Universidade Federal de Santa Catarina - UFSC
Programa de Pós-Graduação em
Engenharia da Produção

MINERAÇÃO DE DADOS DE UM PLANO DE SAÚDE PARA OBTER REGRAS DE ASSOCIAÇÃO

Otávio Roberto Martins de Souza

Dissertação apresentada ao
Programa de Pós-Graduação em
Engenharia de Produção da
Universidade Federal de Santa Catarina
como requisito parcial para obtenção
do título de Mestre em
Engenharia de Produção

Florianópolis
2000

Otávio Roberto Martins de Souza

MINERAÇÃO DE DADOS DE UM PLANO DE SAÚDE PARA OBTER REGRAS DE ASSOCIAÇÃO

Esta dissertação foi julgada e aprovada para a obtenção do título de
Mestre em Engenharia de Produção no
Programa de Pós-Graduação em Engenharia de Produção da
Universidade Federal de Santa Catarina

Florianópolis, 26 de setembro de 2000

Prof. Ricardo Miranda Barcia, Ph.D.
Coordenador do Curso

BANCA EXAMINADORA

_____
Prof. Fernando Alvaro Ostuni Gauthier, Dr
Orientador

_____ _____
Prof. Alejandro Martins, Dr. Prof. João Bosco da Mota Alves, Dr

AGRADECIMENTOS

Meus agradecimentos a todas as pessoas que contribuíram para o desenvolvimento deste trabalho. Em especial agradeço:

À Eliane, minha esposa, incentivadora e colaboradora, pelo carinho e compreensão e por não me deixar desanimar frente aos obstáculos que se apresentaram para a conclusão deste trabalho.

Às minhas filhas Carla e Luciana, pelo apoio ao longo deste trabalho.

Ao Prof. Fernando Gauthier, Dr. Eng, meu orientador, pelo apoio e contribuições para a realização deste trabalho.

À Fundação Eletrosul de Previdência e Assistência Social-ELOS, pelo apoio para freqüentar o curso de pós-.graduação.

Ao ELOSAUDE, na pessoa de José Anastácio Fernandes, pelo apoio e cessão de dados, fundamentais para viabilizar a realização deste trabalho e na avaliação dos resultados obtidos.

Ao atuário da ELOS, Luis Guilherme Valles pelo apoio na escolha de estudos de interesse para o ELOSAUDE e na avaliação de resultados.

Ao pesquisador Eibe Frank, da University of Waikato, Department of Computer Science, New Zealand, pelo apoio via e-mail em detalhes técnicos de seu programa *APRIORI* do projeto Weka (Waikato Environment for Knowledge Analysis).

Nenhum vento sopra a favor
de quem não sabe para onde ir
Sêneca

SUMÁRIO

LISTA DE FIGURAS

RESUMO

As organizações estão investindo cada vez mais na exploração da informação e conhecimento existentes nos dados de suas atividades. A mineração de dados representa um conjunto de técnicas para obtenção de informação que não pode ser obtida através de consultas convencionais. Uma destas técnicas é denominada mineração de regras de associação. Regras de associação são expressões que indicam afinidade ou correlação entre dados. Este trabalho avalia o potencial de utilidade do algoritmo **APRIORI**, um indutor de regras de associação, aplicando-o a dados de um **plano de saúde**, apresentando os resultados obtidos e analisando seu significado.

ABSTRACT

The organizations are increasingly investing in exploring the information and knowledge embedded in the data of their activities. Data mining represents a set of techniques to obtain information that cannot be obtained through conventional queries. One of these techniques is called association rules mining. Association rules are expressions that indicate affinity or correlation among data. This work evaluates the potential of utility of the **APRIORI** algorithm, an association rules inductor, by its application to data from a **health care plan**, showing the results obtained and the analysis of their meaning.

1 INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

A década de 1990 trouxe um crescente problema de abundância de dados para os mundos da ciência, negócios, e governo. A capacidade para colecionar e armazenar dados de toda espécie de longe ultrapassou as habilidades de analisar, sumarizar, e extrair conhecimento destes dados. Métodos tradicionais de análise de dados, baseados principalmente em humanos tratando diretamente com os dados, simplesmente não tem escala para manipular volumosos conjuntos de dados [FAY96].

A convergência de computação e comunicação produziu uma sociedade que se alimenta de informação. A maioria da informação ainda está sob a forma crua: dado. Se dado é caracterizado como fatos registrados, então informação é o conjunto de padrões, ou expectativas, que estão implícitos nos dados[WIT00].

