.

O algoritmo acredita que se um conjunto de itens não está em toda base de dados ele nunca se tornará um candidato para medir um grande conjunto de itens nos passos seguintes. Para evitar a repetição de um conjunto de itens estes são dispostos em ordem lexicográfica. O conjunto de itens A tenta uma extensão para um dos itens em B (i.e., $B=I_1, I_2, \ldots\ldots I_k$) que estão por último na ordem do que qualquer membro de A (i.e = extensão de um conjunto de itens com um item).

Como exemplo temos I={p, q, r, s, t, u, v}, e {p, q} sendo o grande conjunto de itens, para a transação T = {p, q, r, s} o próximo conjunto de itens candidatos são gerados:

- {p, q, r} supostamente grande: continua a extensão
- {p, q, s} supostamente grande: não pode ser estendido
- {p, q, r, s} supostamente pequeno: não pode ser estendido

Podemos ver como o suporte esperado de A+B é calculado. O suporte esperado de A+B é o produto de freqüências individuais relativas dos itens em B e que tem suporte para A, como vemos abaixo [AGR 93]:

- Suporte esperado = $f(I_1) \times f(I_2) \times . . . \times f(I_k) \times (x\text{-}c)$ / tamanho_da_base_de_dados.

Onde $f(I_i)$ representa a freqüência relativa do item I_i na base de dados, e (x-c) / tamanho_da_base_de_dados, representa o suporte atual para A na partição restante da base de dados.

Tal que x = o conjunto de transações que contém o conjunto de itens A, c = número de transações dentro de A que foram processadas no passo corrente e tamanho_da_base_de_dados = número total de transações na base de dados.

A geração de um número enorme de candidatos pode gerar o *overflow* da memória, então um gerenciador de memória adequado deve ser usado. O algoritmo sugere que os grandes conjuntos de itens fiquem residentes no disco e não na memória.

4.2 Algoritmo SETM

O algoritmo SETM foi proposto por [HOU 95] motivado pelo uso do SQL para calcular os grandes conjuntos de itens [SRI 96b]. Neste algoritmo cada membro do conjunto de grandes conjuntos de itens $\overline{L_k}$, é formalizado por <TID, conjunto_de_itens> onde TID é o identificador único da transação. Igualmente cada membro do conjunto de conjuntos candidatos $\overline{C_k}$, é formalizado por <TID, conjunto_de_itens>.

Similar ao AIS, o SETM realiza vários passos sobre a base de dados. No primeiro passo calcula o suporte para cada item determinando o quanto são freqüentes ou grandes dentro da base de dados. Então ele gera os conjuntos de itens candidatos estendendo o conjunto de itens por vários passos. Além disso o SETM recupera os TID da transação com os conjuntos de itens candidatos. A operação de merge_join pode ser usada para gerar os conjuntos de itens candidatos [SRI 96b]. Gerando os conjuntos candidatos, o algoritmo grava a cópia do conjunto de itens candidato junto com o TID da transação, de maneira seqüencial. Posteriormente os conjuntos de itens candidatos são organizados sobre o conjunto de itens e os pequenos conjuntos de itens são apagados através de uma função de agregação. Se a base de dados estiver ordenada pelo TID, no próximo passo o grande conjunto de itens contidos em uma transação é obtido pela organização $\overline{L_k}$ em TID. Este caminho requer vários passos na base de dados. Quando não forem encontrados mais grandes conjuntos de itens o algoritmo termina.

A grande desvantagem deste algoritmo é o número de conjuntos candidatos que ele gera $\overline{C_k}$ [AGR 94].

4.3 Algoritmo *Apriori*

O algoritmo *Apriori* desenvolvido por [AGR 94] é um grande evento na história do processo de *Data Mining* por Regras de Associação [CHE 96], sendo também o mais conhecido. Sua técnica se baseia no uso de vários subconjuntos oriundos de um grande conjunto de itens. A diferença fundamental entre este algoritmo e o AIS SETM está no caminho para geração do conjunto de itens candidatos e da seleção do conjunto candidato para contagem.

Em [AGR 93], [AGR 94], [PAK 95], os problemas de extração de regras de associação são decomposta em duas fases.

4.3.1 Primeira Fase

A primeira fase consiste em descobrir um grande conjunto de itens, que obedeçam a um suporte mínimo determinado em s.

Tabela 4.5: Base de dados de Origem.

TID	Itens
1	A B
2	A B
3	A B C
4	A B C
5	A B C
6	A B C
7	A C
8	A B C
9	B C
10	A B C

Fonte: AGR 93, AGR 94, PAK 95.

Tendo como base de dados a tabela indicada pela Tabela 4.5, na qual cada linha representa uma transação identificada por TID e seus itens, e que o suporte mínimo determinado será 0,6, os conjuntos gerados com este suporte mínimo são dados na Tabela 3.2, definidos por:

Suporte = Nº de registros da tabela que contêm o conjunto / Nº total de registros da tabela

Tabela 4.6: Suporte.

Conjuntos	Suporte
{A}	0,9
{B}	0,9
{C}	0,8
{A B}	0,8
{A C}	0,7
{B C}	0,7
{A B C}	0,6

Fonte: Dados primeira fase.

Assim que a primeira fase é concluída são obtidos os conjuntos que atendem ao suporte mínimo e se passa a segunda fase.

4.3.2 Segunda fase

A segunda fase consiste em utilizar o conjunto com suporte mínimo para gerar as regras de associação para a base de dados.

Tendo como base o indicado pela Tabela 4.6 onde cada linha representa um conjunto com suporte mínimo, serão geradas regras indicadas na Tabela 4.7 com confiança mínima 0,8. A confiança nestas regras é dada por:

Confiança = Número de registros da tabela que contêm todos os itens da regra / Nº de registros da tabela que contêm o antecedente da regra.

Tabela 4.7.: Confiança.

Regras	**Confiança**
{A} => B	0,88
{B} => A	0,88
{A} => C	0,77
{C} => A	0,87
{B} => C	0,77
{C} => B	0,87
{A B} => C	0,75
{A C} => B	0,85
{B C} => A	0,85

Fonte: Experimentos Tabela 4.6.

Supondo o fator de confiança maior ou igual a 0,8, apenas as regras sem sombreamento serão válidas.

O maior volume de processamento exigido no processo total é constatado na primeira fase. Sendo que na segunda, as regras podem ser derivadas de gerenciamento direto.

Com base neste algoritmo, várias extensões foram propostas em adaptação para banco de dados relacionais, com o objetivo de tornar as aplicações independentes de formato ou de Sistemas Gerenciadores de Banco de Dados (SGBD), selecionando um conjunto de dados aplicados ao algoritmo.

O algoritmo *Apriori* identifica os itens mais freqüentes da base de dados através de várias passagens, sendo que no primeiro passo gera um conjunto de itens candidatos, então percorre a base de dados novamente verificando se cada item destes possui o suporte mínimo estabelecido anteriormente. O conjunto de itens gerado é representado por conjunto_1 o qual terá apenas um item, que multiplicado por ele mesmo, resultará em um novo conjunto_2 com dois itens, no qual será verificado o suporte. Este processo recursivo será realizado até que o conjunto resultante, conjunto_*k* com *k* itens, seja um conjunto vazio.

Em um primeiro passo os conjuntos com apenas um item são contados, a descoberta de grandes conjuntos de itens no primeiro passo é usada para gerar os conjuntos candidatos para o segundo passo usando a função apriori_gen() descrita com mais detalhes a seguir. Uma vez que o conjunto candidato é encontrado, seu suporte é calculado para geração do grande conjunto de itens de tamanho 2, No terceiro passo o grande conjunto de itens do segundo passo é considerado para gerar o conjunto de itens candidato deste passo, este processo interativo termina quando não é encontrado um grande conjunto de itens.

Cada passo *i* do algoritmo lê a base de dados uma vez e determina o grande conjunto de dados de tamanho *i* . L_i refere-se ao grande conjunto de itens de tamanho *i* enquanto C_i refere-se ao conjunto de itens candidato de tamanho *i*.

Apriori [AGR 94]

Entrada:

I, D, s

Saida:

L

Algoritmo:

1) C_1: = I; // Conjunto de itens candidato 1

2) Gerar L_1 lendo a base de dados e contando cada ocorrência de um atributo na transação;

3) **for** (k = 2; $L_{k-1} \neq \phi$; k++) **do begin** //Geração do conjunto de itens candidatos, Novos conjuntos de itens candidatos são gerados de (k-1) grande conjunto de itens

4) C_k = apriori-gen(L_{k-1});

5) Count (C_k, D) //Cálculo do suporte para C_k

6) $L_k = \{c \in C_k \mid c.count \geq minsup\}$

7) **end**

9) L:= $\bigcup_k L_k$

Função: count(C: a set of itemsets, D: database)

begin

for each transaction T $\in$ D=$\bigcup D^i$ **do begin**

forall subsets x $\subseteq$ T **do**

if x $\in$ C then

x.count++;

end

end

A função apriori_gen() descrita em [AGR 94] possui dois passos. Durante o primeiro passo L_{k-1} é multiplicado por ele mesmo obtendo C_k . No segundo a função apaga todos os conjuntos de itens da operação que tem (k-1)–subconjunto ou seja que não estão em L_{k-1}.

Função: apriori_gen() [AGR 94]

Entrada: conjunto de todos os grandes (k-1)-conjunto de itens L_{k-1}

Saída: Um super conjunto do conjunto de todos os grandes conjuntos k-itens

//Passo de Junção

I_i = Items i

insert into C_k

Select $p.I_1$, $p.I_2$, ……. , $p.I_{k-1}$, $q.I_{k-1}$

From L_{k-1} is p, L_{k-1} is q

Where $p.I_1 = q.I_1$ and …… and $p.I_{k-2} = q.I_{k-2}$ and $p.I_{k-1} < q.I_{k-1}$.

//Passo de deleção

forall itemsets $c \in C_k$ do

forall (k-1)-subsets s of c do

If ($s \notin L_{k-1}$) then

delete c from C_k

A função subset() retorna o subconjunto dos conjuntos candidatos que aparecem na transação. O cálculo do suporte é um passo que consome muito tempo [CEN 97], para reduzir o número de candidatos ele precisa revisar cada transação, e os candidatos C_k são armazenados na hash tree, onde cada nodo da árvore contém uma lista de conjuntos de itens ou uma tabela hash, sendo identificados como um nodo folha ou nodo interno, respectivamente. A raiz da árvore está definida no nível 1. Um nodo interno de nível *n* aponta para nodos *n*+1 e os nodos folhas armazenam os conjuntos de itens.

Este algoritmo adota como princípio que cada subconjunto de itens freqüentes também deve ser freqüente. Esta regra é utilizada para reduzir o número de candidatos a serem comparados com cada registro do banco de dados, Todos os candidatos que contenham subconjuntos que não sejam freqüentes são eliminados.

Tabela 4.8: Conjuntos de itens candidatos gerados.

TID	Itens
1	A C D
2	B C E
3	A B C E
4	B E

Fonte: Experimentos realizados.

A partir de uma base de dados de itens dada na Tabela 4.8 pode-se verificar na Figura 4.2 os conjuntos de itens candidatos gerados C_k, e os itens, cujo suporte foi igual ou maior que o suporte tido como mínimo igual 2 em L_k, para cada passagem $_k$ na base de dados.

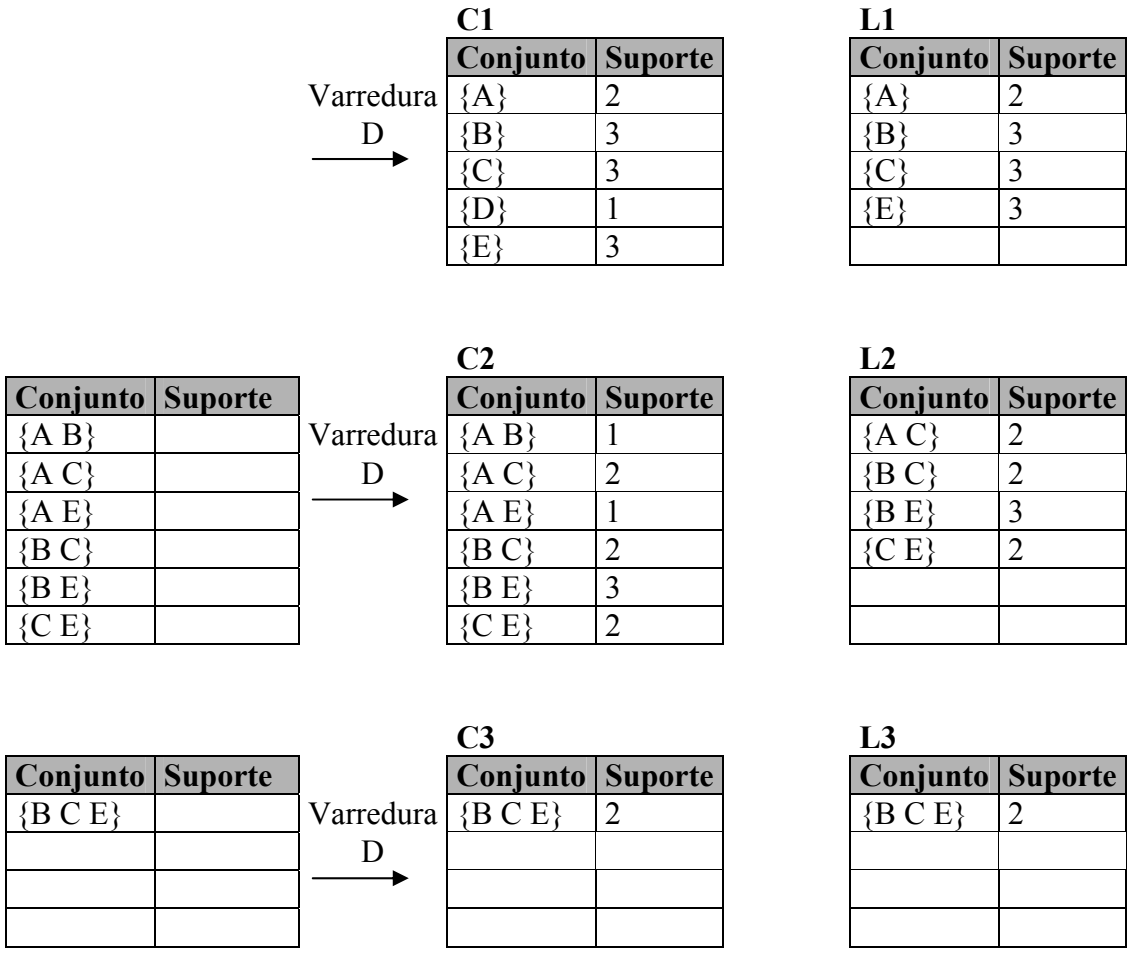


Figura 4.2: Processos a partir de uma base de dados de itens.

Muitas tarefas de mineração de dados, tais como: regras de associação e padrões seqüenciais, usam estruturas de dados complexas baseadas em ponteiros. Dentre estas estruturas a *hash tree* é a mais utilizada [PAR98]. Uma *hash tree* é uma estrutura híbrida com características de estruturas de *tree* e *hash table*.

Um nodo interno de uma *hash tree* em uma profundidade d contém uma *hash table* na qual células apontam para nodos em uma profundidade d+1. Todos os conjuntos de itens (itemsets) são armazenados nas folhas e ordenados em uma lista. A profundidade

máxima 40 de uma árvore em uma iteração k é k. Os diferentes componentes de uma *hash tree* são os seguintes: (HTN) *hash tree node*, (HTNP) *hash table*, (ILN) *Itemset list header*, (LN) *list node* e os *itemsets* (ISET). Um nodo interno em uma hash table aponta para nodos no próximo nível, e uma lista de conjuntos vazios, enquanto um nodo tem uma lista de conjuntos de itens. Para contar o suporte dos conjuntos de itens candidatos em uma passagem k, para cada transação T no banco de dados, são formados todos os conjuntos de itens com k elementos de T em uma ordem lexicográfica. Para cada subconjunto a hash tree é vasculhada em busca do conjunto de itens candidatos e seu contador é atualizado.

4.4 *Apriori-TID* [AGR 94]

O algoritmo *Apriori* desenvolvido por [AGR94] lê a base de dados na sua totalidade para calcular o suporte, sendo que esta leitura pode ser necessária apenas em alguns passos. Baseado neste raciocínio [AGR94] propôs um outro algoritmo chamado Apriori-TID.

Como no algoritmo *Apriori* o *Apriori-TID* usa a mesma função de geração de candidatos para determinar o conjunto de itens candidatos, sendo que a grande diferença entre o algoritmo *Apriori* e o *Apriori-TID* é que este último não usa a base de dados para calcular o suporte depois do passo inicial. Para isso é utilizado um conjunto C_k.

Um conjunto C_k é em um conjunto dos itens candidatos na passagem k associados com o identificador de transação (TID). Cada membro do conjunto C_k está na forma < TID, { X_k } >, onde TID é o identificador da transação e cada X_k é um conjunto de dados candidato presente na transação com o identificador TID. O membro de C_k correspondente à transação t é <t.TID, {c ∈ C_k | c está contido em t}>.

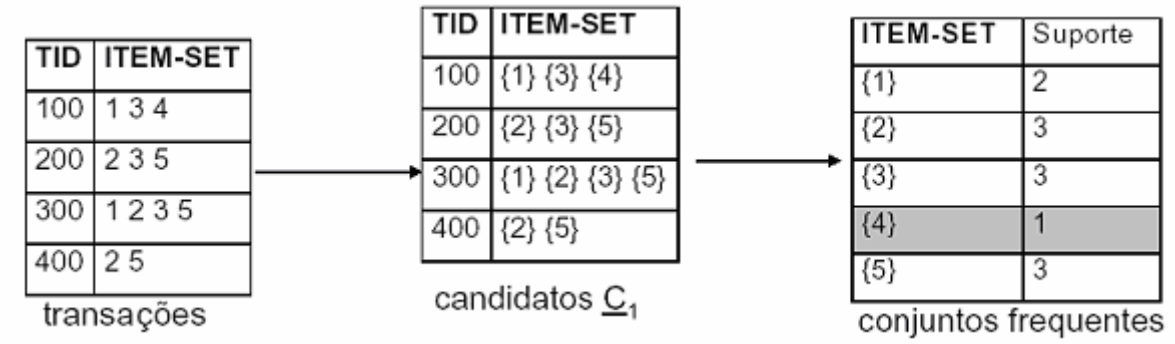


Figura 4.3: Conjunto de itens candidatos.

Como visto na Figura 4.3 a função Apriori-gen gera um conjunto de itens candidatos c_k com k itens pela união de conjunto de dados freqüentes com (k-1) itens. São mantidos dois campos adicionais para cada conjunto de itens candidatos: geradores e extensões. O primeiro campo e conjunto de itens candidato C_k armazena os IDs dos dois conjuntos de itens freqüentes com (k-1) itens que originaram o conjunto de itens com k item gerado. O campo extensões e um conjunto de itens C_k armazenam os IDs de todos os conjuntos de itens candidatos com (k+1) itens que são extensão do conjunto de itens ck.

Quando houver a necessidade de contar o suporte de um conjunto de itens candidato, é então lido o conjunto Ck, e não todo o conjunto de transações, como ocorria com o algoritmo *Apriori*, o que significa a leitura de um conjunto menor de elementos, otimizando o tempo de execução do algoritmo *Apriori-TID*.