Descobrimento de conhecimento é o mais desejável produto final da computação. Encontrar novos fenômenos ou aumentar nosso conhecimento sobre eles tem maior valor a longo prazo que otimizar processos de produção ou inventário, ficando em segundo lugar somente para tarefas que ajudam a preservar nosso planeta e nosso meio ambiente.(Wiederhold em [FAY96])

Existe uma enorme quantidade de informação presa em bases de dados, informação que é potencialmente importante mas que ainda não foi descoberta.

Mineração de dados é a extração de informação implícita, previamente desconhecida, e potencialmente útil, a partir de dados[WIT00]. O processo deve ser automático ou, mais usualmente, semi automático. Os padrões descobertos devem ser significativos, na medida em que direcionam para alguma vantagem, usualmente uma vantagem econômica.

Relatórios exagerados proclamam os segredos que podem ser descobertos pondo algoritmos de aprendizado a trabalhar perdidos em oceanos de dados. Mas não existe mágica em aprendizado por máquina, força oculta, nem alquimia. Ao contrário, existe um conjunto de técnicas simples e práticas que muitas vezes podem extrair informação de dado cru.

1.2 JUSTIFICATIVA

Descobrir informação a partir de dados, ou mineração de dados, vem recebendo cada vez mais atenção de pesquisadores e de grandes corporações. Há diversos enfoques e técnicas para sua consecução, sendo a obtenção de regras de associação uma das alternativas. Este trabalho disponibiliza informações sobre o algoritmo **APRIORI** através de sua utilização a dados reais de um plano de saúde, e análise dos resultados obtidos.

1.3 OBJETIVOS

1.3.1 OBJETIVO GERAL

Utilizar o algoritmo **APRIORI** para avaliar sua potencialidade na indução de regras de associação em dados de um plano de saúde.

1.3.2 OBJETIVOS ESPECÍFICOS

- Identificar estratégias para realizar estudos específicos.
- Induzir regras pertinentes a plano de saúde.
- Verificar a aplicabilidade do algoritmo **APRIORI** para minerar dados de plano de saúde.
- Avaliar a sensibilidade do algoritmo **APRIORI** a seus parâmetros básicos **suporte mínimo** e **confiança mínima**.

1.4 LIMITAÇÕES

Este trabalho limita-se ao estudo do algoritmo **APRIORI**. Desta forma, não faz comparação com outros algoritmos existentes para mineração de dados.

1.5 ESTRUTURA DO TRABALHO

O presente trabalho está estruturado em dez capítulos.

- No primeiro capítulo é apresentada uma introdução ao trabalho desenvolvido, os objetivos e as limitações do mesmo;
- No capítulo 2 são apresentados conceitos relativos à mineração de dados e uma descrição do algoritmo **APRIORI**.

- O capítulo 3 discorre sobre planos de saúde em geral, e sobre o Elosaude em particular, serviços que oferece, critérios para atendimento e reembolso.
- No capítulo 4 é apresentado o modelo proposto para mineração de dados de um plano de saúde para obter regras de associação.
- No capítulo 5 é relatada a apresentada a aplicação prática do trabalho, relatando os estudos feitos, os dados utilizados, os resultados obtidos e análises dos resultados obtidos em cada estudo.
- No capítulo 6 são apresentadas as conclusões e recomendações relativas ao trabalho.
- No capítulo 7, estão anexados:
  - a bibliografia utilizada no desenvolvimento deste trabalho.
  - uma relação de sites da web referentes à mineração de dados.
  - o glossário de termos.
  - uma tabela de tipos de serviços do plano de saúde Elosaude.

2 MINERAÇÃO DE DADOS

Este tópico examina questões fundamentais sobre mineração de dados: o que é mineração de dados, porque é valiosa, e como minerar dados.
Metodologia e terminologia de mineração de dados são discutidos.

2.1 CONCEITO

Mineração de dados(MD) é o processo de automatizar a descoberta de informação[GRO98]. Embora existam em abundância ferramentas para consultar, acessar, e manipular dados, o usuário é deixado abandonado quando precisa encontrar tendências e padrões úteis. A mineração de dados automatiza o processo de descobrimento de tendências e padrões úteis.
No centro da MD está o processo de construção do modelo. Criar um modelo representativo baseado num conjunto existente de dados provou ser útil para compreender tendências, padrões, e correlações.

2.2 FERRAMENTAS DE CONSULTA

Ferramentas de Consulta permitem ao usuário elaborar perguntas típicas de Sistemas de Gerenciamento de Base de Dados (SGBD), obtendo fatos que foram armazenados numa Base de Dados [GRO98], que são óbvios, porque estão explicitamente armazenados. Tipicamente, a SQL (Structured Query Language),

inerente a todo SGBD é uma ferramenta de consulta, como o significado de sua sigla indica, Linguagem Estruturada de Consulta.

2.3 FERRAMENTAS DE MINERAÇÃO DE DADOS

Ferramentas de MD tentam descobrir relacionamentos e padrões ocultos que podem não ser óbvios[GRO98].
Uma curiosidade, que reforça a percepção de uma crescente demanda por ferramentas de MD, é o recente anúncio da Microsoft sobre a inclusão de recursos para MD na versão 2000 do seu SGBD, MS SQLServer a ser lançado no segundo semestre do ano 2000. Isto representa uma tentativa de disponibilizar recursos de MD, mais baratos e mais simples além de integrados ao SGBD.

2.4 REPRESENTAÇÃO DO CONHECIMENTO

Para Ian H. Witten [WIT00] há muitas maneiras diferentes para representar os padrões que podem ser descobertos pelo aprendizado por máquina, e cada um dita o tipo de técnica que pode ser usada para inferir a estrutura de saída a partir de dados.

2.4.1 TABELAS DE DECISÃO

A mais simples, mais rudimentar maneira de representar o resultado de uma máquina de aprendizado é faze-lo parecido com a entrada – uma tabela de decisão[WIT00], que é um modelo alternativo para uma função. Ela representa a função em forma tabular ou matricial; as linhas superiores da tabela especificam as variáveis ou condições a serem avaliadas, e as linhas inferiores especificam a ação correspondente a ser executada quando um teste de avaliação é satisfeito. Uma coluna na tabela é chamada de regra. Cada regra define um procedimento do tipo: se a condição for verdadeira, executar a ação correspondente[MAR91].

2.4.2 ÁRVORES DE DECISÃO

Um enfoque de dividir e conquistar o problema de aprender de um conjunto de instâncias direciona naturalmente para um estilo de representação denominado árvore de decisão[WIT00]. Uma árvore de decisão é um modelo de um função discreta no qual é determinado o valor de uma variável; com base neste valor, alguma ação é executada. Ela dá uma visão gráfica da tomada de decisão necessária. Especificam que variáveis são testadas, que ações devem ser executadas e a ordem em que a tomada de decisão é executada.[MAR91]

2.4.3 REGRAS DE CLASSIFICAÇÃO

Regras de classificação são uma alternativa para árvores de decisão. É fácil ler um conjunto de regras diretamente de uma árvore de decisão. Uma regra é gerada para cada folha. O antecedente da regra inclui uma condição para cada nó na rota desde a raiz até a folha, e o conseqüente da regra é a classe assinalada pela folha.

Se aspecto = ensolarado e umidade > 83 então joga = não

Regras de classificação são do tipo *proposicional*. Envolvem testar o valor de um atributo contra uma constante.

2.4.4 REGRAS DE ASSOCIAÇÃO

Regras de associação não são realmente diferentes de regras de classificação, exceto que podem prognosticar qualquer atributo, não apenas a classe, e isto lhes dá a liberdade para prognosticar combinações de atributos também. Ainda, regras

de associação não se destinam a serem usadas juntas como um conjunto, como são as regras de classificação. Diferentes regras de associação expressam diferentes regularidades intrínsecas ao conjunto de dados, e geralmente prognosticam coisas diferentes.

se umidade = normal e ventoso = falso então joga = sim

2.4.5 REGRAS COM EXCEÇÕES

Retornando às regras de classificação, uma extensão natural é permitir que tenham exceções. Então modificações incrementais podem ser feitas a um conjunto de regras expressando exceções às regras existentes ao invés de reconstruir o conjunto todo.

Se comprimento-petala maior que 2.45 e comprimento-petala
Menor que 4.45 então Iris-versicolor
Exceto Se largura-petala menor que 1.0
Então iris-setosa

2.4.6 REGRAS ENVOLVENDO RELAÇÕES

Em muitas tarefas de classificação, regras *proposicionais* são suficientemente expressivas para descrições de conceitos com concisão e precisão. Entretanto, há situações onde uma forma mais expressiva de regra poderia prover uma descrição de conceitos mais intuitiva e concisa. São situações que envolvem relacionamentos entre exemplos.

Se largura maior que altura então horizontal
Se altura maior que largura então vertical

Os valores dos atributos altura e largura não são importantes, apenas o resultado da comparação dos dois. Regras desta forma são chamadas *relacionais*, porque expressam relacionamentos entre atributos, ao invés de proposicionais, que simbolizam um fato sobre apenas um atributo.