Pseudocódigo do algoritmo *Apriori-TID* [AGR 94].

```
L1 = {large 1-itemsets};
C1 = database D;
for (k=2; Lk-1 ≠ O) ; k++) do begin
        Ck = apriori_gen(Lk-1); // Novos candidatos
        Ck = 0;
        forall entries t ∈ Ck-1 do begin
                // determinar conjuntos de itens candidatos em Ck
                // contidos na transação com identificador t.TID
                Ct = {c ∈ Ck | (c - c[k] ) ∈ t.set-of-itemsets v (c - c[k-1]) ∈ t.set-of-
        itemsets};
                forall candidates c ∈ Ct do
                        c.count++;
                if (Ct ≠ 0) then Ck += <t.TID,Ct>;
        end
        Lk = {c ∈ Ck | c.count >= minsup};
end
Answer = Uk Lk;
```

4.5 Apriori-Hybrid [AGR 94]

Este algoritmo tem como base a idéia de que não é necessário o uso do mesmo algoritmo em todos os passos do processamento. É verificado que o algoritmo *Apriori* executa mais rapidamente o processo de mineração nas passagens iniciais que o algoritmo *Apriori-TID*, enquanto que, após um certo número de passagens, o *Apriori-TID* passa a ser mais rápido que *Apriori*.

Com base em observações experimentais, é proposto o algoritmo *Apriori-Hybrid*, que combina a utilização de *Apriori* nas passagens iniciais e *Apriori-TID* nas passagens subseqüentes, quando o conjunto C_k puder ser alocado em memória. Contudo, a mudança de um algoritmo para o outro envolve um custo que deve ser assumido, que acontece na decisão da mudança no fim da passagem k, na passagem (k+1), após a pesquisa dos conjuntos de itens candidatos contidos em uma transação, deve-se acrescentar o ID da transação ao conjunto Ck+1, e então, a partir da passagem (k+2), é possível mudar para *Apriori-TID*.

4.6 *Dense – Miner* [BAY 99]

O algoritmo Dense - Miner aplica todos os limites definidos pelo usuário durante a execução do algoritmo de mineração, a fim de melhorar a eficiência em dados densos, tais como dados relacionais. Visto que as regras, satisfazendo os limites de suporte e confiança em bancos de dados densos, obtidas através da aplicação de algoritmos

convencionais, são freqüentemente muito numerosas para serem mineradas eficientemente ou compreendidas pelo usuário final, o algoritmo *dense – miner* explora outros limites que eliminam regras que não são interessantes porque elas contêm condições que não contribuem decisivamente para a habilidade preditiva da regra, estes limites considerados seriam a melhoria e o limite de conseqüente.

Como exemplo temos a seguinte regra:

Pão & Manteiga → Leite (Confiança de 80%)

Esta regra tem a confiança de 80%, e é lida da seguinte forma, 80% das pessoas que compram pão e manteiga também compram leite, esta regra possui uma grande confiança, portanto é válida para o objetivo de verificar a população de compradores de leite. Contudo se 85% da população que está sendo examinada compra leite, esta regra não seria interessante para o propósito de verificar o perfil dos compradores de leite, pois sabemos que a maioria compra leite em todas a suas compras.

Ainda levando em conta a população de compradores de leite (85%) temos outra regra:

Ovos & Cereais → Leite (Confiança de 95%)

Esta regra tem a confiança de 95%, portanto, podemos considerá-la válida sendo que a confiança dos compradores de leite é de 85%, menor do que a confiança da regra. Mas supondo que os Cereais sozinhos tendo como conseqüente os compradores de leite seja de 99% a regra descrita acima não seria interessante, pois existem mais pessoas que compram cereais do que pessoas que compram cereais e ovos.

No caso do algoritmo *dense - miner*, há dois modos propostos para se resolver o problema de dados densos, são eles: a limitação de conseqüente e a limitação de melhora mínima ou *minimum improvement*.

A limitação de conseqüente consiste no fato de que uma regra minerada seja considerada, somente se ela tiver um dado conjunto de itens conseqüente C definido pelo usuário. Este conjunto de itens C deve ser composto por no mínimo 1 elemento. Dentre as finalidades deste limite está redução do conjunto de itens freqüentes considerados.

Através da limitação de melhora mínima ou *minimum improvement* é permitido ao usuário especificar um limite de melhora mínima, denominado de minimp, tal que uma regra que tenha uma confiança, com ao menos minimp, maior que a confiança de qualquer uma das suas sub - regras, pois uma sub - regra é uma simplificação da regra formada pela remoção de um ou mais condições de seu antecedente. À medida que o usuário aumenta o valor de minimp torna-se mais conciso o conjunto de regras geradas, sendo tal conjunto formado somente por itens que são decisivamente importantes para a habilidade preditiva da regra.

Um algoritmo que roda eficientemente em bancos de dados densos, sem restrições adicionais, pode fornecer um resultado final com um conjunto extremamente grande de regras, sem indicação de quais destas regras são boas. Tem-se a melhora de uma regra como sendo a diferença mínima entre sua confiança e a confiança de qualquer de suas sub–regras com o mesmo conseqüente.

Uma regra com melhoria negativa é geralmente indesejável, porque ela pode ser simplificada para produzir uma sub–regra que é mais preditiva, e se aplica a uma população igual ou maior. Uma regra com melhoria igual a 0 é, na maioria dos casos, desejável e uma regra com melhoria positiva é completamente desejável, porque a maioria das regras em bancos de dados densos não são úteis apresentando um decremento de confiança.

Suponhamos que uma regra será representada, usando apenas seu conjunto de itens antecedente, pois, assume-se que o conseqüente é um conjunto de itens C. Temos que U é o conjunto de todos os itens presentes no banco de dados exceto aqueles presentes no conseqüente da regra. O problema de mineração é a força do conjunto de U regras que satisfaçam o limite de suporte, confiança e melhoramento mínimos. Este algoritmo baseia-se na estrutura de árvore de enumeração [BAY99], para reduzir o espaço de regras consideradas. A idéia é primeiro impor uma ordem ao conjunto de itens, e então enumerar conjuntos de itens de acordo com a ordenação. A Figura 4.4 mostra uma árvore completamente expandida sobre U = {1,2,3,4,5}, com itens ordenados lexicamente.

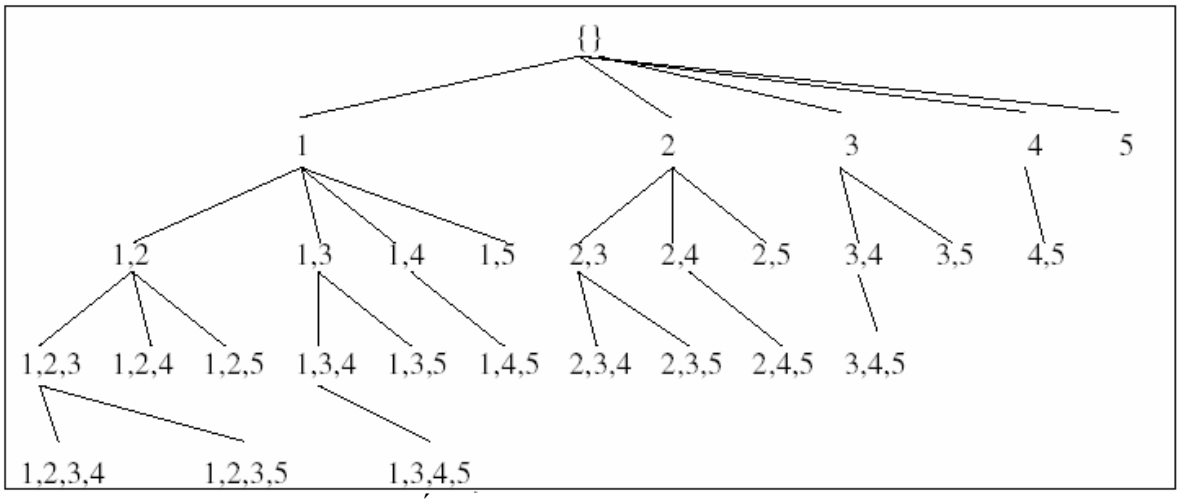


Figura 4.4: Árvore de enumeração de conjunto.

4.7 MiRABIT [CAM 02]

MiRABIT, apresentado em [CAM 02], cuja principal característica é a geração dos conjuntos de itens candidatos para a passagem *k* feita durante a passagem *k* e não durante a passagem (*k*-1) tal qual o algoritmo *Apriori* [AGR 93] e similares. O principal objetivo desta característica é melhorar o desempenho da tarefa de mineração de regras de associação em bancos de dados com um baixo tamanho médio de transações. Como conseqüência desta alteração, o algoritmo MiRABIT gera um conjunto de itens candidatos geralmente entre 0,5 e 2 % do conjunto de itens gerados pelo *Apriori* [AGR93] no caso em estudo, o que resulta em um relevante ganho de desempenho.

```
F1 = {conjuntos de itens freqüentes com 1- elemento};
for (k = 2; (( |Fk-1| ³ 2) or (k £ tamanho_maximo_regra)); k++) do begin
  forall transações t Î D do begin
    Ct = subset (Fk-1,t,k); // Gera possíveis candidatos na transação
    forall candidatos c Î Ct do
      c.count ++;
    Ck=Ck È Ct;
  end
  Fk = {c Î Ck | c.count ³ suporte_minimo}
  Answer=Answer È Fk;
end
```

A primeira linha do pseudocódigo apresentado demonstra que a execução do algoritmo parte do conjunto de itens freqüentes com 1 elemento. Para gerar este conjunto, o algoritmo necessita passar uma vez sobre o banco de dados contando em quantas transações cada item está presente. Para que um item seja considerado freqüente é necessário que seu suporte seja maior que um limiar previamente definido pelo usuário, chamado suporte mínimo.

A segunda linha do pseudocódigo define uma estrutura de controle de iteração. Esta iteração será controlada pela variável *k* que identifica o número da passagem do algoritmo sobre o banco de dados. A variável *k* tem seu valor inicial definido como 2 e a iteração será executada enquanto o conjunto de itens freqüentes gerado na passagem anterior tiver 2 ou mais elementos. A iteração deixará de ocorrer se o conjunto tiver menos de dois elementos pois será feita uma combinação de *n* itens em conjuntos com *p* itens, logo *p* não poderá ser menor que *n*. Outra condição para a iteração ser executada é que *k* seja menor ou igual ao valor definido pelo usuário como tamanho máximo da regra. Esta restrição limita a quantidade de passagens do algoritmo sobre o banco de dados, diminuindo também o tempo para execução do algoritmo. A cada nova iteração, o valor da variável *k* é incrementado em uma unidade.

A cada nova passagem sobre o banco de dados diminui a probabilidade de ser encontrado um conjunto de itens freqüente; por este fato, a utilização de uma restrição de tamanho máximo da regra pode ser muito importante. Tendo-se um banco de dados, a probabilidade de um item qualquer ser freqüente é determinada pela razão entre o número de itens freqüentes na primeira passagem e o número total de itens; este resultado é elevado à ordem da passagem.

A terceira linha do pseudocódigo define uma nova estrutura de controle de iteração que obriga a execução das linhas seguintes para todas as transações no banco de dados. Neste ponto é importante salientar que a seleção, que é um dos passos do pré-processamento, já foi executada, sendo portanto consideradas somente as transações que atendem às restrições pré-estabelecidas.

A quarta linha do pseudocódigo executa a função *subset* que terá seu funcionamento detalhado posteriormente. A função *subset* gera todos os conjuntos de itens candidatos para cada transação.

Na quinta linha do pseudocódigo é definida uma estrutura de controle de iteração que obriga a execução das linhas posteriores para todos os conjuntos de itens gerados anteriormente pela função *subset*.

Na sexta linha é definida a instrução para que o contador de cada conjunto de itens encontrado na transação, seja incrementado em uma unidade. Na sétima linha o conjunto de itens candidatos encontrado na transação (*Ct*) é inserido no conjunto de itens candidatos da passagem.

Na oitava linha é finalizada a estrutura de controle iniciada na terceira linha. Na nona linha do pseudocódigo é definida a instrução que irá incluir os conjuntos de itens candidatos com suporte maior ou igual ao suporte mínimo no conjunto de itens freqüentes da passagem. Esta linha será executada após a análise de todas as transações do banco de dados selecionadas para serem mineradas.

Na décima linha do pseudocódigo é definida a instrução que irá inserir os conjuntos de itens freqüentes na passagem *k* ao conjunto resposta. Ao final da execução do algoritmo o conjunto resposta conterá todos os conjuntos de itens freqüentes encontrados em cada uma das passagens do algoritmo sobre o banco de dados juntamente com seus respectivos contadores de suporte.

Na décima primeira linha do algoritmo é definido o fim da estrutura de controle iniciada na segunda linha e finalizada a execução do algoritmo. A partir dos conjuntos de itens freqüentes são geradas as regras de associação que satisfaçam a restrição de confiança mínima definida pelo usuário.

No algoritmo MiRABIT a função *subset* tem como finalidade principal identificar que conjuntos de itens em uma transação são candidatos para posterior contagem destes conjuntos de itens. Esta função recebe como argumentos o conjunto de itens freqüentes na passagem anterior *Lk-1* e a transação atual *t*. Cada conjunto de itens *c* contido na transação que seja derivado de um conjunto de itens freqüentes na passagem anterior é incluído no conjunto de itens candidatos da transação *Ct* para sua posterior contagem.

if (**all** (*k*-1) – subsets **of** *c*)Î *Fk-1* **then**

c Î *Ct*;

A instrução anterior apresenta a função *subset*.

A função *subset* gera as possíveis combinações de itens em conjuntos de *k* itens e também implementa uma função de poda, sendo que para cada conjunto de itens gerado em uma transação, é verificado se algum de seus subconjuntos não é freqüente; em caso positivo, este conjunto de itens não é considerado candidato, e por conseqüência, seu suporte não é contado.

Visto que o conjunto de itens gerados é muito menor que no algoritmo *Apriori*, conforme demonstrado em experimentos, a quantidade de memória utilizada durante o processo também é menor.

5 PROPOSTA

Neste capítulo são apresentadas as etapas para a obtenção da base de dados unindo os dados socioeconômicos aos dados de internações hospitalares, bem como as descrições de suas origens. É apresentado também o protótipo de uma ferramenta de mineração de dados baseada no algoritmo *Apriori,* que implementa o uso de *Lift* e de *Improvement* para elevar a capacidade preditiva da ferramenta. É implementada ainda uma característica que permite ao usuário estabelecer um campo como objetivo, eliminando assim associações indesejáveis.

5.1 Objetivos da Mineração

As minerações têm o objetivo principal de encontrar associações que fugissem do padrão esperado pelo especialista e analisá-las em profundidade a nível granular verificando, possíveis desvios de entendimento.

Para se realizar minerações mais consistente foram selecionados alguns atributos que as nortearam:

- Verificar as associações entre as características físicas e socioeconômicas relacionadas com os procedimentos realizados nas internações hospitalares.
- Verificar as associações entre as características físicas e socioeconômicas relacionadas com o caráter das internações hospitalares.
- Verificar as associações entre as características físicas e socioeconômicas relacionadas com a clínica médica das internações hospitalares.

5.2 Base de Dados

A base de dados a ser analisada contém informações de duas fontes distintas de dados que são:

- Base de Dados do CadSUS (Cartão Nacional de Saúde), que registra os dados socioeconômicos da população.
- Base de Dados do CLEITOS (Sistema de Central de Leitos), que registra alguns dados referentes à internação do paciente.

Os dados utilizados nos experimentos são provenientes das internações dos três últimos anos, e foram fornecidos pela Autarquia do Serviço Municipal de Saúde de Londrina – PR, responsável pelo gerenciamento destas informações.

Na seção 5.2.1 serão explicadas mais detalhadamente as fontes de dados utilizadas em conjunto com o protótipo, e também os procedimentos realizados nos dados para torná-los passíveis de mineração.

5.2.1. CadSUS (Cartão Nacional de Saúde)

O Cadastro Nacional de Usuários é o primeiro passo para a implantação do Cartão Nacional de Saúde em todo o território nacional. O Cartão será uma importante ferramenta para a consolidação do Sistema Único de Saúde (SUS), facilitando a gestão do sistema e contribuindo para o aumento da eficiência no atendimento direto ao usuário.

A realização de um cadastramento domiciliar de base nacional, aliado à possibilidade de manutenção dessa base cadastral atualizada, pode permitir aos gestores do SUS a construção de políticas sociais integradas e intersetoriais (educação, trabalho, assistência social, tributos e outros) nos diversos níveis de governo.

Nos campos de atendimento aos usuários do SUS e de organização do sistema de saúde, o cadastramento é condição para a implantação do Cartão Nacional de Saúde. O Cartão contribui para o desenvolvimento de ações programáticas estratégicas, ações de vigilância epidemiológica, assistência ambulatorial e hospitalar, fortalecimento dos sistemas de referência e contra-referência, controle e avaliação, dentre outras.

O cadastramento consiste no processo por meio do qual são identificados os usuários do Sistema Único de Saúde e seus domicílios de residência. Por meio do cadastro será possível a emissão do Cartão Nacional de Saúde para os usuários e a vinculação de cada um ao seu domicílio de residência, permitindo uma maior eficiência na realização das ações de natureza individual e coletiva, desenvolvidas nas áreas de abrangência dos serviços de saúde.

O cadastramento permite ainda a construção de um banco de dados para diagnóstico, avaliação, planejamento e programação das ações de saúde.

Para a atenção básica, este formato de cadastramento tem como vantagens:

- Fortalecimento do vínculo entre indivíduos e Unidade Básica de Saúde, por meio da oferta organizada de serviços e do acompanhamento pelos profissionais da rede básica na trajetória dos indivíduos na mesma.
- Possibilidade de trabalhar com enfoque na vigilância à saúde, por meio do diagnóstico das condições de saúde e do perfil epidemiológico da população adscrita da área de abrangência da unidade básica, favorecendo a intervenção sobre os fatores e grupos de riscos existentes.
- Melhoria da qualidade do atendimento, pelo acesso às informações sobre utilização de serviços pelos pacientes nos diversos níveis de atenção.
- Potencialização das atividades de vigilância epidemiológica e sanitária, por meio da localização espacial de casos e contatos domiciliares, bem como de faltosos aos programas desenvolvidos na rede básica, facilitando a realização de ações de busca ativa, vacinação de bloqueio, acompanhamento domiciliar, tratamento supervisionado, entre outras, de modo ágil e oportuno.

- Melhoria da qualidade dos sistemas de informação cujos dados são gerados no âmbito das unidades básicas, pela possibilidade de individualização dos registros e delimitação da população, com a produção de indicadores com maior precisão e conseqüente potencialização da avaliação dos processos e resultados das ações desenvolvidas no âmbito da atenção básica.

O sistema CadSUS foi desenvolvido com a ferramenta de desenvolvimento Borlad Delphi, utilizando como base de dados Borland Interbase.

5.2.1.1 Modelo ER do Sistema CadSUS

O modelo de entidade e relacionamento do sistema foi construído com base nas principais tabelas do sistema, outras estruturas foram suprimidas, pois não tinham utilidade para o projeto.

Para se construir o modelo, a base de dados original do banco de dados Borland Interbase foi migrada para uma base de dados local do Microsoft Access onde foi possível obter uma melhor documentação das estruturas.