2.4.7 ÁRVORES PARA PROGNÓSTICO NUMÉRICO

As árvores de decisão e regras apresentadas anteriormente foram projetadas para prognóstico de categorias e não de quantidades numéricas. Quando se trata de prognóstico de quantidades numéricas, a mesma espécie de representação em árvore ou regra pode ser usada, mas os nós folha da árvore, ou o lado direito das regras, contém um valor numérico que é a média de todos os valores do conjunto de treinamento a que a folha, ou regra, se aplicam.

2.4.8 REPRESENTAÇÃO BASEADA EM INSTÂNCIA

A mais simples forma de aprendizado é a plena memorização, ou aprendizado de rotina (rote learning). A representação de conhecimento baseada em instância usa a própria instância para representar o que é aprendido, ao invés de inferir um conjunto de regras ou árvore de decisão.

Uma vez que um conjunto de instâncias de treinamento tenham sido memorizados, ao encontrar uma nova instância a memória é pesquisada em busca da instância que mais se assemelhe à nova. O único problema é como interpretar semelhança. É também chamado de método de classificação vizinho mais próximo (nearest-neighbor).

2.4.9 AGRUPAMENTOS

Quando é aprendido agrupamentos ao invés de um classificador, a saída toma a forma de um diagrama que mostra como as instâncias caem dentro dos

agrupamentos. No caso mais simples isto envolve associar um número do agrupamento com cada instância.

2.5 TERMINOLOGIA DE MINERAÇÃO DE DADOS

Diversos conceitos e objetos são fundamentais para MD.
Existe uma guerra de terminologia em MD, sendo conveniente estar atento ao fato de não haver um consenso sobre como as palavras são usadas.
A seguir é apresentada uma lista de conceitos largamente aceitos atualmente.

2.5.1 CLASSIFICAÇÃO (APRENDIZADO SUPERVISIONADO)

Classificação provê um mapeamento prévio a partir de atributos para grupamentos especificados[GRO98]. É também denominado aprendizado supervisionado porque a entrada e a saída desejadas são fornecidas previamente por um supervisor externo[FAU94] . Por exemplo, pessoas podem ser previamente grupadas nas classificações de bebes, crianças, adolescentes, adultos, e idosos. Dois anos ou menos pode ser mapeado para a categoria bebê.

2.5.2 AGRUPAMENTO (APRENDIZADO NÃO SUPERVISIONADO)

Agrupamento (Clustering) ou segmentação, é um método no qual dados parecidos são grupados juntos[BER97]. No caso de reivindicações fraudulentas, os registros podem naturalmente ser separados em duas classes. Uma das classes pode corresponder a reivindicações normais e outra pode corresponder a reivindicações fraudulentas.

2.5.3 REGRESSÃO LINEAR

Regressão Linear é a técnica estatística para descobrir como dados de entrada, ou variáveis independentes, podem afetar uma certa saída, ou variável dependente[GRO98]. Quando a saída, ou classe, e todos os atributos são numéricos, a regressão linear é uma técnica natural a considerar. A idéia é expressar a classe como uma combinação linear dos atributos, com pesos predeterminados.

$x = w_0 + w_1a_1 + w_2a_2 + ... + w_ka_k$

Onde x é a classe, a_1, a_2, ... , a_k são os valores dos atributos, e w_0, w_1, ... w_k, são os pesos.

2.5.4 MODELAGEM

Modelagem é o processo de criar um modelo para representar um conjunto de dados[GRO98]. Um modelo em geral não representará um conjunto de dados com 100% de precisão. Pode ser criado um modelo que seja 100% preciso em alguns enfoques usando um conjunto de treinamento, mas se um modelo estiver sendo usado para previsão, pode ocorrer excesso de treinamento do modelo, tornando-o específico[GRO98]. Para casos futuros, o modelo pode ser menos preciso porque, com o tempo, tendências gerais são mais importantes que casos específicos.

Isto traz uma importante questão em MD:
O conhecimento derivado de um conjunto de treinamento será aplicável a outros dados não vistos durante o processo de treinamento?
O modelo criado com um conjunto de treinamento não preverá corretamente usando outro conjunto de dados em que as pessoas tenham hábitos diferentes [GRO98].

2.5.5 VISUALIZAÇÃO

Algumas vezes dados podem ser melhor compreendidos através de gráfico.
A mente humana está no seu melhor desempenho quando processando imagens. Desenhos podem transportar informação muito mais sucintamente que descrições textuais. Isto se aplica a mineração de dados também: ferramentas de visualização são extremamente úteis durante a preparação da entrada de dados para um esquema de aprendizado e quando tentando compreender seu resultado[WIT00].

Visualização da entrada

Visualização interativa é uma ferramenta poderosa para seleção de atributos e percepção rápida de padrões, o que em muitas situações não seria possível de outra maneira. Em muitos casos práticos de mineração há atributos demais para visualizar simultaneamente, e não há alternativa para algoritmos automáticos de seleção de atributos. Apesar disto, mesmo aqui as visualizações podem ser muito úteis para compreender melhor o que os algoritmos encontram. Além disto, podem prover pistas úteis sobre quais métodos de aprendizado são adequados para produzir bons resultados para os dados.

Visualização da saída

Visualizar a saída de um esquema de aprendizado pode ser igualmente útil. Alguns modelos se prestam naturalmente para uma representação gráfica. Assim podemos ter árvores de decisão representadas como grafos estruturados em árvore; tabelas de decisão tem uma representação natural como uma estrutura em grade bidimensional, cada célula representando uma entrada na tabela.

Na figura abaixo, a altura das barras representa a confiança , enquanto a cor representa o suporte. O plano horizontal pode ser movido pelo usuário para

esconder parcialmente todas as regras cuja confiança esteja abaixo de um certo limiar. O usuário também pode ter a altura mostrando o suporte e a cor a confiança.

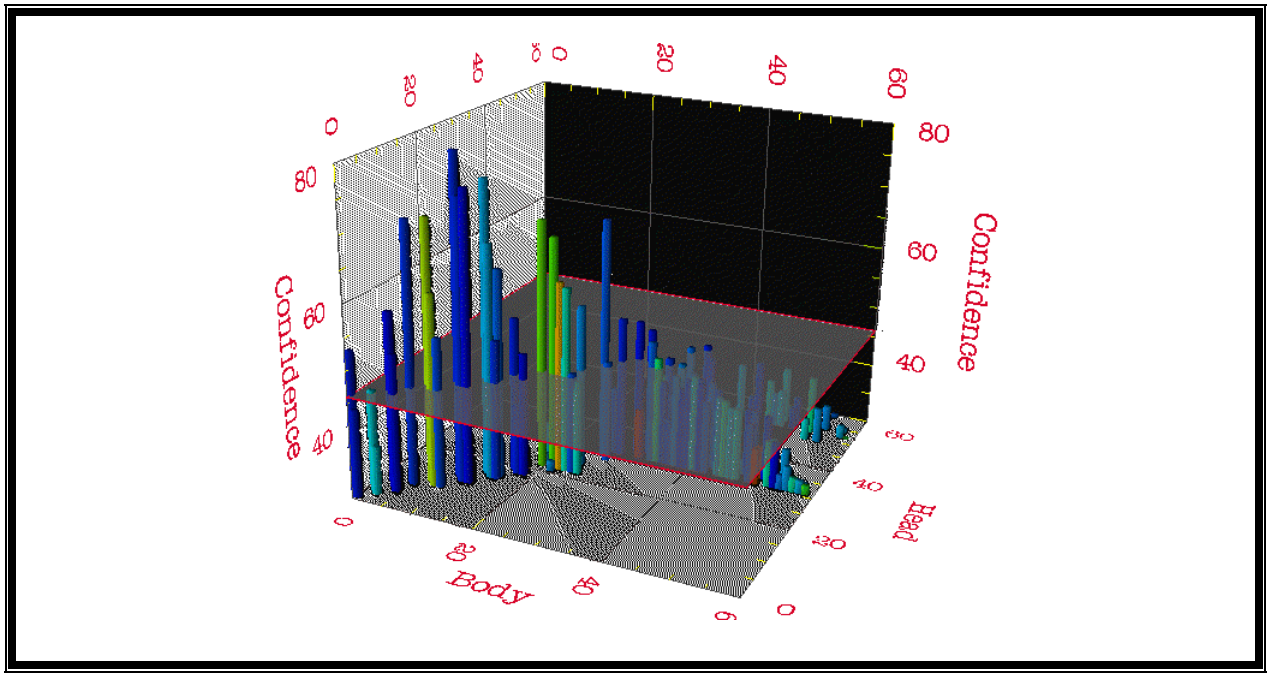

Figura 1: Visualização 3-D de Regras de dois itens (fonte projeto Quest da IBM)

A figura a seguir visualiza regras com dois itens no corpo e um na cabeça.