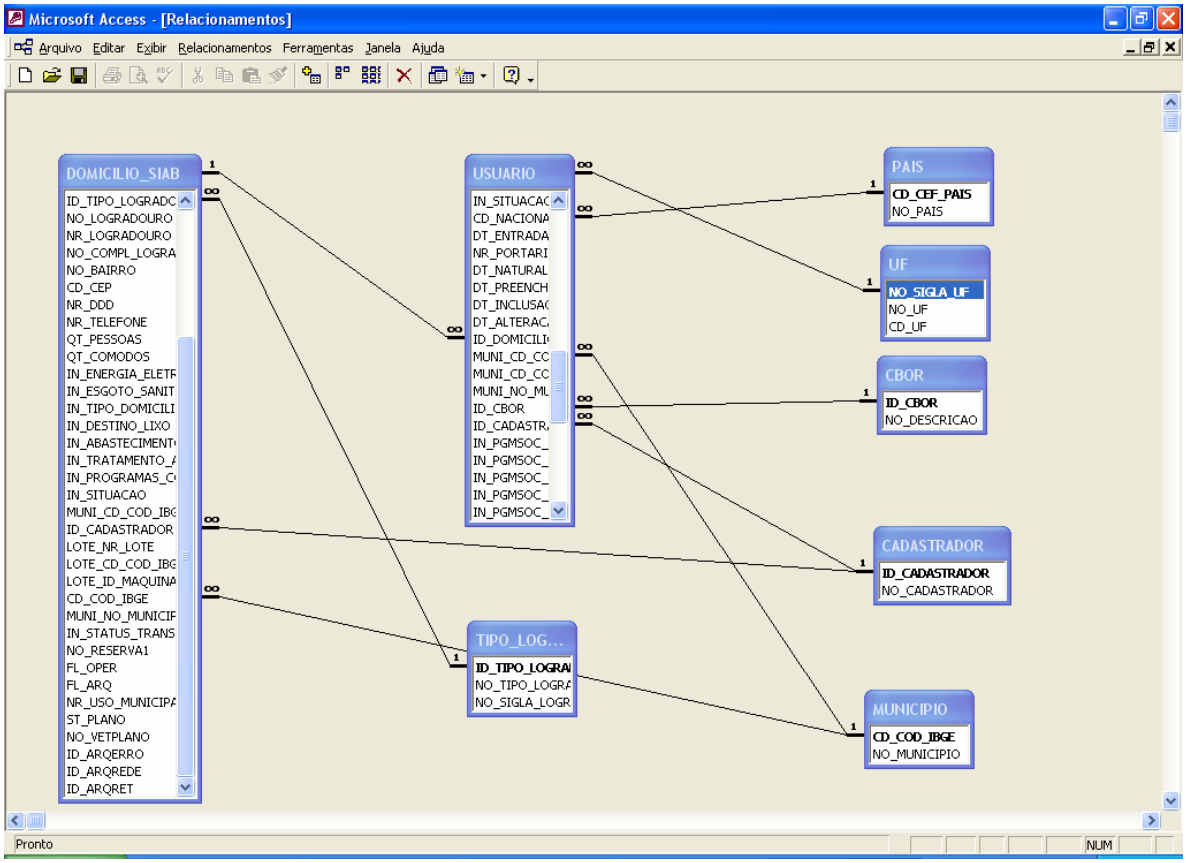


Figura 5.5: ER do sistema CadSUS.

As tabelas do modelo da Figura 5.5 foram documentadas e detalhadas e são mostradas no Anexo I, utilizando a ferramenta de análise do próprio banco de dados.

5.2.2. CLEITOS (Sistema de Central de Leitos)

O CLEITOS - Central de Leitos foi desenvolvido pela Polo de Software de Curitiba S/A para ser implantado na Secretaria Municipal de Saúde daquela cidade.

Este sistema é destinado a apoiar a administração municipal no controle do estoque de leitos em hospitais e no acompanhamento de internações hospitalares(AIH) em hospitais conveniados ao SUS. É composto de 5 módulos: Cadastros, Laudos, AIHs, Relatórios, Utilidades.

O sistema CLEITOS foi desenvolvido em linguagem de programação CA Clipper, utilizando como fonte de dados arquivos dBASE (*.dbf).

5.2.2.1 Modelo ER do Sistema CLEITOS

O modelo de entidade e relacionamento do sistema foi construído com base nas tabelas das principais tabelas do sistema. Outras estruturas foram suprimidas do sistema, pois não tinham utilidade para o projeto.

Para se construir o modelo a base de dados original dos arquivos dBASE foi migrada para uma base de dados local do Microsoft Access, onde foi possível apresentar uma melhor documentação das estruturas.

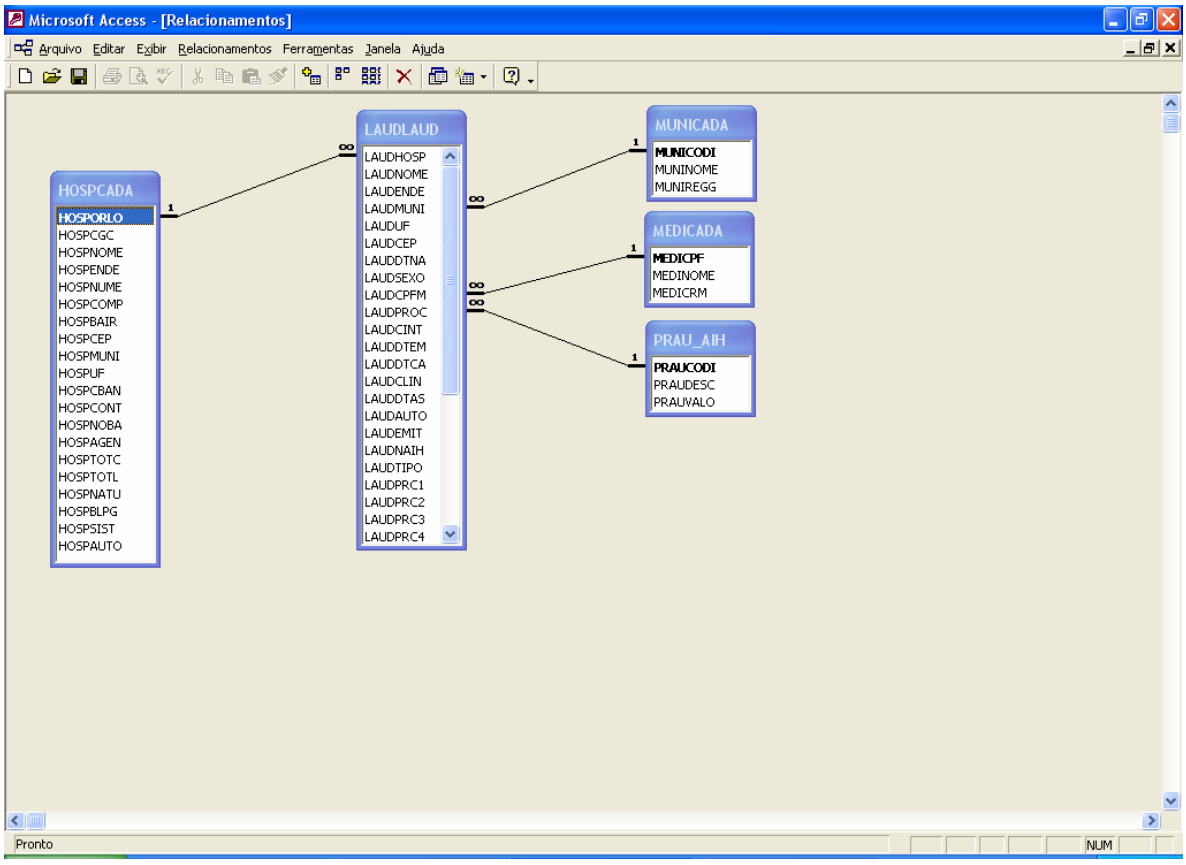


Figura 5.6: ER do sistema CLEITOS.

As tabelas do modelo da Figura 5.6 foram documentadas e detalhadas como estão mostrado no Anexo II, utilizando a ferramenta de análise do próprio banco de dados.

5.2.3. Fases para Obtenção da Base de Dados

Nesta seção serão explicadas as fases para obtenção da base de dados utilizada no protótipo, como a união dos dados dos dois sistemas envolvidos e a discretização dos campos com faixas de valores muito amplos.

5.2.3.1 Fase de Seleção

Nesta fase foi realizada a segmentação dos dados dos sistemas. Para este intuito, os dados originais do sistema CadSUS foram importados para uma base de dados Microsoft Access chamada de repositório temporário como mostra o esquema na Figura 5.7.

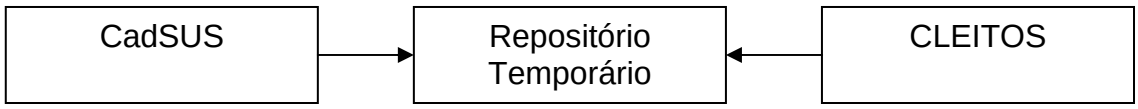


Figura 5.7: Esquema de seleção da base de dados para o protótipo.

Um estudo preliminar realizado junto com um especialista indicou quais seriam as tabelas onde existiria a possibilidade de realização de cruzamentos em ambos os sistemas. O passo seguinte seria a união dos dados dos dois sistemas, mas a falta de um identificador comum entre ambos poderia ser um problema.

O problema do identificador comum foi resolvido utilizando a seguinte estratégia:

- Os campos **Nome do Paciente,** em ambos os sistemas foram segmentados em dois outros campos, **Primeiro Nome** e **Ultimo Nome**.
- Foi realizada então uma consulta SQL fazendo a junção das tabelas dos dois sistemas pelo **Primeiro Nome**, **Ultimo Nome** e **Data de Nascimento**.
- O resultado da consulta foi armazenado em uma nova tabela.

Como resultado da consulta foram gerados 56.539 registros, originados de 236.540 registros do sistema CadSUS e 270.496 registros do sistema CLEITOS. Com a base em mãos o próximo passo seria a verificação da integridade dos registros.

5.2.3.2 Pré-processamento

Nesta fase foram verificados todos os campos gerados pela união dos sistemas, a fim de encontrar dados incompletos ou que não tinham valor para o projeto. Foram eliminados todos os campos que tinham utilidade interna para os sistemas, e também os campos cujos conteúdos não tinham representatividade dentro da base.

Os registros que possuíam dados incompletos ou ruídos foram excluídos para não causar interferência nas minerações, como conseqüência desta exclusão foi registrado o seguinte resultado para cada campo:

- RACA_COR, 14 registros contendo ruídos ou dados incompletos.
- FREQUENTA_ESCOLA, 615 registros contendo ruídos ou dados incompletos.
- CD_SEGMENTO, 17 registros contendo ruídos ou dados incompletos.
- LAUDCINT, 159 registros contendo ruídos ou dados incompletos.

Como saldo, após as exclusões, restaram 55.734 registros.

Também nesta fase foram realizadas as discretizações nos campos onde os valores assumiam um domínio muito vasto. Os campos em que a discretização foi realizada são:

- Campo FAIXAETARIA, este campo foi obtido primeiramente pelo cálculo da idade do paciente que foi realizado, cruzando os campos de data do nascimento, e data do cadastro do laudo, posteriormente foram estabelecidas faixas para este campo com base no sistema SIAB do Ministério da Saúde:

    - 0/4A – 0 a 4 Anos
    - 5/9A – 5 a 9 Anos
    - 10/14A – 10 a 14 Anos
    - 15/19A – 15 a 19 Anos
    - 20/24A – 20 a 24 Anos
    - 25/29A – 25 a 29 Anos
    - 30/34A – 30 a 34 Anos
    - 35/39A – 35 a 39 Anos
    - 40/44A – 40 a 44 Anos
    - 45/49A – 45 a 49 Anos
    - 50/54A – 50 a 54 Anos
    - 55/59A – 55 a 59 Anos
    - 60/64A – 60 a 64 Anos
    - 65/69A – 65 a 69 Anos
    - 70/74A – 70 a 74 Anos
    - 75/79A – 75 a 79 Anos
    - 80/+A – 80 ou mais Anos

- Campo FAIXAPESSOAS, este campo foi obtido com base no campo do sistema CadSUS de quantidade de pessoas no domicílio, foram estabelecidas faixas para ele, com se segue:

    - 1/2P – 1 a 2 Pessoas
    - 3/4P – 3 a 4 Pessoas
    - 5/7P – 5 a 7 Pessoas
    - 8+P – 8 ou mais Pessoas

As faixas foram obtidas através da distribuição dos valores para este campo dentro da base de dados, onde cada faixa possui quantidades de registros aproximadas.

- Campo FAIXACOMODOS, este campo foi obtido com base no campo do sistema CadSUS de quantidade de cômodos no domicílio, Foram estabelecidas faixas para este campo com base no sistema SIAB do Ministério da Saúde:

    - 1C – 1 Cômodo
    - 2C – 2 Cômodos
    - 3C – 3 Cômodos
    - 4+C – 4 ou mais Cômodos

Como resultados dos passos anteriores e também através da interação com o especialista, foram selecionados os seguintes campos para se realizar as minerações:

Tabela 5.9: Fonte de Dados para o Protótipo.

Campo	Descrição	Origem
RACA_COR	Raça (cor) do paciente.	CadSUS
SITUACAO_CONJUGAL	Situação familiar ou conjugal do paciente.	CadSUS
ESCOLARIDADE	Escolaridade do paciente.	CadSUS
FAIXAETARIA	Faixa etária do paciente	CadSUS
SEXO	Sexo do paciente.	CadSUS
FREQUENTA_ESCOLA	Se o paciente freqüenta da escola.	CadSUS
CBOR	Código brasileiro de ocupações.	CadSUS
BAIRRO	Bairro de residência do paciente.	CadSUS
CEP	Código de endereçamento postal.	CadSUS
FAIXAPESSOAS	Faixa de número de pessoas que reside no domicílio.	CadSUS
FAIXACOMODOS	Faixa de número de cômodos da residência.	CadSUS
ENERGIA_ELETRICA	Se possui energia elétrica.	CadSUS
ESGOTO_SANITARIO	Tipo de esgotamento sanitário	CadSUS
TIPO_DOMICILIO	Tipo de domicílio.	CadSUS
DESTINO_LIXO	Destino do lixo.	CadSUS
IN_ABASTECIMENTO_AGUA	Tipo de abastecimento de água.	CadSUS
TRATAMENTO_AGUA	Tipo de tratamento de água	CadSUS
LAUDHOSP	Hospital em que foi internado.	CLEITOS
CAUDCPFM	CPF do médico atendente.	CLEITOS
LAUDCINT	Caráter da internação.	CLEITOS
LAUDCLIN	Clínica médica.	CLEITOS
LAUDPROC	Procedimento principal realizado.	CLEITOS

Fonte: Resultados dos passos anteriores.

5.3 Ferramenta Desenvolvida

Para a realização dos experimentos descritos no capítulo 6, foi desenvolvida uma ferramenta que implementa o algoritmo *Apriori* apresentada em [AGR 94] e também implementa o uso de *Lift* e *Improvement* apresentados em [BAY 99].

A ferramenta se baseia em um banco de dados, em que os dados já foram selecionados, pré-processados e transformados. Devido à natureza dos dados é utilizado o método Regras de Associação Quantitativas apresentado por [SRI 96], cujo domínio quantitativo, não binário, pode ser utilizado para mineração de regras, bastando mapear este domínio quantitativo e/ou categorizado para o domínio binário.

A ferramenta apresenta as regras mineradas e ao final a quantidade de regras encontradas, baseada nos limites estabelecidos pelo usuário. A seguir são apresentadas as características do protótipo da ferramenta desenvolvida.

5.3.1. Requisitos de Hardware

O protótipo implementado tem como requisitos mínimos de hardware um microcomputador IBM-PC compatível 486, com 16Mb de memória RAM, e espaço disponível em disco de 50% do tamanho do banco de dados analisado. Sendo

recomendada a utilização de um Pentium com 32 Mb de memória RAM (ou mais) para possibilitar a exploração de bases de dados de maior porte.

A adoção da plataforma IBM-PC se deu porque este tipo de maquina é tido como padrão para os sistemas do Ministério da Saúde.

5.3.2. Requisitos de Software

O protótipo necessita de um equipamento com ambiente Windows e o Microsoft Data Access, este último é uma ferramenta de acesso às bases de dados Microsoft Access. Já foram executados testes de compatibilidade com os sistemas operacionais Windows 98 e Windows ME, Windows 2000 e Windows XP. A resolução de vídeo recomendada é de 800 x 600 pixels.

A adoção do Windows se deu porque este tipo de sistema operacional é tido como padrão para os sistemas do Ministério da Saúde.

5.3.3. Arquitetura da Ferramenta

O protótipo foi construído com base na ferramenta de desenvolvimento Borland Delphi, utilizando como base de dados relacional o Microsoft Access. As tabelas utilizadas pela ferramenta de mineração são as seguintes:

- Tabela de fonte de dados;
- Tabela de seleção dos dados;
- Tabelas de controle dos níveis das regras.
- Tabela de regras.

A tabela de fonte de dados possui todos os campos que seriam potencialmente interessantes para a mineração.

A tabela de seleção dos dados funciona de forma dinâmica a partir do uso do protótipo, onde o usuário seleciona os campos que deseja minerar e a ferramenta gera a tabela de seleção com base na tabela de fonte de dados somente com os campos selecionados.

Este procedimento foi feito para não sobrecarregar a memória do equipamento, pois se fosse utilizada uma subconsulta, o uso da memória e do processamento seria mais intensivo.

As tabelas de controle de níveis de regra são geradas dinamicamente pelo protótipo multiplicando seus conteúdos para obter o próximo nível das regras e têm como objetivo armazenar as regras em potencial com o seu suporte dentro da base de dados.

As tabelas de controle de níveis podem ser descritas da seguinte maneira:

Tabela 5.10: Controle de Níveis de Regras.

Campo	Descrição
CONJUNTO_1	Identificador do conjunto da regras.
CAMPO_1	Primeiro campo da regra na base selecionada.
VALOR_1	Valor do primeiro campo na base selecionada.
...	
CONJUNTO_n	Identificador do enésimo conjunto de regras.
CAMPO_n	Enésimo campo da regra na base selecionada.
VALOR_n	Valor do enésimo campo na base selecionada.
SUPORTE	Suporte a esta regra.

Fonte: Tabelas de controle de níveis.

A tabela de regras, armazena a regra, seu nível e confiança para o uso de *Improvement.*

5.3.4. Operação do Protótipo

O protótipo do sistema possui uma interface extremamente simples, sendo necessário ao usuário apenas um pequeno conhecimento sobre os conceitos de regras de associação para que o mesmo consiga utilizar o sistema. O usuário deve fornecer ao sistema os campos que deseja minerar, o valor de suporte mínimo, confiança mínima, *Lift*, *Improvement* (melhoramento mínimo) e o número de níveis (tamanho máximo da regra). A tela principal do protótipo é apresentada na Figura 5.8.

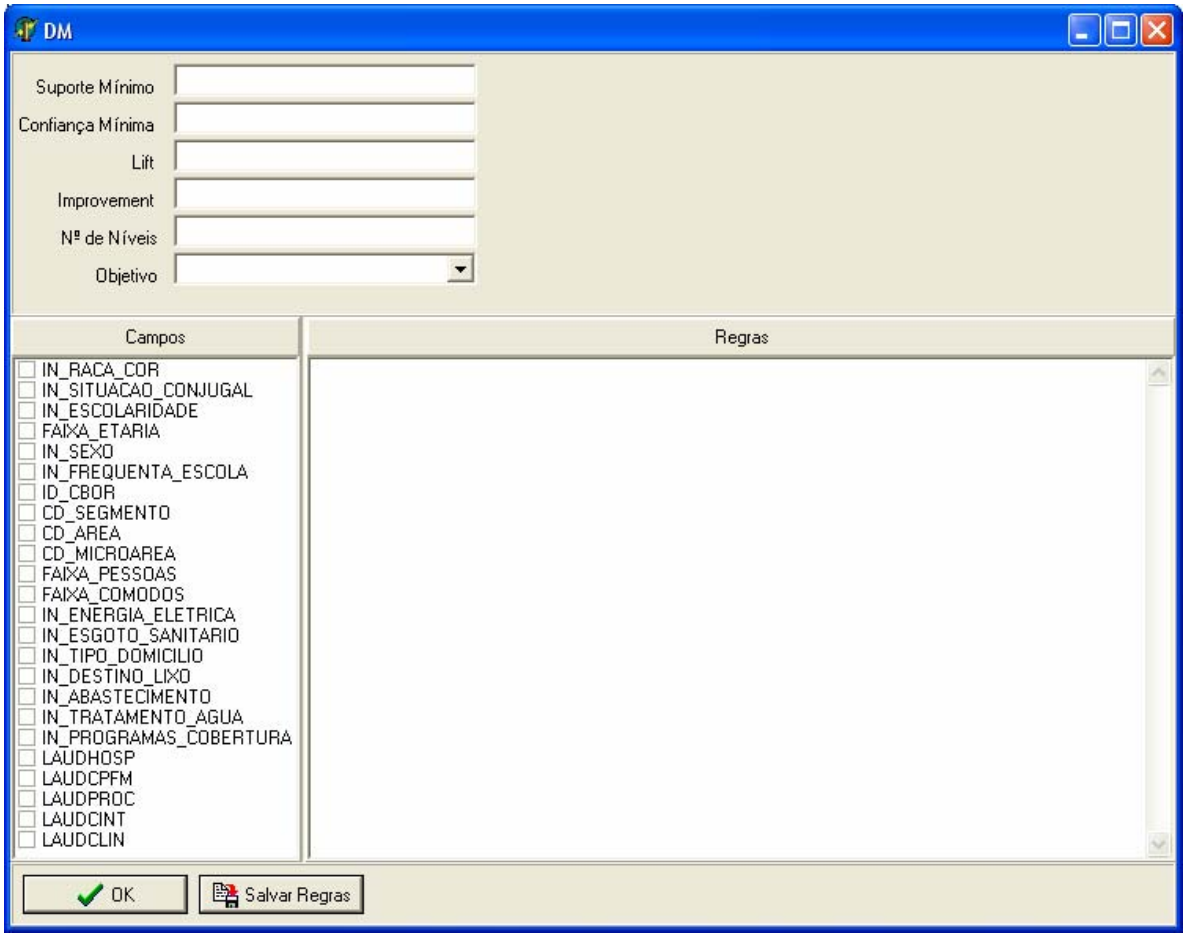


Figura 5.8: Tela principal do protótipo.

O cabeçalho da tela ou área de parâmetros é onde o usuário informa o suporte mínimo para que um itemset seja considerado freqüente. Confiança mínima é o valor mínimo de confiança para que uma regra seja gerada. *Lift* e *Improvement* para regra ser gerada. Tamanho máximo da regra é o número máximo de itens que devem existir em uma regra, este parâmetro também delimita o número máximo de passagens sobre o banco de dados, delimita os dados que devem ser selecionados e os parâmetros para o processo de mineração. Para que o número de regras seja reduzido, o usuário pode selecionar sua meta de mineração no campo objetivo, caso não seja selecionado o objetivo, o protótipo funciona normalmente encontrando todas as possíveis regras. O usuário pode selecionar os campos que deseja minerar na área situada à esquerda da tela, este recurso é útil nas situações em que se tem um domínio pré-estabelecido para mineração. Para iniciar o processo de mineração o comando está localizado no rodapé da tela.

Assim que o processo de mineração estiver concluído o protótipo mostrará na área da tela, localizada à direita, todas as regras que foram encontradas, obedecendo aos critérios pré-estabelecidos, todas a regras se apresentam na seguinte descrição:

{ANTECEDENTE} ==> CONSEQUENTE --- conf(CONFIANÇA) --- sup(SUPORTE) --- lift(LIFT)

A Figura 5.9 mostra um exemplo das regras geradas pelo protótipo com base nos parâmetros selecionados.

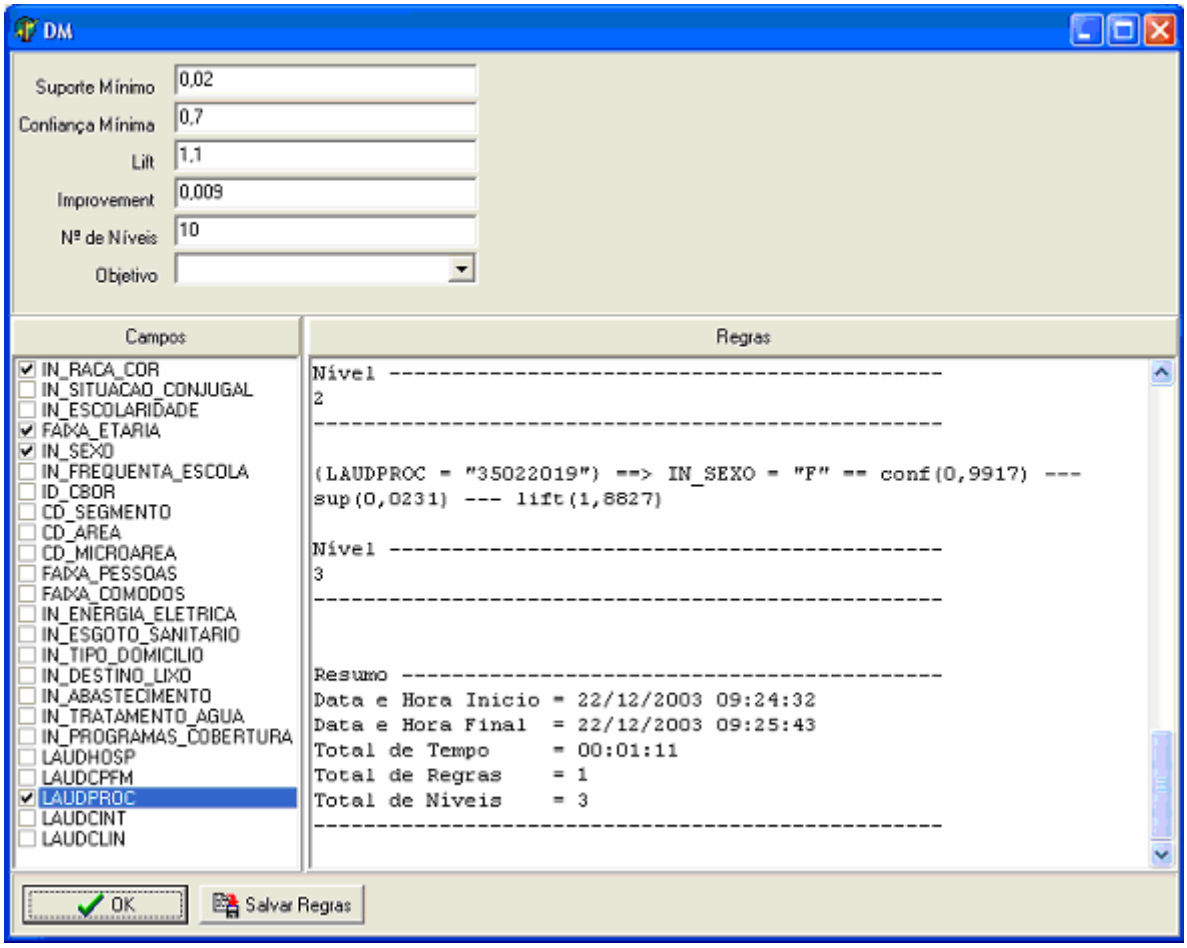


Figura 5.9: Tela do protótipo com regras geradas.

A Figura 5.10 demonstra o relatório gerado pelos parâmetros, suporte mínimo igual 0,02, confiança mínima igual 0,7, *lift* igual 1, *improvement* igual 0,009 e número de níveis igual a 10 e com os campos selecionados RACA_COR, FAIXAETARIA, SEXO, LAUDPROC, como descritos na Figura 5.9 e gravados em um arquivo texto pela opção Salvar Regras. As linhas geradas no nível 1 apenas relatam o suporte para cada valor de campo selecionado, cujo suporte supere ou iguale o suporte mínimo estabelecido.

```
Parâmetros ----------------------------------------
Suporte Mínimo     = 0,02
Confiaça Mínima    = 0,7
Lift Mínimo = 1,1
Improvement Mínimo = 0,009
Nº de Níveis       = 10
-----------------------------------------------------

Data e Hora do Início da Mineração ----------------

Data e Hora        = 22/12/2003 09:24:32

-----------------------------------------------------

Nível ----------------------------------------------
1
-----------------------------------------------------

{IN_RACA_COR = "1"} == --- sup(0,7748)
{IN_RACA_COR = "2"} == --- sup(0,0509)
{IN_RACA_COR = "4"} == --- sup(0,1642)
{FAIXA_ETARIA = "0/4A"} == --- sup(0,1307)
{FAIXA_ETARIA = "10/14A"} == --- sup(0,0384)
{FAIXA_ETARIA = "15/19A"} == --- sup(0,0565)
{FAIXA_ETARIA = "20/24A"} == --- sup(0,068)
{FAIXA_ETARIA = "25/29A"} == --- sup(0,0694)
{FAIXA_ETARIA = "30/34A"} == --- sup(0,0719)
{FAIXA_ETARIA = "35/39A"} == --- sup(0,0688)
{FAIXA_ETARIA = "40/44A"} == --- sup(0,0662)
{FAIXA_ETARIA = "45/49A"} == --- sup(0,0641)
{FAIXA_ETARIA = "5/9A"} == --- sup(0,0575)
{FAIXA_ETARIA = "50/54A"} == --- sup(0,0544)
{FAIXA_ETARIA = "55/59A"} == --- sup(0,0524)
{FAIXA_ETARIA = "60/64A"} == --- sup(0,0546)
{FAIXA_ETARIA = "65/69A"} == --- sup(0,0483)
{FAIXA_ETARIA = "70/74A"} == --- sup(0,0434)
{FAIXA_ETARIA = "75/79A"} == --- sup(0,0291)
{FAIXA_ETARIA = "80/+A"} == --- sup(0,0262)
{IN_SEXO = "F"} == --- sup(0,5268)
{IN_SEXO = "M"} == --- sup(0,4732)
{LAUDPROC = "35022019"} == --- sup(0,0233)
{LAUDPROC = "77500113"} == --- sup(0,021)

Nível ----------------------------------------------
2
-----------------------------------------------------

{LAUDPROC = "35022019"} ==> IN_SEXO = "F" == conf(0,9917) --- sup(0,0231) --- lift(1,8827)

Nível ----------------------------------------------
3
-----------------------------------------------------

Resumo ---------------------------------------------
Data e Hora Inicio = 22/12/2003 09:24:32
Data e Hora Final  = 22/12/2003 09:25:43
Total de Tempo     = 00:01:11
Total de Regras    = 1
Total de Níveis    = 3
-----------------------------------------------------
```

Figura 5.10: Regras geradas pelo protótipo.

5.3.5 Utilização por Outras Bases de Dados

O protótipo foi desenvolvido com o intuito de suprir as informações específicas da área de saúde pública, mas pode ser utilizado por qualquer domínio, bastando que este seja um domínio quantitativo.

A base de dados que se desejar aplicar ao protótipo deverá ter as seguintes características obrigatórias:

- Conter um campo de identificação única para cada linha da tabela, chamado de ID.
- Ter passado pelo processo de seleção onde se segmenta os dados para uma mineração mais apurada.
- Ter passado pelo processo de pré-processamento onde a compatibilidade entre os dados será resolvida e também os problemas com domínios, que devem ser transformados em dados discretos.

6 MINERAÇÕES

Neste capítulo são apresentados os passos que antecederam as minerações, bem como os seus resultados, visando os objetivos já relatados na seção 5.1. em um banco de dados real da Autarquia do Serviço Municipal de Saúde Londrina –PR.

6.1 Processo de Trabalho

O primeiro passo foi a realização da distribuição dos valores para cada campo, a fim de verificar possíveis acúmulos de registros com determinadas características.

Com base nos objetivos descritos na seção 5.1., o campo **Procedimento** foi observado com maior ênfase, pois os campos **Caráter da Internação** e **Clínica Médica** são dependentes do primeiro. Não houve uma observação muito rigorosa para os outros campos neste primeiro momento, pois qualquer regra de associação significativa encontrada seria útil.

Após a análise da distribuição, verificou-se a existência de uma alta concentração de registros com os mesmos procedimentos e as mesmas características, o que levaria o protótipo a se concentrar nestes procedimentos. Foi decidido junto ao especialista que os estes registros deveriam ser retirados da mineração.

Assim, com a segmentação da base de dados, retirando-se os registros que continham os procedimentos identificados, a base que originalmente era de 55.734 registros ficou com 41.636 registros. Este processo resultou em uma melhor distribuição dos campos de objetivo.

Posteriormente foi decidido junto com o especialista que as minerações deveriam iniciar apenas com dois campos, sendo um deles um campo objetivo. Esta abordagem se deve ao fato da enorme possibilidade de correlação entre os campos, evitando assim uma explosão combinatória.

Realizadas as minerações com dois campos, os resultados foram examinados e apenas os campos que possuíssem alguma informação útil minerada passariam a etapa seguinte, na qual seriam minerados todos os campos significativos, mais o objetivo.

Os parâmetros utilizados na mineração foram:

- Suporte mínimo igual 0,005 ou 0,5% o que significa aproximadamente 208
- registros;
- Confiança mínima igual 0,7 ou 70%;
- *Lift* igual 1,1, pois resultados abaixo de 1,1 não possuem valor para pesquisa;
- *Improvement* igual 0,009.

6.2 Regras

Apesar de muitas regras serem comentadas com o auxilio do especialista e uma equipe composta de diversos segmentos da Secretaria de Saúde de Londrina, outras regras foram excluídas, pois as correlações encontradas eram esperadas.

6.2.1. Minerações objetivando o campo Procedimento.

1ª Raça X Procedimento

Devido ao grande número de registros de pessoas com raça branca, os resultados provaram o trivial, todos os procedimentos com o suporte mínimo possuíam associação com confiança significativa com a raça branca. Portanto as regras geradas foram descartadas.

2ª Situação Familiar Conjugal X Procedimento

- 72,04% das pessoas nas quais foram realizados o procedimento de laqueadura tubária convive com companheiro e filhos.(suporte 0,53%), (*lift* 1,9702)

Esta regra se mostra importante, pois ela revela a princípio um desvio de aproximadamente 27,96% quanto ao procedimento, especialistas na área presumem que realizem este procedimento somente pessoas que já tenham filhos.

Analisado o conteúdo granular da regra foi identificado que, este desvio se deve em parte a variedade de formas cadastrais mostradas no Anexo I (Tabela **USUARIO**), pois:

1 - 7,57% das pessoas nas quais foram realizadas o procedimento de laqueadura tubária convive com companheiro, com filho e/ou outros familiares;

2 - 3,62% das pessoas nas quais foram realizadas o procedimento de laqueadura tubária convive com companheiro e sem filhos;

3 - 0,32% das pessoas nas quais foram realizadas o procedimento de laqueadura tubária convive com outras pessoas sem laços consangüíneos e/ou laços conjugais;

4 -16,45% das pessoas nas quais foram realizadas o procedimento de laqueadura tubária convive com familiares e sem companheiros.

Com base nesta análise ficam ressaltados os itens 2 e 4, no item 2 fica explícito que as pessoas realmente não tinham filhos e no item 4 a situação é mais complexa pois é a única alternativa para pessoas que não convivem com companheiros e possuam ou não filhos.

Fazendo outra análise baseado no procedimento e no item 2 o atributo que se destacou foi o da faixa etária onde a grande maioria estava entre 35 e 40 anos e os outros casos se aproximavam bem desta faixa etária, justificando em parte a realização do procedimento, pois nesta faixa estaria a gravidez é considerada de risco.

- 70,15% das pessoas nas quais foram realizados o procedimento de cesariana convive com companheiro e filhos. (suporte 0,68%), (*lift* 1,9185)

Esta a princípio pode significar, pela situação conjugal, que a maioria das Cesarianas acontece a partir da segunda gestação.

Seria preciso uma análise de padrões seqüenciais para estudar mais a fundo esta regra, verificando a validade da regra.

- 79,66% das pessoas tratadas em psiquiatria em hospitais convivem com familiares e sem companheiro. (suporte 1,11%), (*lift* 1,687).

Estudando os registros foi verificado que esta situação familiar conjugal indica um crescimento no tratamento de doentes psiquiátricos em hospitais dia, onde o paciente fica durante o dia para tratamento e a noite volta para o convívio dos familiares. Tal crescimento é natural em resposta a triagem que feita para estes pacientes, pois a maioria dele eram internados em sanatórios apesar da sua situação não ser tão grave, esta situação além de gerar custos para Secretaria de Saúde, era prejudicial ao paciente.

Foram reveladas outras regras triviais, relacionando procedimentos pediátricos com a situação familiar dos que convivem com familiares e sem companheiro, situação esta em que as crianças normalmente se encontram.

3ª Escolaridade X Procedimento

Nesta mineração foram reveladas regras triviais, relacionando procedimentos pediátricos com características de pessoas que não sabem ler e escrever, características estas em que as crianças normalmente se encontram.

4ª Faixa Etária X Procedimento

Nesta mineração foram reveladas regras triviais relacionando procedimentos pediátricos com pessoas de faixas etárias de 0 a 4 anos.

5ª Sexo X Procedimento

- 88,85% das safenectomias interna radical são realizadas em pessoas do sexo feminino. (suporte 0,55%), (*lift* 1,6867)

A princípio esta regra foi considerada interessante, pois poderia ser indício de procedimento relacionado a fatores estéticos.

Estudando-se mais profundamente foi verificado que o procedimento era concentrado em mulheres que trabalhavam no lar com mais de 40 anos, levando o especialista a concluir que o procedimento se justifica, pois as dores que levam a realização do procedimento podem ser agravadas pelo tipo de atividade física (ocasionada pelo trabalho no lar) e também pela idade.

- 78,78% das cirurgias múltiplas são realizadas em pessoas do sexo masculino. (suporte 0,59%), (*lift* 1,6646)

Esta regra se mostra interessante, pois a concentração para o sexo masculino não condiz com a distribuição dos sexos no geral da base de dados que na sua maioria e do sexo feminino.

Ao se estudar esta regra foi encontrada além da relação do procedimento com o sexo masculino também com a faixa etária de 0 a 9 anos. Não foram encontradas razões, dentro dos atributos da base de dados, que justificassem a concentração deste procedimento neste sexo e faixa etária. Esta situação foi encaminhada para o setor de avaliação e controle para um melhor estudo.

- 80,45% das herniorrafias inguinais (unilateral) múltiplas são realizadas em pessoas do sexo masculino. (suporte 1,45%), (*lift* 1,7)

Estudando mais profundamente foi verificado que este procedimento estava concentrado sexo masculino também em crianças de 0 a 4 anos, caracterizando um erro de nomenclatura nos procedimentos pois nesta idade um procedimento de urologia muito comum é o tratamento da hidrocele comunicante muito parecido com a herniorrafia inguinal.

Foram tomadas medidas para que tal procedimento fosse registrado de forma correta pelos hospitais, pois a herniorrafia inguinal pode levar a uma internação de urgência ou emergência, aumentado seu custo, já o tratamento da hidrocele comunicante e um procedimento eletivo.

- 72,81% das internações em psiquiatria são realizadas em pessoas do sexo masculino. (suporte 0,96%), (*lift* 2,2567)

Ao se estudar mais profundamente a regra não foi encontrada nenhuma razão dentro dos atributos da base de dados que justificasse o procedimento se concentrar no sexo masculino, o que está levando a um estudo em loco dos casos.

A colecistectomias, colpoperineoplastia, histerectomia total, laqueadura tubária, parto normal, cesariana, curetagem pós-aborto, procedimentos que deveriam compor 100% para sexo feminino, nem sempre correspondem a essa lógica indicando um desvio.