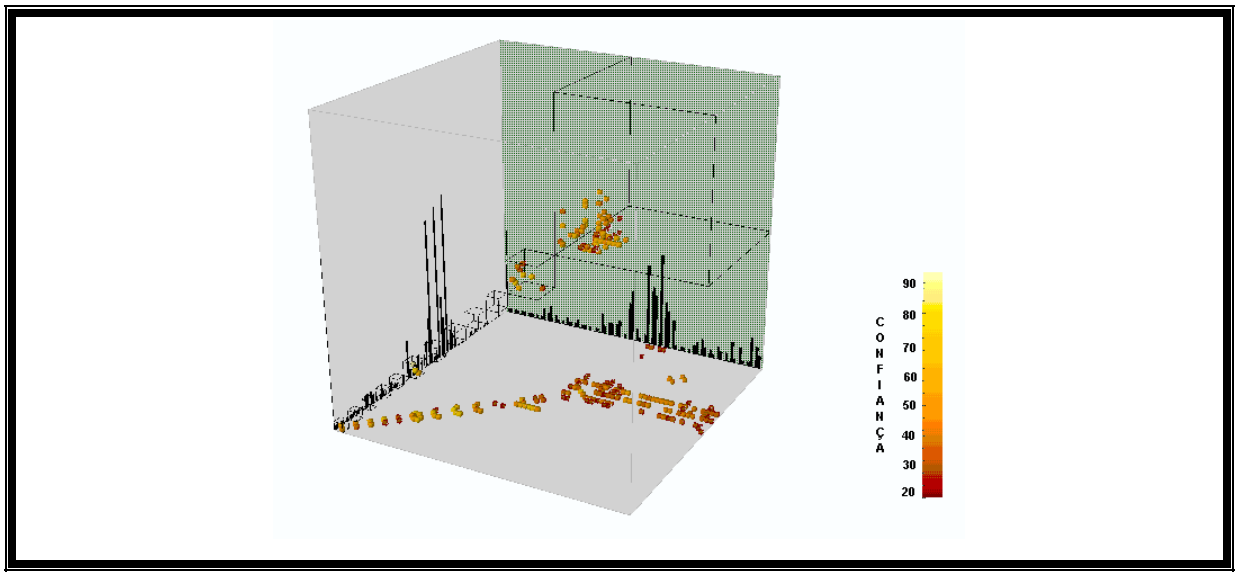

Figura 2: Visualização 3-D de Regras de três itens (fonte projeto Quest da IBM)

O plano do fundo mostra regras de dois itens como cubos. As caixas centrais mostram regras de três itens. Usuários podem “voar” por este espaço para explorar grupamentos, ou clicar numa caixa para grifar todas as regras contendo qualquer dos itens daquela caixa.

2.5.6 MODELAGEM PARA PREVISÃO

Um modelo pode ser usado com sucesso para prever resultados de eventos futuros. Embora dados históricos não possam profetizar o futuro, padrões tendem a se repetir, de forma que se um modelo representativo de um conjunto de dados pode ser construído, previsões podem ser feitas a partir deles[GRO98].

2.6 FATORES CRÍTICOS NUM SISTEMA DE MINERAÇÃO DE DADOS

Segundo [BER97] três medidas são os fatores mais críticos para construir um sistema de mineração de dados usável em ambiente de negócios:

- Automação – a técnica utilizada deve ser fácil de usar, o mais transparente possível, tão automatizada quanto for possível.
- Clareza – as respostas devem ser compreensíveis e fazer sentido.
- Retorno do Investimento – as respostas precisam ser conversíveis em análise de retorno do investimento.

2.7 MINERAÇÃO DE DADOS VERSUS ESTATÍSTICA

Uma idéia que surge naturalmente é questionar em que as técnicas de mineração de dados diferem das técnicas estatísticas. A estatística foi usada durante muitos anos para atingir objetivos de marketing. É preciso haver um forte motivo para

abandonar ferramentas estatísticas já consagradas e adotar outra baseada em tecnologia de mineração de dados.

De fato, a estatística apresenta os mesmos resultados que a mineração de dados. A regressão é usada em estatística para criar modelos de previsão do comportamento do cliente, a partir de longas séries históricas de dados armazenados. Um usuário final dificilmente tem domínio de estatística suficiente para trabalhar diretamente com a ferramenta de análise.

As ferramentas com a tecnologia de mineração de dados são mais fáceis de usar, viabilizando que o usuário final interaja diretamente com a ferramenta de análise.

A principal diferença entre mineração de dados e estatística então, é que a mineração de dados se destina a ser usada diretamente pelo usuário final, o gerente de marketing, enquanto a estatística requer um estatístico que, a partir da solicitação do usuário final monte o modelo[BER97].

Pelo exposto acima a mineração de dados é uma tecnologia que deve ser considerada.

2.8 MOTIVOS PARA USAR MINERAÇÃO DE DADOS

O melhor argumento sobre a utilidade da MD é o número de empresas que estão minerando dados e se negando a falar sobre isto[GRO98].