Ao se estudar estes procedimentos mais profundamente analisado o desvio para sexo masculino, foi concluído que os procedimentos tinham sido preenchidos de forma incorreta.

6ª Freqüenta a escola X Procedimento

Esta relação foi descartada, pois a distribuição para o campo **freqüenta a escola** foi extremamente desigual sendo que 86% das pessoas não freqüenta a escola, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente **freqüenta a escola** como não.

7ª CBOR X Procedimento

Todas a regras encontradas fizeram referência aos procedimentos pediátricos relacionados ao CBOR XX4 que faz referência a crianças, portanto regras triviais.

8ª Segmento X Procedimento

Esta relação foi descartada, pois a distribuição para o campo **segmento** foi extremamente desigual, sendo que 95% das pessoas são do **segmento** urbano, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o **segmento** urbano.

9ª Área X Procedimento

Apesar de apresentar áreas com bom suporte não foi possível obter regras significativas para esta relação, devido ao baixo suporte das mesmas.

10ª Micro-área X Procedimento

Apesar de apresentar micro-áreas com bom suporte não foi possível obter regras significativas para esta relação, devido ao baixo suporte das mesmas.

11ª Faixa de Pessoas X Procedimento

Apesar de apresentar Faixa de Pessoas com bom suporte não foi possível obter regras significativas para esta relação, devido à baixa confiança das mesmas.

12ª Faixa de Cômodos X Procedimento

Esta relação foi descartada, pois a distribuição para o campo **faixa de cômodos** foi extremamente desigual sendo que 85% das pessoas possuem 4 cômodos ou mais,

justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente a **faixa de cômodos** de 4 ou mais.

13ª Energia elétrica X Procedimento

Esta relação foi descartada, pois a distribuição para o campo **energia elétrica** foi extremamente desigual sendo que praticamente 100% das pessoas possuem energia elétrica, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente à **energia elétrica** como sim.

14ª Esgoto sanitário X Procedimento

Esta relação foi descartada, pois a distribuição para o campo **esgoto sanitário** foi desigual sendo que 64% das pessoas usam a rede pública de esgoto, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o uso da rede pública de esgoto.

15ª Tipo de Domicilio X Procedimento

Esta relação foi descartada, pois a distribuição para o campo **tipo de domicilio** foi extremamente desigual sendo que 84% das pessoas moram em domicílios de alvenaria, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o domicilio de alvenaria.

16ª Destino do Lixo X Procedimento

Esta relação foi descartada, pois a distribuição para o campo **destino do lixo** foi extremamente desigual sendo que 96% das pessoas têm seu lixo coletado, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o lixo coletado.

17ª Abastecimento de água X Procedimento

Esta relação foi descartada, pois a distribuição para o campo **abastecimento de água** foi extremamente desigual sendo que 97% das pessoas têm abastecimento de água pela rede pública, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente à rede pública com abastecimento de água.

18ª Tratamento de água X Procedimento

Esta relação foi descartada, pois a distribuição para o campo **tratamento de água** foi extremamente desigual sendo que 71% das pessoas não têm tratamento de água, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o não tratamento de água.

19ª Programas de Cobertura X Procedimento

Esta relação foi descartada, pois a distribuição para o campo **programas de cobertura** foi extremamente desigual sendo que praticamente 100% das pessoas possuem programas de cobertura como PSF, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o programa de cobertura PSF.

20ª Hospital X Procedimento

- 76,67% das adenoidectomias são realizadas no Hospital Zona Norte. (suporte 0,77%), (*lift* 5,2808)

Esta regra se destaca devido a concentração deste procedimento nesse hospital. Não foi encontrada nenhuma razão dentro dos atributos da base de dados, sendo que a questão foi encaminhada ao setor de avaliação e controle.

- 78,96% das amigdalectomia com adenoidectomia são realizadas no Hospital Zona Norte. (suporte 0,91%), (*lift* 4,7895)

Esta regra se destaca devido a concentração deste procedimento nesse hospital. Não foi encontrada nenhuma razão dentro dos atributos da base de dados, sendo que a questão foi encaminhada ao setor de avaliação e controle.

21ª CPF do Médico X Procedimento

Todas as regras descobertas revelaram a relação trivial entre o médico e os procedimentos de sua especialidade.

22ª Regras de tamanho 3 para Procedimento

- 90,76% das pessoas que convivem com companheiro e filhos e realizam o procedimento de herniorrafia inguinal (unilateral) são do sexo masculino. (suporte 0,52%), (*lift* 1,9177)

Esta regra é interessante, pois ao contrário da regra encontrada na 5ª mineração para o procedimento, esta representa o que era esperado para a herniorrafia inguinal, indicando que pessoas já com filhos, ou seja, com idade normalmente acima de 18 anos realizam este procedimento. Este indicativo foi comprovado fazendo-se uma análise mais granular na base de dados onde além de provar que as faixas etárias estavam entre 45 e 65 anos foi encontrada uma relação com a profissão onde a maioria das pessoas exercia profissões que exigem muito esforço físico normalmente executado por homens, como pedreiros e estucadores.

- 74,68% das pessoas com 1º grau incompleto e que se submeteram a cesariana com atendimento ao recém nascido sala de parto são atendidas na Maternidade Municipal. (suporte 0,85%), (*lift* 21,6541)

Esta regra revela a característica socioeconômica dos pacientes atendidos na Maternidade Municipal.

Algumas minerações trazem resultados óbvios relacionando procedimentos pediátricos a faixas etárias e escolaridades de crianças, portanto foram descartadas.

23ª Regras de tamanho 4 para Procedimento

- 76,45% das pessoas convivem com companheiro e filhos e possuem o 1º grau incompleto e se submeteram a cesariana com atendimento ao recém nascido na sala de parto são internadas na Maternidade Municipal. (suporte 0,57%), (*lift* 22,1667)

Esta regra revela, além de uma característica socioeconômica relacionada aos pacientes normalmente internados na Maternidade Municipal, característica de pessoas que já tiveram um filho.

Algumas minerações foram descartadas devido às associações obvias, como relacionar procedimentos obstétricos com o sexo feminino e relacionar procedimentos pediátricos a faixas etárias e escolaridades de crianças.

6.2.2. Minerações objetivando o campo Caráter da Internação.

1ª Raça X Caráter da Internação

Devido ao grande número de registros de pessoas da raça branca, os resultados provaram o trivial, todo o Caráter da Internação com o suporte mínimo possuía associação com confiança significativa com a raça branca. Portanto as regras geradas forma descartadas.

2ª Situação Familiar Conjugal X Caráter da Internação

- 75,06% das pessoas que convivem com outras pessoas sem laços consangüíneos e ou laços conjugais são internadas como urgência ou emergência. (suporte 0,79%), (*lift* 1,2213)

Esta regra é interessante, pois traz um perfil de atendimento de urgência ou emergência.

Estudando a regra mais profundamente verificou-se que as pessoas atendidas com as características trazidas na regra possuíam idades bem avançadas, entre 65 e 80 ou mais, o que justifica a princípio o atendimento em urgência ou emergência e o fato do convívio com outras pessoas sem laços consangüíneos e ou laços conjugais, pois na maioria estas pessoas residem em asilos, o que foi confirmado verificando sua área e micro-área. Foi realizada mais uma pesquisa granular na base de dados a fim de verificar qual o procedimento mais comum seguindo as características da regra, encontrando-se então que o procedimento mais comum era de insuficiência cardíaca reforçado então a justificativa de caráter de internação com urgência ou emergência.

3ª Escolaridade X Caráter da Internação

Apesar de apresentar escolaridade com bom suporte não foi possível obter regras significativas para esta relação, devido à baixa confiança das regras.

4ª Faixa Etária X Caráter da Internação

- 70,58% das pessoas entre 15 e 19 anos são internadas como urgência ou emergência. (suporte 3,99%), (*lift* 1,1484)

Esta regra se mostra interessante, pois coloca uma faixa etária relacionada com o caráter de internação de urgência ou emergência, mas estudando-se mais profundamente verificou-se que os procedimentos relacionados eram na sua maioria voltados para parto o que justifica a idade, onde se encontram as mulheres em idade fértil.

- 70,37% das pessoas entre 20 e 24 anos são internadas como urgência ou emergência. (suporte 4,79%), (*lift* 1,1451)

Como a regra anterior esta se mostrou interessante, pois coloca uma faixa etária relacionada com o caráter de internação de urgência ou emergência, mas estudando-se mais profundamente verificou-se que os procedimentos relacionados eram na sua maioria voltados para parto o que justifica a idade, onde se encontram as mulheres em idade fértil.

- 70,32% das pessoas entre 75 e 79 anos são internadas como urgência ou emergência. (suporte 2,05%), (*lift* 1,1442)

Esta regra se mostra interessante, pois coloca uma faixa etária relacionada com o caráter de internação de urgência ou emergência, mas estudando-se mais profundamente verificou-se que os procedimentos relacionados eram na sua maioria insuficiência cardíaca ou doenças pulmonares o que justifica a idade, onde se encontram as pessoas com maiores riscos de desenvolver estes tipos de doença.

- 75,78% das pessoas com idade maior ou igual a 80 são internadas como urgência ou emergência. (suporte 1,98%), (*lift* 1,233)

Como a regra anterior esta se mostrou interessante, pois coloca uma faixa etária relacionada com o caráter de internação de urgência ou emergência, mas estudando-se mais profundamente verificou-se que os procedimentos relacionados eram na sua maioria insuficiência cardíaca ou doenças pulmonares o que justifica a idade, onde se encontram as pessoas com maiores riscos de desenvolver estes tipos de doença.

5ª Sexo X Caráter da Internação

Apesar de apresentar o campo sexo com bom suporte não foi possível obter regras significativas para esta relação, devido à baixa confiança das regras.

6ª Freqüenta a escola X Caráter da Internação

Esta relação foi descartada, pois a distribuição para o campo freqüenta a escola foi extremamente desigual sendo que 86% das pessoas não freqüentam a escola, justifica os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente freqüenta a escola como não.

7ª CBOR X Caráter da Internação

- 75,06% dos trabalhadores agropecuários polivalentes ou assemelhados são internados como urgência ou emergência. (suporte 0,74%), (*lift* 1,2213)

Esta regra a princípio que a atividade descrita possui uma relação direta com o caráter da internação, mas estudando mais profundamente foi verificado que os principais procedimentos atrelados ao caráter da internação eram insuficiência cardíaca e broncopneumonia o que levantou suspeita sobre a real relação com as atividades. Verificou-se estão a faixa etária do grupo em questão e encontrou-se que a população estudada tinha entre 65 e 80 anos ou mais, justificando sua internação no caráter urgência ou emergência. A questão agora era verificar a razão pela qual a atividade (CBOR) era evidenciada, e em discussão com o grupo de ação social foi colocada a hipótese que estes trabalhadores ao se aposentarem migram para as cidades e como possuíam pouco cuidado preventivo antes de migrarem acabam por adquirir as patologias acima descritas. A confirmação de tal hipótese ficou sobre a responsabilidade da ação social.

Todas as regras contendo CBOR como A99 ou 999 (trabalhadores não classificados) foram desconsideradas por sua falta de significado.

8ª Segmento X Caráter da Internação

Esta relação foi descartada, pois a distribuição para o campo segmento foi extremamente desigual sendo que 95% das pessoas são do segmento urbano, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o segmento urbano.

9ª Área X Caráter da Internação

- 72,37% das pessoas residentes na área de abrangência da unidades básica de saúde Panissa, são internados como urgência ou emergência. (suporte 0,58%), (*lift* 1,1776)

Esta regra é interessante, pois revela a principio uma característica da área de abrangência da unidade básica de saúde.

E analisando mais profundamente verificamos que os procedimentos referentes a parto e atendimentos pediátricos são o que compõem a maioria destes atendimentos de urgência ou emergência. Tal situação foi encaminhada a gerência das unidades básicas de saúde para providências relacionadas a prevenção e controle dos procedimentos descobertos.

- 70,28% das pessoas residentes nas áreas de abrangência das unidades básicas de saúde Itapua, são internados como urgência ou emergência. (suporte 1,14%), (*lift* 1,1436)

Esta regra é interessante, pois revela a principio uma característica da área de abrangência da unidade básica de saúde.

E analisando mais profundamente verificamos que os procedimentos referentes a parto e atendimentos pediátricos são os que compõem a maioria destes atendimentos de urgência ou emergência. Tal situação foi encaminhada a gerência das unidades básicas de saúde para providências relacionadas a prevenção e controle dos procedimentos descobertos.

- 70,91% das pessoas residentes nas áreas de abrangência das unidades básicas de saúde União da Vitória, são internados como urgência ou emergência. (suporte 0,9%), (*lift* 1,1538)

Esta regra é interessante, pois revela a principio uma característica da área de abrangência da unidade básica de saúde.

E analisando mais profundamente verificamos que os procedimentos referentes a parto e atendimentos pediátricos são os que compõem a maioria destes atendimentos de urgência ou emergência. Tal situação foi encaminhada a gerência das unidades básicas de saúde para providências relacionadas a prevenção e controle dos procedimentos descobertos.

As três unidades básicas de saúde anteriores Panissa, Itapoã e União da Vitória estão em áreas de baixa renda o que pode implicar na demanda por atendimento de urgência e emergência. Esta situação também foi passada para a ação social.

- 70,24% das pessoas residentes nas áreas de abrangência das unidades básicas de saúde Guanabara, são internados como urgência ou emergência. (suporte 0,71%), (*lift* 1,1428)

Esta regra é interessante, pois revela a principio uma característica da área de abrangência da unidade básica de saúde.

E analisando mais profundamente verificamos que os procedimentos cardiovasculares e pulmonares são os que compõem a maioria destes atendimentos de urgência ou emergência, também foi verificada a faixa etária entre 65 e 80 anos na população em questão. Tal situação é justificada pela presença um grande asilo na região.

10ª Micro-área X Caráter da Internação

Apesar de apresentar micro-áreas com bom suporte não foi possível obter regras significativas para esta relação, devido ao baixo suporte das regras.

11ª Faixa de Pessoas X Caráter da Internação

Apesar de apresentar Faixa de Pessoas com bom suporte não foi possível obter regras significativas para esta relação, devido à baixa confiança das regras.

12ª Faixa de Cômodos X Caráter da Internação

- 76,07% das pessoas que residem em 1 cômodo são internadas como urgência ou emergência. (suporte 0,94%), (*lift* 1,2377)

Esta regra é interessante por revelar que pessoas de nível social mais baixo são internadas como urgência ou emergência com mais freqüência. Posteriormente foi confirmado que as suspeitas em relação ao nível social estavam corretas, pois a grande maioria é residente em áreas de população mais simples, também foi verificado que os procedimentos mais comuns com as características da regra eram procedimentos de parto e pediátricos. Esta situação também foi passada para a ação social.

- 70,01% das pessoas que residem em 2 cômodos são internadas como urgência ou emergência. (suporte 3,13%), (*lift* 1,1391)

Esta regra é interessante por revelar que pessoas de nível social mais baixo são internadas como urgência ou emergência com mais freqüência. Posteriormente foi confirmado que as suspeitas em relação ao nível social estavam corretas, pois a grande maioria é residente em áreas de população mais simples, também foi verificado que os procedimentos mais comuns com as características da regra eram procedimentos de parto e pediátricos. Esta situação também foi passada para a ação social.

Outra relação, tendo como conseqüente, residem em 4 cômodos ou mais foi descartada, pois a distribuição para o campo faixa de cômodos foi extremamente desigual, sendo que 85% das pessoas possuem 4 cômodos ou mais.

13ª Energia elétrica X Caráter da Internação

Esta relação foi descartada, pois a distribuição para o campo energia elétrica foi extremamente desigual, sendo que praticamente 100% das pessoas possuem energia elétrica, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente a energia elétrica como sim.

14ª Esgoto sanitário X Caráter da Internação

Apesar de apresentar Faixa de Pessoas com bom suporte não foi possível obter regras significativas para esta relação, devido à baixa confiança das regras.

15ª Tipo de Domicilio X Caráter da Internação

- 73,29% das pessoas que moram em domicílios de madeira são internadas como urgência e emergência. (suporte 0,8%), (*lift* 1,1925)

Esta regra é interessante por revelar que pessoas de nível social mais baixo são internadas como urgência ou emergência com mais freqüência, confirmado a 12ª mineração para o Caráter da Internação. Posteriormente foi confirmado que as suspeitas em relação ao nível social estavam corretas, pois a grande maioria é residente em áreas de população mais simples, também foi verificado que os procedimentos mais comuns com as características da regra eram procedimentos de parto e pediátricos. Esta situação também foi passada para a ação social.

A outra regra foi descartada, pois a distribuição para o campo tipo de domicílio foi extremamente desigual sendo que 84% das pessoas moram em domicílios de alvenaria, justificando o resultado da mineração, uma vez que a regra encontrada coloca como conseqüente o domicilio de alvenaria.

16ª Destino do Lixo X Caráter da Internação

- 70,95% das pessoas que joga o seu lixo a céu aberto são internadas como urgência e emergência. (suporte 0,72%), (*lift* 1,1545)

Esta regra é interessante por revelar que pessoas de nível social mais baixo são internadas como urgência ou emergência com mais freqüência, confirmado as 12ª e 15ª minerações para o Caráter da Internação. Posteriormente foi confirmado que as suspeitas em relação ao nível social estavam corretas, pois a grande maioria é residente

em áreas de população mais simples, também foi verificado que os procedimentos mais comuns com as características da regra eram procedimentos de parto e pediátricos. Esta situação também foi passada para a ação social.

A outra regra foi descartada, pois a distribuição para o campo destino do lixo foi extremamente desigual, sendo que 96% das pessoas têm seu lixo coletado, justificando o resultado da mineração, uma vez que a regra encontrada coloca como conseqüente o lixo coletado.

17ª Abastecimento de água X Caráter da Internação

Esta relação foi descartada, pois a distribuição para o campo abastecimento de água foi extremamente desigual sendo que 97% das pessoas têm abastecimento de água pela rede pública, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente a rede pública com abastecimento de água.

18ª Tratamento de água X Caráter da Internação

Esta relação foi descartada, pois a distribuição para o campo tratamento de água foi extremamente desigual sendo que 71% das pessoas não têm tratamento de água, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o não tratamento de água.

19ª Programas de Cobertura X Caráter da Internação

Esta relação foi descartada, pois a distribuição para o campo programas de cobertura foi extremamente desigual, sendo que praticamente 100% das pessoas possuem programas de cobertura como PSF, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o programa de cobertura PSF.

20ª Hospital X Caráter da Internação

- 77,94% das internações SUS do Hospital Evangélico são urgência ou emergência. (suporte 6,43%), (*lift* 1,2682)

Esta regra levanta suspeitas sobre a cobrança irregular, pois o atendimento como urgência ou emergência tem um valor maior que o eletivo. Mas fazendo a pesquisa granular na base de dados foi verificado que os procedimentos realizados neste hospital se justificam sendo que na sua maioria são insuficiência coronária aguda, apendicite, intercorrencia para renal crônico, hemorragias digestivas e insuficiência cardíaca.

- 80,15% das internações SUS na Santa Casa são de urgência ou emergência. (suporte 14,73%), (*lift* 1,3042)

Esta regra levanta suspeitas sobre a cobrança irregular, pois o atendimento como urgência ou emergência tem um valor maior que o eletivo. Mas fazendo a pesquisa granular na base de dados foi verificado que os procedimentos realizados neste hospital se justificam sendo que na sua maioria são insuficiência coronária aguda, acidente vascular cerebral e insuficiência cardíaca.

- 99,66% das internações SUS na Clínica Psiquiátrica são de urgência ou emergência. (suporte 1,4%), (*lift* 1,6215)
- 98,97% das internações SUS no Hospital Maxwell são eletivas. (suporte 1,38%), (*lift* 3,2362)

Estas duas regras devem ser analisadas conjuntamente, pois se referem a tratamentos psiquiátricos, a primeira regra coloca os atendimentos realizados na Clínica Psiquiátrica que são de urgência ou emergência, atendimentos estes justificados por ser uma clínica que funciona 24 horas atendendo e internando pacientes em estados mais delicados durante períodos maiores que um dia. Já o Hospital Maxwell é um hospital dia onde o tratamento dos pacientes é feito durante o dia e este retorna ao convívio dos familiares a noite e finais de semana.

21ª CPF do Médico X Caráter da Internação

Todas as regras descobertas revelaram a relação trivial entre o médico e o caráter da internação de sua especialidade médica.

Todas as regras contendo faixa de cômodos de 4 ou mais cômodos, tipo de domicilio de alvenaria e têm seu lixo coletado foram descartadas devido a sua distribuição desigual na base de dados.

Todas as regras contendo CBOR como A99 (trabalhadores não classificados) foram desconsideradas por sua falta de significado.

22ª Regras de tamanho 3 para Caráter da Internação

- 85,9% das pessoas com faixa etária de 15 a 19 anos e que convivem com companheiro e filhos são internados com urgência ou emergência. (suporte 1,27%), (*lift* 1,3977)
- 79,56% das pessoas com faixa etária de 20 a 24 anos e que convivem com companheiro e filhos são internados com urgência ou emergência. (suporte 2,17%), (*lift* 1,2945)

As duas regras acima devem ser analisa conjuntamente, pois colocam uma faixa etária relacionada com a situação conjugal e o caráter de internação de urgência ou emergência, e quando estudadas mais profundamente verifica-se que os procedimentos relacionados eram na sua maioria voltados para parto o que justifica a idade, onde se encontram as mulheres em idade fértil e a situação conjugal.

- 81,93% das pessoas que convivem com companheiro, com laços conjugais e sem filhos e são internadas no Hospital Evangélico, são internadas em urgência ou emergência. (suporte 0,65%), (*lift* 1,3331)

Esta regra a principio levo a acreditar de que se trata-se de pessoas recém casadas e sem filhos que eram internados no hospital em questão com caráter de urgência ou emergência, mas se fazendo um estudo mais aprofundados verificou-se de que se tratavam de pessoas idosas que moravam com o companheiro e que realizavam procedimentos considerados comuns para sua faixa etária como procedimentos cardíacos, renais e pulmonares.

- 83,33% das pessoas que convivem com companheiro com laços conjugais com filhos e/ou outros familiares e são internadas na Santa Casa, são internadas como urgência ou emergência. (suporte 0,84%), (*lift* 1,3559)

Esta regra leva a acreditar que os procedimentos realizados em urgência e emergência eram provenientes de partos devido a situação conjugal e o hospital em questão, o que foi confirmado posteriormente por uma pesquisa mais granular.

- 70,33% das pessoas de 70 a 74 anos que convivem com familiares e sem companheiros são internados em urgência ou emergência. (suporte 0,86%), (*lift* 1,1444)
- 76% das pessoas de 75 a 79 anos que convivem com familiares e sem companheiros são internados em urgência ou emergência. (suporte 0,86%), (*lift* 1,249)
- 77,88% das pessoas com 80 anos ou mais que convivem com familiares e sem companheiros são internados em urgência ou emergência. (suporte 0,99%), (*lift* 1,2672)

As três regras anteriores convertem-se a mesma questão, que são: Quais foram os procedimentos realizados?

Fazendo-se um estudo granular verificou-se que os procedimentos mais comuns eram cardíacos e ou vasculares, renais ou pulmonares.

- 70,43% das pessoas que residem na área de abrangência da unidade de saúde Marabá e convivem com familiares e sem companheiros são internados em urgência ou emergência. (suporte 0,82%), (*lift* 1,146)

Ao se fazer uma analise mais granular verificou-se que para esta regra existe uma concentração de procedimentos pediátricos e faixa etária esta entre 0 a 9 anos,

justificando a situação conjugal da regra, quanto a área de abrangência e o caráter de internação esta situação já foi relatada na 9ª mineração para o caráter de internação.

- 71,04% das pessoas que residem na área de abrangência da unidade de saúde Novo Amparo e convivem com familiares e sem companheiros são internados em urgência ou emergência. (suporte 0,51%), (*lift* 1,156)

Ao se fazer uma analise mais granular verificou-se que para esta regra existe uma concentração de procedimentos pediátricos e faixa etária esta entre 0 a 9 anos, justificando a situação conjugal da regra, quanto a área de abrangência e o caráter de internação esta situação já foi relatada na 9ª mineração para o caráter de internação.

- 73,97% das pessoas que residem na área de abrangência da unidade de saúde Itapoa e convivem com familiares e sem companheiros são internados em urgência ou emergência. (suporte 0,69%), (*lift* 1,2036)

Ao se fazer uma analise mais granular verificou-se que para esta regra existe uma concentração de procedimentos pediátricos e faixa etária esta entre 0 a 9 anos, justificando a situação conjugal da regra, quanto a área de abrangência e o caráter de internação esta situação já foi relatada na 9ª mineração para o caráter de internação.

- 71,12% das pessoas de 0 a 4 anos com 3 cômodos na casa são internadas em urgência ou emergência. (suporte 1,51%), (*lift* 1,1572)

Esta regra pode ser agregada as três anteriores, analisando os registros que compõem a regra foi verificado que eram relativos as na sua maioria da área de abrangência das regras anteriores, deixando claro que estas áreas são regiões de baixa renda onde a incidência de procedimentos pediátricos de emergência é alta. Tal situação foi encaminhada a gerência das unidades básicas de saúde para providências relacionadas a prevenção e controle dos procedimentos descobertos.

- 72,29% das pessoas de 25 a 29 anos que moram em casa de taipa não revestida são internadas com urgência ou emergência. (suporte 0,69%), (*lift* 1,1763)

Fazendo uma analise nos registros que compõem esta regra verificou-se que se tratavam de pessoas que residiam na zona rural, e os procedimentos mais comuns são partos e anemia profunda. Este caso foi levado ao conhecimento da equipe de médicos da família que atende a zona rural para que estes tomem as devidas providências.

Toda regra que relaciona a situação familiar, como convive com familiares e sem companheiros e o Hospital Infantil foi descartada por ser trivial.

Toda regra que relaciona a situação familiar, como convive com familiares e sem companheiros e faixa etária de 0 a 19 foi descartada por ser trivial.

Toda regra que relaciona a situação familiar, como convive com familiares e sem companheiros e CBOR como estudante ou criança foi descartada por ser trivial.

Toda regra que relaciona a faixa etária de 0 a 14 e CBOR como estudante ou criança foi descartada por ser trivial.

6.2.3. Minerações objetivando o campo Clínica Médica.

1ª Raça X Clínica Médica

Devido ao grande número de registros de pessoas de raça branca, os resultados provaram o trivial, todo o Caráter da Internação com o suporte mínimo possui associação com confiança significativa com a raça branca. Portanto as regras geradas foram descartadas.

2ª Situação Familiar Conjugal X Clínica Médica

As regras descobertas nesta mineração se mostraram triviais para o especialista.

3ª Escolaridade X Clínica Médica

As regras descobertas nesta mineração se mostraram triviais para o especialista.

4ª Faixa Etária X Clínica Médica

As regras descobertas nesta mineração se mostraram triviais para o especialista.

5ª Sexo X Clínica Médica

As regras descobertas nesta mineração se mostraram triviais para o especialista.

6ª Freqüenta a escola X Clínica Médica

Esta relação foi descartada, pois a distribuição para o campo **freqüenta a escola** foi extremamente desigual sendo que 86% das pessoas não freqüenta a escola, justifica os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente **freqüenta a escola** como não.

7ª CBOR X Clínica Médica

As regras descobertas nesta mineração se mostraram triviais para o especialista.

8ª Segmento X Clínica Médica

Esta relação foi descartada, pois a distribuição para o campo **segmento** foi extremamente desigual sendo que 95% das pessoas são do segmento urbano, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o segmento urbano.

9ª Área X Clínica Médica

Apesar de apresentar áreas com bom suporte não foi possível obter regras significativas para esta relação, devido à baixa confiança das regras.

10ª Micro-área X Clínica Médica

Apesar de apresentar micro-áreas com bom suporte não foi possível obter regras significativas para esta relação, devido ao baixo suporte das regras.

11ª Faixa de Pessoas X Clínica Médica

Apesar de apresentar Faixa de Pessoas com bom suporte não foi possível obter regras significativas para esta relação, devido à baixa confiança das regras.

12ª Faixa de Cômodos X Clínica Médica

Esta relação foi descartada, pois a distribuição para o campo **faixa de cômodos** foi extremamente desigual, sendo que 85% das pessoas possuem 4 cômodos ou mais, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente a **faixa de cômodos** de 4 ou mais.

13ª Energia elétrica X Clínica Médica

Esta relação foi descartada, pois a distribuição para o campo **energia elétrica** foi extremamente desigual, sendo que praticamente 100% das pessoas possuem energia elétrica, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente a **energia elétrica** como sim.

14ª Esgoto sanitário X Clínica Médica

Esta relação foi descartada, pois a distribuição para o campo **esgoto sanitário** foi extremamente desigual sendo, que 64% das pessoas usam a rede pública de esgoto, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o uso da rede pública de esgoto.

15ª Tipo de Domicilio X Clínica Médica

Esta relação foi descartada, pois a distribuição para o campo **tipo de domicilio** foi extremamente desigual sendo, que 84% das pessoas moram em domicílios de alvenaria, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o domicilio de alvenaria.

16ª Destino do Lixo X Clínica Médica

Esta relação foi descartada, pois a distribuição para o campo **destino do lixo** foi extremamente desigual sendo, que 96% das pessoas têm seu lixo coletado, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o lixo coletado.

17ª Abastecimento de água X Clínica Médica

Esta relação foi descartada, pois a distribuição para o campo **abastecimento de água** foi extremamente desigual sendo, que 97% das pessoas têm abastecimento de água pela rede pública, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente à rede pública com abastecimento de água.

18ª Tratamento de água X Clínica Médica

Esta relação foi descartada, pois a distribuição para o campo **tratamento de água** foi extremamente desigual sendo, que 71% das pessoas não têm tratamento de água, justificando os resultados da mineração, uma vez que todas as regras encontradas colocam como conseqüente o não tratamento de água.

19ª Programas de Cobertura X Clínica Médica

Esta relação foi descartada, pois a distribuição para o campo **programas de cobertura** foi extremamente desigual sendo, que praticamente 100% das pessoas possuem programas de cobertura como PSF, justificando os resultados da mineração,

uma vez que todas as regras encontradas colocam como conseqüente o programa de cobertura PSF.

20ª Hospital X Clínica Médica

As regras descobertas nesta mineração se mostraram triviais para o especialista.

21ª CPF do Médico X Clínica Médica

Todas as regras descobertas revelaram a relação trivial entre o médico e o clínica médica de sua especialidade.

Algumas minerações trazem resultados óbvios relacionando clínica médica pediátrica a faixas etárias e escolaridades de crianças, portanto foram descartadas.

Toda regra que relaciona a situação familiar como convive com familiares e sem companheiros e faixa etária de 0 a 19 foi descartada por ser trivial.

7 CONCLUSÕES E TRABALHOS FUTUROS

Data Mining é uma área com crescimento rápido em que, novos sistemas, algoritmos ou protótipos são desenvolvidos. Muitos pesquisadores e desenvolvedores têm contribuído muito para o estado da arte. Alguns motivos vêm contribuindo para o crescimento desta área, como a facilidade de armazenamento dos dados, o aumento da capacidade computacional a baixo custo, e também a introdução de novos métodos de aprendizado de máquina para representação do conhecimento baseado em programação lógica, além da tradicional análise estatística de dados. Tais métodos tendem a ser usados intensivamente, conforme a demanda de processamento.

Contudo, apesar de estar crescendo, o campo de Mineração de Dados ainda tem um longo caminho a seguir. Existem vários desafios a superar, mas alguns sucessos têm sido alcançados. Devido ao potencial das aplicações de Mineração de dados estarem crescendo, existe uma pressa para oferecer produtos e serviços para o mercado.

Os dados no seu formato bruto possuem beneficio para os seus sistemas de origem, mas seu grande valor está na possibilidade de se extrair informações úteis para suporte de decisão ou exploração e compreensão do fenômeno de gerenciamento de fonte de dados.

Os Sistemas de Administração de Banco de Dados dão acesso aos dados armazenados, mas essa é apenas uma parte do que se pode obter desses dados. Os sistemas tradicionais de processamento de transações on-line, OLTPs, são boas ferramentas no que se refere à reposição de dados de forma rápida, segura e eficaz em bancos de dados, mas não o são para retornar uma análise significativa relativa aos dados.

Dados analisados podem prover conhecimento adicional sobre um negócio, indo explicitamente além dos armazenados, para derivar conhecimento. Isto posto, é possível afirmar que o *Data mining*, ou *Knowledge Discovery in Databases* (KDD), tem benefícios óbvios para qualquer empreendimento.

Com base neste raciocínio e devido ao grande número e variedades de bases de dados, uma grande perspectiva existe para a área de saúde pública na questão da obtenção de conhecimento útil para ser utilizado em tratamentos curativos e preventivos.

Com esta motivação foi iniciado este projeto, obtendo uma base de dados complexa, composta de dados socioeconômicos, pessoais e de internação, possibilitando uma grande variedade de associações e implementando um protótipo de uma ferramenta de mineração de dados baseada no algoritmo *Apriori*. Os méritos deste trabalho são:

- Obtenção da base de dados proposta,
- Mineração e obtenção de regras baseadas em 3 atributos, procedimentos, caráter da internação e clínica médica, com o objetivo principal de encontrar associações que fugissem do padrão esperado pelo especialista e analisá-las em profundidade a nível granular verificando, possíveis desvios de entendimento,
- Avaliação das regras obtidas juntamente com o especialista, verificando-se a nível granular as informações da regra,
- Implementação de um protótipo baseado no algoritmo *Apriori*, utilizando os conceitos de *Lift* e *Improvement* para melhorar a compreensão das regras.

A obtenção de uma base consistente que refletisse as características do município, gerou uma série de dificuldades como:

- Unificar as fontes de dados. Esta tarefa teve como requisição um identificador único, foram realizados estudos baseados em dados de documentos pessoais, mas a falta de obrigatoriedade e a inconsistência destes dados causaram o impedimento do seu uso. Foi decidido então que a unificação se daria pela união dos primeiros e últimos nomes mais a data de nascimento;
- Dados incompletos, inconsistentes e ou não confiáveis deveriam ser retirados, para realizar tal tarefa foram necessárias varias interações com a base de dados, devido a grande gama de variações das inconsistências e dados incompletos.

Outras dificuldades foram encontradas ao realizar a mineração:

- Ao se realizar as primeiras minerações foi observado que as regras se concentravam em certos procedimentos e por conseqüência o caráter da internação e clínica médica. Então foi decidido junto ao especialista que os procedimentos que concentravam os registros fossem retirados da base, para que ela se tornasse mais homogênea;
- Ao se realizar as minerações, após a retirada dos registros baseada na distribuição do campo procedimento foi observado que as regras geradas se concentravam em alguns outros valores de campos, foi realizada então a distribuição de registros para os demais campos;
- Devido a má distribuição de registros para alguns campos, foi decidido junto ao especialista a adoção da estratégia de se fazer as primeiras minerações em pares e em seguida restringir as minerações somente para os campos que tiveram resultados significativos no passo anterior; em paralelo foi pesquisado o conceito de *Lift* o qual ajuda a diminuir as regras baseadas em campos com má distribuição, tal conceito foi implementado na ferramenta desenvolvida;

- Após as minerações em pares foram realizadas as minerações com os campos mais significativos, mas muitas regras geradas eram na verdade especializações de regras mais simples porém menos confiáveis, foi pesquisado e implementado na ferramenta o conceito de *Improvement*, para evitar este tipo de problema.

Superadas as dificuldades e os trabalhos de mineração realizados, os resultados surtiram efeitos em diversos setores da Secretaria Municipal de Saúde de Londrina, dentre os resultados podemos destacar:

- 88,85% das safenectomias interna radical são realizadas em pessoas do sexo feminino. A princípio esta regra foi considerada interessante, pois poderia ser indício de procedimento relacionado a fatores estéticos. Estudando-se mais profundamente foi verificado que o procedimento era concentrado em mulheres que trabalhavam no lar com mais de 35 anos, levando o especialista a concluir que o procedimento se justifica, pois as dores que levam a realização do procedimento podem ser agravadas pelo tipo de atividade física (ocasionada pelo trabalho no lar) e também pela idade.
- 78,78% das cirurgias múltiplas são realizadas em pessoas do sexo masculino. Esta regra se mostra interessante, pois a concentração para o sexo masculino não condiz com a realidade da base de dados que na sua maioria e do sexo feminino. Ao se estudar esta regra foi encontrada além da relação do procedimento com o sexo masculino também com a faixa etária de 0 a 9 anos. Não foram encontradas razões, dentro dos atributos da base de dados, que justificassem a concentração deste procedimento neste sexo e faixa etária. Esta situação foi encaminhada para o setor de avaliação e controle para um melhor estudo.
- 80,45% das herniorrafias inguinais (unilateral) múltiplas são realizadas em pessoas do sexo masculino. Estudando mais profundamente foi verificado que este procedimento estava concentrado sexo masculino também em crianças de 0 a 4 anos, caracterizando um erro de nomenclatura nos procedimentos pois nesta idade um procedimento de urologia muito comum é o tratamento da hidrocele comunicante muito parecido com a herniorrafia inguinal. Foram tomadas medidas para que tal procedimento fosse registrado de forma correta pelos hospitais, pois a herniorrafia inguinal pode levar a uma internação de urgência ou emergência, aumentado seu custo, já o tratamento da hidrocele comunicante e um procedimento eletivo.
- Foi confirmado que os procedimentos mais comuns paras as faixas etárias 65 a 80 anos são cardíacos ou vasculares, renais e pulmonares internados como urgência ou emergência.
- Verificou-se que em ares menos favorecidas a incidência de procedimentos de parto e pediátricos de urgência ou emergência é alta.
- Foi confirmado que o caráter de internação que consome a maioria dos leitos são de urgência ou emergência.

- Para o campo Clínica Médica foram encontradas apenas regras triviais que não levam a conclusão nenhuma.

As regras geradas relativas à base de dados obtidas não se restringirão apenas a este trabalho, levaram e continuarão levando conhecimento útil à Autarquia do Serviço Municipal de Saúde de Londrina mesmo após o término deste projeto.

Como trabalho futuro, pretende-se aplicar os conceitos de unificação da base e mineração de regras de associação a outras bases de dados de saúde pública da Secretaria Municipal de Saúde de Londrina, unindo a base do cadastro único como as bases de mortalidade e nascidos vivos, observando que os problemas de identificador único e a falta de distribuição de registro para certos campos serão minimizadas devido ao conhecimento obtido com este trabalho.

Pretende-se também implementar a ferramenta de mineração mecanismos para facilitar e agilizar a pesquisa mais aprofundada nas regras obtidas.

REFERÊNCIAS

[ADR 96] ADRIAANS, P.; ZANTINGE, D. **Data Mining**. Harlow: Addison-Wesley, 1996.

[AGR 93] AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining Associations between Sets of Items in Massive Databases. In: ACM SIGMOD INT. CONFERENCE ON MANAGEMENT OF DATA, 1993. **Proceedings**... Washington D.C.: ACM Press, 1993. p. 207-216.