Algumas áreas nas quais a MD está sendo usada para benefício estratégico:

Marketing Direto: para reduzir despesa selecionando clientes potenciais melhor qualificados, reduzindo o número de correspondências enviadas pelo sistema de mala direta. Há casos de redução de vinte vezes no custo, em que baixou de um milhão de dólares para cinqüenta mil dólares.

Detecção de Fraude: para montar um modelo de quais reclamações de seguro, chamadas de telefone celular, ou compras com cartão de crédito parecem ser fraudulentas.

Previsão no Mercado Financeiro: para modelar o mercado de ações, usando redes neurais para obter ganhos financeiros.

2.9 MODELOS DE MINERAÇÃO DE DADOS

No coração da MD está o processo de construção do modelo para representar um conjunto de dados, característica comum a todos os produtos encontrados no mercado[GRO98]; o que não é comum a todos os produtos de MD é a maneira pela qual o modelo é construído. Para confundir esta situação, há centenas de enfoques derivados sob um rótulo genérico de nomes como redes neurais, redes de agentes, árvores de decisão. Um exemplo é um produto da NeuralWare que oferece mais de vinte e cinco diferentes enfoques de redes neurais. [GRO98]
As tecnologias disponíveis para mineração de dados podem ser grupadas nas seguintes vertentes principais:

- Árvores de Decisão
- Vizinho mais próximo e agrupamentos
- Algoritmos Genéticos
- Redes Neurais Artificiais
- Indução de Regras
- Agentes Inteligentes

A seguir serão descritas sucintamente as diversas tecnologias de modelagem citadas:

2.9.1 ÁRVORES DE DECISÃO

Uma árvore de decisão é um modelo de previsão(classificação) que pode ser visto como uma árvore. Cada ramo da árvore é uma questão classificatória, e as folhas da árvore são partições do conjunto de dados com suas classificações[BER97].

O maior benefício do enfoque de árvores de decisão é sua compreensibilidade; entretanto, para modelar dados com sucesso usando árvores de decisão, diversas quebras podem ser necessárias.
A partir de uma árvore de decisão podem ser obtidas regras caracterizadas por serem mutuamente exclusivas e coletivamente exaustivas.

Exemplos de algoritmos de árvore de decisão são: CART, CHAID, ID3, C4.5, SLIQ.

- CART – (Classification And Regression Trees), é um algoritmo de exploração e previsão de dados que usa métricas de entropia para escolher ramificações ótimas[BER97]. Leo Breiman, 1984.
- CHAID – (Chi Square Automatic Interaction Detection), é similar ao CART no fato de construir uma árvore de decisão, mas difere no modo como escolhe suas ramificações, pelo teste do chi quadrado em tabelas de contingência[BER97].
- ID3 - é um algoritmo que classifica os objetos testando suas propriedades[BER97]. Foi proposto por J Ross Quinlan, em 1978.
- C4.5 – surgiu como uma evolução do ID3, introduzindo ausência de valores, a poda e a derivação de regras[BER97]. Quinlan, 1993.
- SLIQ – (Supervised Learning In Quest) desenvolvido pela equipe do projeto Quest da IBM(1995). Usa pré-classificação na fase de crescimento da árvore. Calcula valores de entropia de todos os nós. Posteriormente(1996) a equipe desenvolveu um novo algoritmo SPRINT(**S**calable **P**a**R**allelizable **IN**duction of Decision **T**rees) que resolveu as restrições de memória existentes no SLIQ.

Os algoritmos ID3, C4.5 e SLIQ tiveram origem na inteligência artificial, enquanto o CHAID originou-se na estatística. O CART é um híbrido entre IA e estatística[BER97].

2.9.2 VIZINHO MAIS PRÓXIMO E AGRUPAMENTO

Segundo [BER97] as técnicas de previsão e classificação vizinho mais próximo e agrupamento estão entre as mais antigas técnicas usadas em mineração de dados. Agrupamento é um método em que dados parecidos são grupados juntos.
A técnica do vizinho mais próximo está entre as técnicas mais fáceis de usar e compreender porque trabalha de um modo similar ao modo como as pessoas pensam – detectando exemplos semelhantes.
A principal diferença entre as duas técnicas é que o agrupamento é uma técnica de aprendizado não supervisionado, enquanto vizinho mais próximo é de aprendizado supervisionado.

2.9.3 ALGORITMOS GENÉTICOS

Algoritmos genéticos são métodos de otimização combinatorial baseados em processos da evolução biológica[GRO98], simulam combinação genética, mutação e seleção dos melhor adaptados. A idéia é que através do tempo, a evolução selecionou espécies melhor adaptadas.
Segundo [BER97], os AG tem sido aplicados junto de outras técnicas de mineração de dados, tais como redes neurais, para encontrar os pesos ótimos das ligações, ou a técnica do vizinho mais próximo, para encontrar os pesos a serem aplicados a cada previsor. Até o momento não foi demonstrado que o uso de AG provê soluções mais rápidas ou melhores do que os algoritmos de busca específicos de cada técnica de mineração de dados.