[AGR 94] AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules. In: ACM VLDB INT. CONFERENCE ON VERY LARGE DATABASES, 20., 1994. **Proceedings**... Hove: Morgan Kaufmann, 1994. p.487-499.

[AGR 96] AGRAWAL, R.; SHAFER, J. C. Parallel Mining of Association Rules. **IEEE Transactions on Knowledge and Data Engineering**. New York, 1996.

[BAY 99] BAYARDO JR, R. J.; AGRAWAL, R.; GUNOPULOS, D. Constraint-Based Rule Mining in Large, Dense Databases. In: INT. CONFERENCE ON DATA ENGINEERING, 15., 1999, Sidney. **Proceedings**... [S.l.:s.n.], 1999. p. 188-197.

[BER 97] BERRY, M. J.A.; LINOFF, G. **Data Mining Techniques.** New York : John Wiley, 1997.

[BRA 96] BRACHMAN, R. J.; ANAND, T. The Process of Knowledge Discovery in databases. In: **Advances in Knowledge Discovery and Data Mining**. Menlo Park: AAAI Press, 1996. p.37-57.

[CAM 2002] CAMARGO, S. S. **Mineração de regras de associação no problema da cesta de compras aplicada ao comércio varejista de confecção**. 2002. 105 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

[CAR 2001] CARVALHO, L. A.V. **Data Mining A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração**. São Paulo: Érica, 2001.

[CEN 97] CENGIZ, I.. Mining Association Rules, Bilkent University, **Department of Computer Engineering and Information Sciences**, Ankara, v. 7, n. 3, p.302-353. Dec. 1997, Disponível em: <http://www.cs.bilkent.edu.tr/~icegiz/datamone/mining.html.> Acesso em: jan. 2003.

[CHE 96] CHEN, M.; HAN, J.; YU, P. S. Data Mining: An Overview from a Database Perspective. **IEEE Transactions On Knowledge and Data Engineering**: New York, v. 8, n. 6, p.866-883, Dec. 1996. Disponível em: <http://www.informatik.uni-trier.de/~ley/db/journals/tkde/tkde8.html>. Acesso em jan. 2003.

[FAY 96] FAYYAD, U. M. et al. From data mining to knowledge discovery: an overview. In: FAYYAD, U. M. et al. **Advances in Knowledge discovery and data mining**. Menlo Park: MIT Press, 1996.

[HER 95] HERRMANN, S. L.; GOLENDZINER, L. G.; SANTOS, C. S. dos**. Estudo sobre Mineração de Banco de Dados**. [s.n]. Porto Alegre: CPGCC da UFRGS, TI-499, 1995.

[HOU 95] HOUTSMA, M.; SWAMI, A. Set-Oriented Mining for Association Rules in Relational Databases. In: IEEE INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 11., 1995. **Proceeedings...** Los Alamitos: IEEE Computer Society Press, 1995.

[INM 2002] INMON, Bill. Billinmon.com. **Data Mining**. [S.l.]. Disponível em: <http://www.billinmon.com/>. Acesso em jan. 2002.

[PAR 89] PARSAYE, K.; et al. **Intelligent Databases**: object-oriented, deductive and hypermedia technologies. New York: John Willey, 1989.

[PAK 95] PARK, J.-S.; CHEN, M.-S. C.; YU, P.S. An Effective Hash Based Algorithm for Mining Association Rules. In: ACM SIGMOD INT. CONFERENCE ON MANAGEMENT OF DATA, 1995. **Proceedings**... San Jose: [s.n.], 1995. p. 175-186.

[MEN 98] MENESES, C. J.; GRINSTAIN, G. G. Categorization and Evaluation of Data Mining Techniques. In: **Data Mining**. Southampton: WIT Press, 1998. p.53-80.

[SHA 91] SHAPIRO, G. P. **Discovery, Analysis, and Presentation of Strong Rules. 1991** [S.l.]: AAAI Press: The MIT Press, 1991.

[SRI 96] SRIKANT, R.; AGRAWAL, R. **Mining Quantitative Association Rules in Large Relational Tables.** 1996. Trabalho apresentado na ACM SIGMOD International Conference on Management of Data.

[SRI 96b] SRIKANT, R. **Fast Algorithms for Mining Association Rules and Sequential Patterns**. 1996 (Ph.D Dissertation) - University of Wisconsin, Madison.

[WEI 98] WEISS, S. M.; INDURKHYA, N. **Predictive Data Mining. A practical guide**. San Francisco: Morgan Kaufmann Publisher Inc., 1998.

[ZAK 97] ZAKI, M. J.; PARTHASARATHY, S.; LI, W. A Localized Algorithm for Parallel Association Mining. In: ACM SYMPOSIUM ON PARALLEL ALGORITHMS AND ARCHITECTURES, 9., 1997. **Proceedings**... Newport: ACM Press, 1997. p. 321-330.

[ZAK 97a] ZAKI, M. J.; PARTHASARATHY, S.; LI, W. New Algorithms for Fast Discovery of Association Rules. In: ACM KDD INT. CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 3., 1997. **Proceedings**... Newport: AAAI Press, 1997. p. 283-286.

ANEXO A CAMPOS E RELACIONAMENTOS: CADASTRADOR, CBOR,MUNICÍPIO, DOMICÍLIO, PAÍS, TIPO LOGRADOURO, USUÁRIO

Tabela: CADASTRADOR

Propriedades

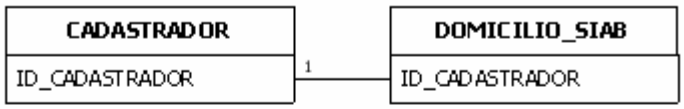
Descrição: Tabela de Cadastradores, o cadastrador é a pessoa responsável pela coleta dos dados para o sistema CadSUS nos domicílios.

Colunas

Nome		Tipo	Tamanho
ID_CADASTRADOR		Texto	6
	Descrição:	Identificador do cadastrador	
NO_CADASTRADOR		Texto	35
	Descrição:	Nome do cadastrador	

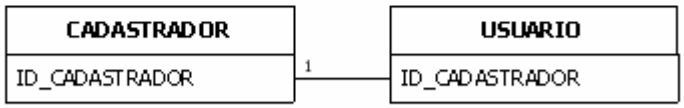
Relacionamentos

CADASTRADORDOMICILIO_SIAB



Atributos: Imposto
RelationshipType: Um-para-muitos

CADASTRADORUSUARIO



Atributos: Imposto
RelationshipType: Um-para-muitos

Tabela: CBOR

Propriedades

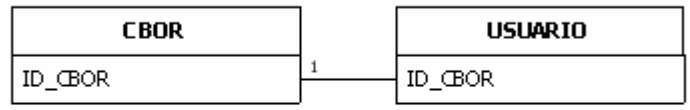
Descrição: Tabela de CBOR, Código Brasileiro de Ocupações

Colunas

Nome		Tipo	Tamanho
ID_CBOR		Texto	3
	Descrição:	Identificador do Código Brasileiro de Ocupações	
NO_DESCRICAO		Texto	200
	Descrição:	Descrição do Código Brasileiro de Ocupações	

Relacionamentos

CBORUSUARIO



Atributos: Imposto
RelationshipType: Um-para-muitos

Tabela: DOMICILIO_SIAB

Propriedades

Descrição: Tabelas de Domicílios e dados pertinentes a ele, cadastrados pelos sistema

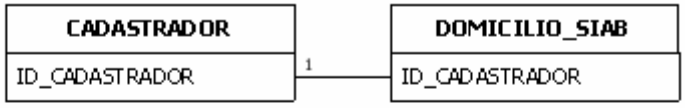
Colunas

Nome		Tipo	Tamanho
ID_DOMICILIO		Texto	48
	Descrição:	Identificador do domicílio (Para uso interno no sistema)	
CD_DOMICILIO		Texto	16
	Descrição:	Código do domicílio	
NR_FICHA		Texto	4
	Descrição:	Número da ficha domiciliar	
DT_PREENCHIMENTO_FORM		Data/Hora	8
	Descrição:	Data do preenchimento da ficha	
DT_INCLUSAO		Data/Hora	8
	Descrição:	Data da inclusão no sistema	
DT_ALTERACAO		Data/Hora	8
	Descrição:	Data da alteração no sistema	
CD_SEGMENTO		Texto	2
	Descrição:	Segmento (PSF ou outro programa do Ministério da Saúde)	
CD_AREA		Texto	4
	Descrição:	Área (PSF ou outro programa do Ministério da Saúde)	
CD_MICROAREA		Texto	2
	Descrição:	Microárea (PSF ou outro programa do Ministério da Saúde)	
CD_FAMILIA		Texto	3
	Descrição:	Família (PSF ou outro programa do Ministério da Saúde)	
ID_TIPO_LOGRADOURO		Texto	3
	Descrição:	Tipo de logradouro	
NO_LOGRADOURO		Texto	50
	Descrição:	Nome do logradouro	
NR_LOGRADOURO		Texto	7
	Descrição:	Número do domicilio no logradouro	
NO_COMPL_LOGRADOURO		Texto	15
	Descrição:	Complemento do endereço	
NO_BAIRRO		Texto	30
	Descrição:	Bairro	
CD_CEP		Texto	8
	Descrição:	CEP	
NR_DDD		Texto	3
	Descrição:	Código DDD	
NR_TELEFONE		Texto	9
	Descrição:	Número do telefone	
QT_PESSOAS		Texto	2
	Descrição:	Quantidade de pessoas no domicílio	
QT_COMODOS		Texto	2
	Descrição:	Quantidade de cômodos no domicílio	
IN_ENERGIA_ELETRICA		Texto	1
	Descrição:	Se possui Energia elétrica (S - Sim ou N - Não)	
IN_ESGOTO_SANITARIO		Texto	1
	Descrição:	Tipo de esgotamento sanitário (1 - Rede pública, 2 - Fossa, 3 - Céu aberto)	

IN_TIPO_DOMICILIO		Texto	1
	Descrição:	Tipo de Domicilio (1 - Tijolo / Alvenaria, 2 - Adobe, 3 - Taipa revestida, 4 - Taipa não revestida, 5 - Madeira, 6 - Material aproveitado, 7 - Outros)	
IN_DESTINO_LIXO		Texto	1
	Descrição:	Destino do lixo (1 - Coletado, 2 - Queimado/enterrado, 3 - Céu aberto)	
IN_ABASTECIMENTO_AGUA		Texto	1
	Descrição:	Tipo de abastecimento de água (1 - Rede pública, 2 - Poço ou Nascente, 3 - Outros)	
IN_TRATAMENTO_AGUA		Texto	1
	Descrição:	Tipo de tratamento de água (1 - Filtração, 2 - Fervura, 3 - Cloração, 4 - Sem tratamento)	
IN_PROGRAMAS_COBERTURA		Texto	1
	Descrição:	Domicílio coberto por programas do Ministério da Saúde (1 - PACS, 2 - PSF, 3 - Similares ao PSF, 4 - Outros)	
IN_SITUACAO		Texto	2
	Descrição:	Situação do cadastro (Para uso interno no sistema)	
MUNI_CD_COD_IBGE		Texto	7
	Descrição:	Município	
ID_CADASTRADOR		Texto	6
	Descrição:	Cadastrador	
LOTE_NR_LOTE		Texto	5
	Descrição:	Número do Lote para envio ao Ministério da Saúde (Para uso interno no sistema)	
LOTE_CD_COD_IBGE		Texto	7
	Descrição:	(Para uso interno no sistema)	
LOTE_ID_MAQUINA		Texto	15
	Descrição:	(Para uso interno no sistema)	
CD_COD_IBGE		Texto	7
	Descrição:	(Para uso interno no sistema)	
MUNI_NO_MUNICIPIO		Texto	40
	Descrição:	(Para uso interno no sistema)	
IN_STATUS_TRANS		Texto	1
	Descrição:	(Para uso interno no sistema)	
NO_RESERVA1		Texto	20
	Descrição:	(Para uso interno no sistema)	
FL_OPER		Texto	1
	Descrição:	(Para uso interno no sistema)	
FL_ARQ		Texto	1
	Descrição:	(Para uso interno no sistema)	
NR_USO_MUNICIPAL		Texto	20
	Descrição:	(Para uso interno no sistema)	
ST_PLANO		Texto	1
	Descrição:	(Para uso interno no sistema)	
NO_VETPLANO		Texto	30
	Descrição:	(Para uso interno no sistema)	
ID_ARQERRO		Duplo	8
	Descrição:	(Para uso interno no sistema)	
ID_ARQREDE		Duplo	8
	Descrição:	(Para uso interno no sistema)	
ID_ARQRET		Duplo	8
	Descrição:	(Para uso interno no sistema)	

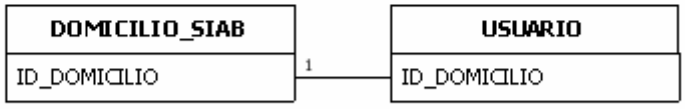
Relacionamentos

CADASTRADORDOMICILIO_SIAB



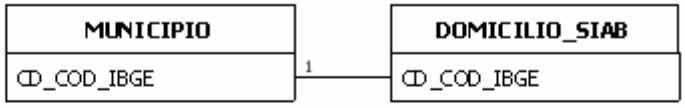
Atributos: Imposto
RelationshipType: Um-para-muitos

DOMICILIO_SIABUSUARIO



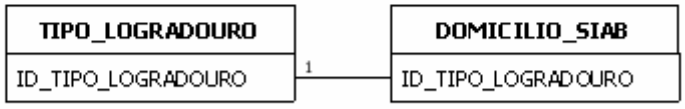
Atributos: Imposto
RelationshipType: Um-para-muitos

MUNICIPIODOMICILIO_SIAB



Atributos: Imposto
RelationshipType: Um-para-muitos

TIPO_LOGRADOURODOMICILIO_SIAB



Atributos: Imposto
RelationshipType: Um-para-muitos

Tabela: MUNICIPIO

Propriedades

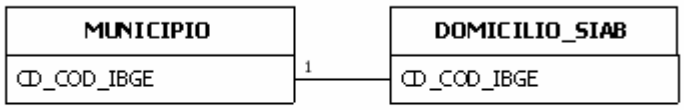
Descrição:Tabela de municípios.

Colunas

Nome		Tipo	Tamanho
CD_COD_IBGE		Texto	7
	Descrição:	Código do IBGE do município.	
NO_MUNICIPIO		Texto	40
	Descrição:	Nome do município	

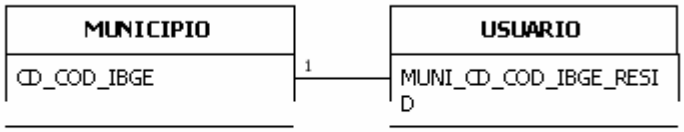
Relacionamentos

MUNICIPIODOMICILIO_SIAB



Atributos: Imposto
RelationshipType: Um-para-muitos

MUNICIPIOUSUARIO



Atributos: Imposto
RelationshipType: Um-para-muitos

Tabela: PAIS

Propriedades

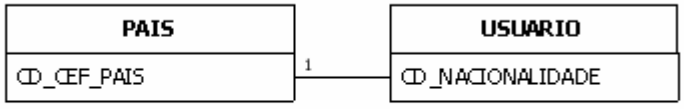
Descrição: Tabela de Paises

Colunas

Nome		Tipo	Tamanho
CD_CEF_PAIS		Texto	3
	Descrição:	Código do Pais	
NO_PAIS		Texto	50
	Descrição:	Nome do Pais	

Relacionamentos

PAISUSUARIO



Atributos: Imposto
RelationshipType: Um-para-muitos

Tabela: TIPO_LOGRADOURO

Propriedades

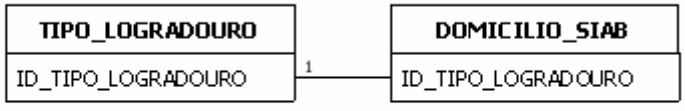
Descrição: Tabela de Tipos de Logradouro

Colunas

Nome		Tipo	Tamanho
ID_TIPO_LOGRADOURO		Texto	3
	Descrição:	Identificador do Tipo de Logradouro	
NO_TIPO_LOGRADOURO		Texto	30
	Descrição:	Nome do tipo de logradouro	
NO_SIGLA_LOGRADOURO		Texto	10
	Descrição:	Sigla do tipo de logradouro	

Relacionamentos

TIPO_LOGRADOURODOMICILIO_SIAB



Atributos: Imposto
RelationshipType: Um-para-muitos

Tabela: UF

Propriedades

Descrição: Tabela de Unidades da Federação

Colunas

Nome		Tipo	Tamanho
NO_SIGLA_UF		Texto	2
	Descrição:	Sigla da Unidade da Federação	
NO_UF		Texto	20
	Descrição:	Nome da Unidade da Federação	
CD_UF		Texto	2
	Descrição:	Código da Unidade da Federação	

Relacionamentos

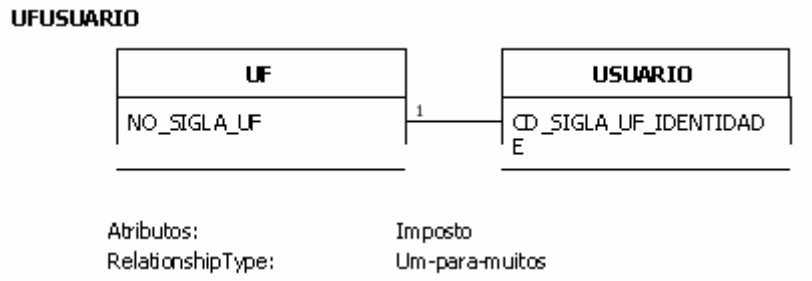


Tabela: USUARIO

Propriedades

Descrição: Tabela de Usuários do SUS

Colunas

Nome		Tipo	Tamanho
ID_USUARIO		Texto	58
	Descrição:	Identificador do usuário (Para uso interno no sistema)	
NR_USO_MUNICIPAL		Texto	20
	Descrição:	(Para uso interno no sistema)	
NR_FICHA		Texto	4
	Descrição:	Número da ficha domiciliar	
NO_USUARIO		Texto	70
	Descrição:	Nome do usuário	
NO_FONETICO_USUARIO		Texto	70
	Descrição:	Nome fonético do usuário (Para uso interno no sistema) não implementado	
IN_RACA_COR		Texto	1
	Descrição:	Raça ou cor do usuário (1 - Branca, 2 - Preta, 3 - Amarela, 4 - Parda, 5 - Indígena)	
IN_SITUACAO_CONJUGAL		Texto	2
	Descrição:	Situação familiar/conjugal (1 - Convive com companheira(o) e filho(s), 2 - Convive com companheira(o) com laços conjugais e sem filhos, 3 - Convive com companheira(o) com filho(s) e/ou outro(s) familiares, 4 - Convive com familiar(es) sem companheira(o), 5 - Convive com outra(s) pessoa(s), sem laços consangüíneos e/ou laços conjugais, 6 – Vive só	
IN_ESCOLARIDADE		Texto	2
	Descrição:	Escolaridade do usuário (1 -Não sabe ler/escrever, 2 - Alfabetizado – indivíduo que sabe ler e escrever no mínimo um bilhete. O indivíduo que apenas assina o nome não é considerado alfabetizado, 3 - Nível Fundamental Incompleto (1° grau incompleto), 4 - Nível Fundamental Completo (1°	

grau completo), 5 - Nível Médio Incompleto (2° grau incompleto), 6 - Nível Médio Completo (2º grau completo), 7 - Superior incompleto, 8 - Superior completo, 9 - Especialização, 10 - Mestrado, 11 - Doutorado.

DT_NASCIMENTO Data/Hora 8
Descrição: Data de Nascimento do usuário.
IN_SEXO Texto 1
Descrição: Sexo do usuário (M – Masculino, F – Feminino)
NO_PAI Texto 70
Descrição: Nome do pai do usuário.
NO_FONETICO_PAI Texto 70
Descrição: Nome fonético do pai usuário (Para uso interno no sistema) não implementado
NO_MAE Texto 70
Descrição: Nome da mão do usuário
NO_FONETICO_MAE Texto 70
Descrição: Nome fonético da mãe usuário (Para uso interno no sistema) não implementado
NR_CNS Texto 15
Descrição: Número do Cartão Nacional de Saúde (Para uso interno no sistema)
NR_PIS_PASEP Texto 11
Descrição: Número do PIS/PASEP do usuário
NR_CPF Texto 11
Descrição: Número do CPF do usuário.
CD_TIPO_CERTIDAO Texto 2
Descrição: Tipo de certidão (1 - Certidões de nascimento, 2 - Certidões de casamento, 3 - Certidões de separação ou divórcio).
NO_CARTORIO_CERTIDAO Texto 20
Descrição: Nome do cartório.
NR_LIVRO_CERTIDAO Texto 8
Descrição: Número do livro.
NR_FOLHA_CERTIDAO Texto 4
Descrição: Número da folha.
NO_TERMO_CERTIDAO Texto 8
Descrição: Número da certidão constante do livro em que foi lavrada naquele cartório.
IN_FREQUENTA_ESCOLA Texto 1
Descrição: Freqüenta a escola (S – sim, N – Não).
DT_EMISSAO_CERTIDAO Data/Hora 8
Descrição: Data da emissão da certidão.
NR_IDENTIDADE Texto 11
Descrição: Número da identidade.
NO_COMPL_IDENTIDADE Texto 4
Descrição: Complemento da Identidade.
CD_SIGLA_UF_IDENTIDADE Texto 2
Descrição: Estado de emissão da identidade.
CD_ORGAO_EMISSOR_IDENTIDADE Texto 2
Descrição: Órgão emissor da identidade (
10- SSP – Secretaria de Segurança Pública;
41- Ministério da Aeronáutica;
42- Ministério do Exército;
43- Ministério da Marinha;
44- Polícia Federal;
60- Carteira de Identidade Classista;
61- Conselho Regional de Administração;
62- Conselho Regional de Assistentes Sociais;
63- Conselho Regional de Biblioteconomia;
64- Conselho Regional de Contabilidade;
65- Conselho Regional de Corretores de Imóveis;

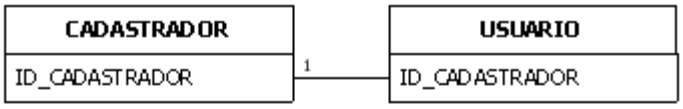
66- Conselho Regional de Enfermagem;
67- Conselho Regional de Engenharia, Arquitetura e Agronomia;
68- Conselho Regional de Estatística;
69- Conselho Regional de Farmácia;
70- Conselho Regional de Fisioterapia e Terapia Ocupacional;
71- Conselho Regional de Medicina;
72- Conselho Regional de Medicina Veterinária;
73- Ordem dos Músicos do Brasil;
74- Conselho Regional de Nutrição;
75- Conselho Regional de Odontologia;
76- Conselho Regional de Profissionais de Relações Públicas;
77- Conselho Regional de Psicologia;
78- Conselho Regional de Química;
79- Conselho Regional de Representantes Comerciais;
80- Ordem dos Advogados do Brasil;
81- Outros Emissores;
82- Documentos Estrangeiros).

DT_EMISSAO_IDENTIDADE		Data/Hora	8
	Descrição:	Data da emissão da identidade.	
NR_CTPS		Texto	7
	Descrição:	Número da Carteira de Trabalho e Previdência Social.	
NR_SERIE_CTPS		Texto	5
	Descrição:	Número da Série da Carteira de Trabalho e Previdência Social.	
CD_SIGLA_UF_CTPS		Texto	2
	Descrição:	Estado da Carteira de Trabalho e Previdência Social.	
DT_EMISSAO_CTPS		Data/Hora	8
	Descrição:	Data da emissão da Carteira de Trabalho e Previdência Social.	
NR_TITULO_ELEITOR		Texto	13
	Descrição:	Número do título de eleitor.	
NR_ZONA_TITULO_ELEITOR		Texto	4
	Descrição:	Número da zona do título de eleitor.	
NR_SECAO_TITULO_ELEITOR		Texto	4
	Descrição:	Número da seção do título de eleitor.	
IN_SITUACAO		Texto	2
	Descrição:	(Uso interno do sistema)	
CD_NACIONALIDADE		Texto	3
	Descrição:	Nacionalidade (Brasileiro ou Estrangeiro).	
DT_ENTRADA_PAIS		Data/Hora	8
	Descrição:	Data de Entrada no pais.	
NR_PORTARIA_NATURALIZACAO		Texto	16
	Descrição:	Número da portaria de naturalização.	
DT_NATURALIZACAO		Data/Hora	8
	Descrição:	Data de naturalização.	
DT_PREENCHIMENTO_FORM		Data/Hora	8
	Descrição:	Data do preenchimento do formulário.	
DT_INCLUSAO		Data/Hora	8
	Descrição:	Data de inclusão (Uso interno do sistema).	
DT_ALTERACAO		Data/Hora	8
	Descrição:	Data de alteração (Uso interno do sistema).	
ID_DOMICILIO		Texto	48
	Descrição:	Número do domicílio (Uso interno do sistema).	
MUNI_CD_COD_IBGE_RESID		Texto	7
	Descrição:	Município (Uso interno do sistema).	
MUNI_CD_COD_IBGE_NASC		Texto	7
	Descrição:	Município de Nascimento.	
MUNI_NO_MUNICIPIO_NASC		Texto	40
	Descrição:	Nome do município de nascimento.	
ID_CBOR		Texto	3
	Descrição:	Classificação Brasileira de Ocupações – Reduzida.	
ID_CADASTRADOR		Texto	6
	Descrição:	Código do cadastrador.	

IN_PGMSOC_BOLSA_ALIM		Texto	1
	Descrição:	(Uso interno do sistema).	
IN_PGMSOC_PRODEA		Texto	1
	Descrição:	(Uso interno do sistema).	
IN_PGMSOC_5		Texto	1
	Descrição:	(Uso interno do sistema).	
	Posição ordinal:	55	
	Requerido:	Falso	
IN_PGMSOC_6		Texto	1
	Descrição:	(Uso interno do sistema).	
NO_RESERVA1		Texto	20
	Descrição:	(Uso interno do sistema).	
IN_ERRO_LOG		Texto	1
	Descrição:	(Uso interno do sistema).	
IN_STATUS_TRANS		Texto	1
	Descrição:	(Uso interno do sistema).	
FL_OPER		Texto	1
	Descrição:	(Uso interno do sistema).	
FL_ARQ		Texto	1
	Descrição:	(Uso interno do sistema).	
NR_PESSOA_NO_DOMIC		Texto	2
	Descrição:	Número da pessoa no domicílio.	
FL_ERROS		Texto	1
	Descrição:	(Uso interno do sistema).	
FL_VINC_DOMICILIO		Texto	1
	Descrição:	(Uso interno do sistema).	
FL_VINC_SIAB		Texto	1
	Descrição:	(Uso interno do sistema).	
CD_DOMICILIO		Texto	16
	Descrição:	Código do domicílio (Uso interno do sistema).	
NO_USERNAME		Texto	8
	Descrição:	(Uso interno do sistema).	
NR_VERSAO		Texto	5
	Descrição:	(Uso interno do sistema).	
ST_PLANO		Texto	1
	Descrição:	(Uso interno do sistema).	
NO_VETPLANO		Texto	30
	Descrição:	(Uso interno do sistema).	
FL_ARQBA		Texto	1
	Descrição:	(Uso interno do sistema).	
FL_PLANOSCEF		Texto	1
	Descrição:	(Uso interno do sistema).	
ID_ARQERRO		Duplo	8
	Descrição:	(Uso interno do sistema).	
ID_ARQREDE		Duplo	8
	Descrição:	(Uso interno do sistema).	
ID_ARQRET		Duplo	8
	Descrição:	(Uso interno do sistema).	

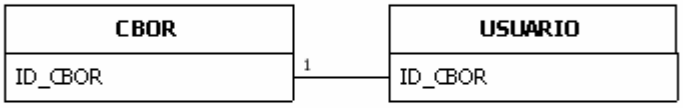
Relacionamentos

CADASTRADORUSUARIO



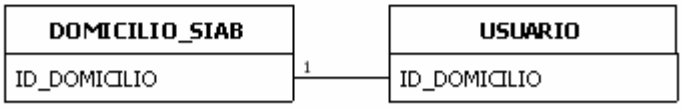
Atributos: Imposto
RelationshipType: Um-para-muitos

CBORUSUARIO



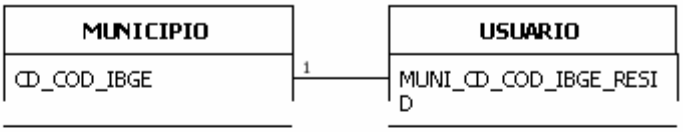
Atributos: Imposto
RelationshipType: Um-para-muitos

DOMICILIO_SIABUSUARIO



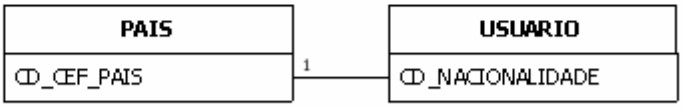
Atributos: Imposto
RelationshipType: Um-para-muitos

MUNICIPIOUSUARIO

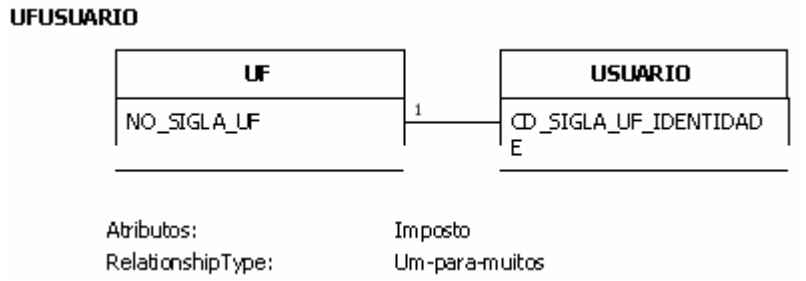


Atributos: Imposto
RelationshipType: Um-para-muitos

PAISUSUARIO



Atributos: Imposto
RelationshipType: Um-para-muitos


UFUSUARIO
UF
NO_SIGLA_UF
1
USUARIO
CD_SIGLA_UF_IDENTIDADE
Atributos:
Imposto
RelationshipType:
Um-para-muitos

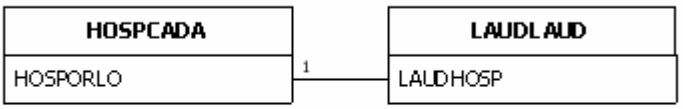
ANEXO B CAMPOS E RELACIONAMENTOS: HOSPITAIS, LAUDOS, MEDICOS, MUNICÍPIOS E PROCEDIMENTOS

Tabela: HOSPCADA
Propriedades
Descrição: Tabela de Hospitais
Colunas

Nome			Tipo	Tamanho
HOSPORLO			Texto	7
	Descrição:	Código do hospital		
HOSPCGC			Texto	14
	Descrição:	CGC ou CNPJ do hospital		
HOSPNOME			Texto	60
	Descrição:	Nome do hospital		
HOSPENDE			Texto	20
	Descrição:	Endereço do hospital		
HOSPNUME			Texto	5
	Descrição:	Número do endereço do hospital		
HOSPCOMP			Texto	6
	Descrição:	Complemento do endereço do hospital		
HOSPBAIR			Texto	12
	Descrição:	Bairro do hospital		
HOSPCEP			Texto	8
	Descrição:	CEP do hospital		
HOSPMUNI			Texto	20
	Descrição:	Município do hospital		
HOSPUF			Texto	2
	Descrição:	Unidade da federação do hospital		
HOSPCBAN			Texto	5
	Descrição:	Número do banco utilizados pelo hospital		
HOSPCONT			Texto	14
	Descrição:	Conta corrente no banco utilizado pelo hospital		
HOSPNOBA			Texto	12
	Descrição:	Nome do banco utilizado pelo hospital		
HOSPAGEN			Texto	12
	Descrição:	Agência no banco utilizado pelo hospital		
HOSPTOTC			Inteiro	2
	Descrição:	(uso interno do sistema)		
HOSPTOTL			Inteiro	2
	Descrição:	(uso interno do sistema)		
HOSPNATU			Texto	2
	Descrição:	(uso interno do sistema)		
HOSPBLPG			Texto	1
	Descrição:	(uso interno do sistema)		
HOSPSIST			Texto	10
	Descrição:	(uso interno do sistema)		
HOSPAUTO			Texto	1
	Descrição:	(uso interno do sistema)		

Relacionamentos

HOSPCADALAUDLAUD



Atributos:	Imposto
RelationshipType:	Um-para-muitos

Tabela: LAUDLAUD

Propriedades

Descrição: Tabela de armazenamento de laudos

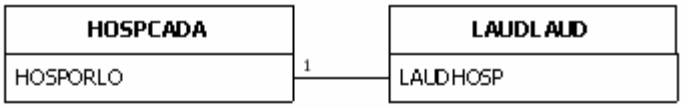
Colunas

Nome			Tipo	Tamanho
LAUDHOSP			Texto	7
	Descrição:	Hospital em que foi realizada a internação.		
LAUDNOME			Texto	40
	Descrição:	Nome do paciente.		
LAUDENDE			Texto	40
	Descrição:	Endereço do paciente.		
LAUDMUNI			Texto	7
	Descrição:	Município do paciente.		
LAUDUF			Texto	2
	Descrição:	Estado do paciente.		
LAUDCEP			Texto	8
	Descrição:	CEP do paciente.		
LAUDDTNA			Data/Hora	8
	Descrição:	Data de nascimento do paciente.		
LAUDSEXO			Texto	1
	Descrição:	Sexo do paciente (1 - Masculino e 3 - Feminino).		
LAUDCPFM			Texto	11
	Descrição:	CPF do médico responsável.		
LAUDPROC			Texto	8
	Descrição:	Procedimento principal realizado.		
LAUDCINT			Texto	1
	Descrição:	Caráter da internação (1 – Eletivo, 2 – Urgência / Emergência em hospital de referência, 3 – Urgência / Emergência (extinto), 4 – Câmara de compensação (outros municípios), 5 – Urgência / Emergência, 6 – Acidente no trabalho, 7 – Acidente a caminho do trabalho, 8 – Acidente de transito quando a caminho do trabalho, 9 – Lesões e envenenamentos causados por agentes físicos ou químicos no trabalho).		
	Posição ordinal:	11		
	Requerido:	Falso		
LAUDDTCA			Data/Hora	8
	Descrição:	Data do cadastro do laudo.		
LAUDCLIN			Texto	1
	Descrição:	Clínica média (1 - Cirúrgica, 2 - Obstétrica, 3 - Clínica Médica, 4 - Tisiopneumologica, 5 -Psiquiatria, 7 - Pediatria)		
LAUDNAIH			Texto	10
	Descrição:	Número da AIH (carregado assim que for apresentado o disquete de faturamento)		
LAUDDCMO			Texto	40
	Descrição:	(uso interno do sistema)		
LAUDDH53			Data/Hora	8
	Descrição:	(uso interno do sistema)		
LAUDDH54			Data/Hora	8
	Descrição:	(uso interno do sistema)		
LAUDPRIM			Texto	4
	Descrição:	(uso interno do sistema)		

LAUDLIBE		Texto	4
	Descrição:	(uso interno do sistema)	
LAUDDTDI		Data/Hora	8
	Descrição:	(uso interno do sistema)	
LAUDOPER		Texto	1
	Descrição:	(uso interno do sistema)	

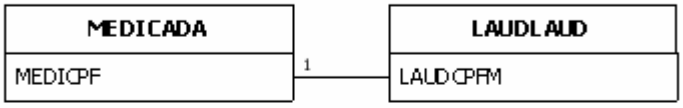
Relacionamentos

HOSPCADALAUDLAUD



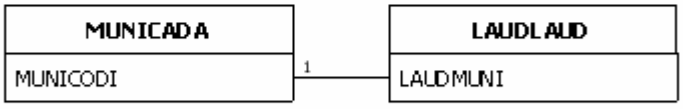
Atributos: Imposto
RelationshipType: Um-para-muitos

MEDICADALAUDLAUD



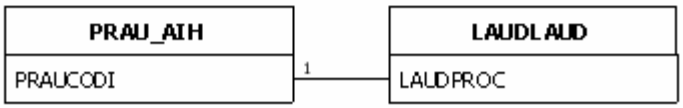
Atributos: Imposto
RelationshipType: Um-para-muitos

MUNICADALAUDLAUD



Atributos: Imposto
RelationshipType: Um-para-muitos

PRAU_AIHLAUDLAUD



Atributos: Imposto
RelationshipType: Um-para-muitos

Tabela: MEDICADA

Propriedades

Descrição: Tabelas de Médicos

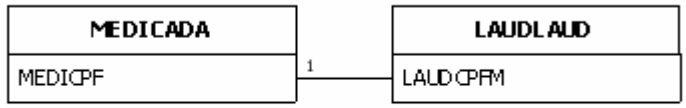
Colunas

Nome	Tipo	Tamanho

MEDICPF			Texto	11
	Descrição:	CPF do médico		
MEDINOME			Texto	40
	Descrição:	Nome do médico		
MEDICRM	Texto		5	
	Descrição:	CRM do médico		

Relacionamentos

MEDICADALAUDLAUD



Atributos: Imposto
RelationshipType: Um-para-muitos

Tabela: MUNICADA

Propriedades

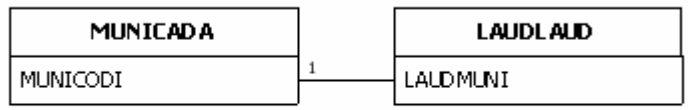
Descrição: Tabelas de muncípios

Colunas

Nome		Tipo	Tamanho
MUNICODI		Texto	7
	Descrição:	Código do IBJE do município	
MUNINOME		Texto	33
	Descrição:	Nome do município	
MUNIREGG		Texto	2
	Descrição:	Regional de Saúde do município	

Relacionamentos

MUNICADALAUDLAUD



Atributos: Imposto
RelationshipType: Um-para-muitos

Tabela: PRAU_AIH

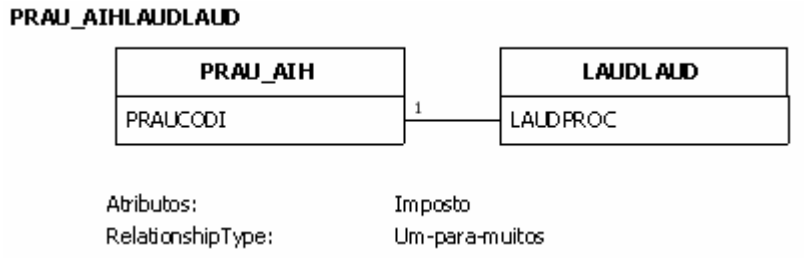
Propriedades

Descrição: Tabela de procedimentos.

Colunas

Nome		Tipo	Tamanho
PRAUCODI		Texto	8
	Descrição:	Código do procedimento	
PRAUDESC		Texto	50
	Descrição:	Descrição do procedimento	
PRAUVALO		Duplo	8
	Descrição:	Valor do procedimento	

Relacionamentos


PRAU_AIHLAUDLAUD
PRAU_AIH
PRAUCODI
1
LAUDLAUD
LAUDPROC
Atributos:
Imposto
RelationshipType:
Um-para-muitos