2.9.4 REDES NEURAIS

Em 1943, McCulloch e Pitts apresentaram uma unidade lógica de limiar, criando o conceito de neurônio artificial, que faz uma mímica do processo de um neurônio no cérebro humano.[GRO98]

Em 1982 , John Hopfield mostrou como redes neurais poderiam ser usadas para fins computacionais. Em 1984, Teuvo Kohonen introduziu um novo algoritmo que ele denominou Organizing Feature Map, Mapa Organizador de Característica, que possibilitou usar redes neurais para aprendizado não supervisionado . Isto abriu um novo ramo de pesquisa de redes neurais onde uma resposta **correta** não é requerida para aprender ou treinar uma rede[BER97].

As RNA imitam a capacidade do cérebro humano de identificar padrões, sendo considerados os mais complicados algoritmos de classificação e regressão, mas são muito úteis como modelos de previsão, por exemplo na detecção de fraudes em tempo real. Uma rede treinada com milhões de transações, incluindo algumas que sabidamente foram fraudulentas, forma um modelo que permite classificar transações em boas e más[BER97].

Por não disporem de um componente descritivo, fica difícil compreender as escolhas feitas pela RNA, sendo freqüentemente referida como tecnologia caixa preta. Uma diferença fundamental entre redes neurais e outras técnicas é o fato de as RNA somente operarem diretamente sobre números. Embora a aproximação conseguida com outras tecnologias seja suficientemente boa, quando a precisão é importante, a rede neural é a melhor opção. Mais freqüentemente é usado o algoritmo de retropropagação, sendo usados também redes de Kohonen e RBF (função de base radial)[BER97].

2.9.5 INDUÇÃO DE REGRAS

A indução de regras é uma das principais formas de mineração de dados e possivelmente a forma mais comum de descobrimento de conhecimento em sistemas de aprendizado não supervisionado[BER97].

Sistemas de indução de regras são altamente automatizados, sendo provavelmente as melhores técnicas para expor todos os padrões previsíveis implícitos numa base de dados.

As regras produzidas por um sistema de indução de regras não são mutuamente exclusivas, podendo ser coletivamente exaustivas ou não. A diferença para as

regras obtidas a partir de árvores de decisão é que estas são mutuamente exclusivas e coletivamente exaustivas[BER97].

Comparando técnicas de mineração de dados com ênfase na capacidade de explicação, as RNA ficariam num extremo enquanto os sistemas de indução de regras ficariam no outro. As RNA são extremamente competentes para dizer exatamente o que deve ser feito numa tarefa de previsão, por exemplo, a quem conceder crédito, com pouca ou nenhuma explicação. Sistemas de indução de regras, ao contrário, são como um comitê de consultores, cada um com uma leve diferença de opinião sobre o que fazer, mas com relativamente bem fundamentadas razões e uma boa explicação[BER97].

2.9.6 REDE DE AGENTES

Esta tecnologia foi desenvolvida por Dr Khai Min Pham, na França em 1990, sendo descrita como um enfoque polimórfico híbrido, significando que usa características de diferentes algoritmos, dependendo de como é usada.

Este método de construção de modelo trata todos os elementos de dado, ou categorias de elementos de dado definidos, como agentes que são conectados um ao outro de maneira significativa.[GRO98]

2.10 PREPARAÇÃO PARA MINERAR DADOS

O processo de MD foi descrito anteriormente como um processo de construção de modelo. Construindo um modelo de um conjunto de dados, os dados podem ser compreendidos de maneiras que não tenham sido previamente consideradas.

Embora diferentes literaturas descrevam o processo de MD de maneiras diversas, há cinco passos principais para MD[GRO98]:

.Preparação de dados

.Definição de um estudo

.Leitura dos dados e construção de um modelo

.Compreensão do modelo

.Previsão
Cada um destes passos será descrito a seguir:

2.10.1 PREPARAÇÃO DE DADOS

MD não é um processo mágico que parte de dado cru e destila informações valiosas. A preparação dos dados está no coração deste processo.

2.10.1.1 LIMPEZA DE DADOS

Dado não é sempre limpo. Pepsi não é igual a Pepsi Cola. Os valores se referem à mesma bebida, mas são dados diferentes para o programa de computador.
Outra questão de limpeza se refere a dados desatualizados. Listas de correspondência precisam ser continuamente atualizadas porque as pessoas se mudam e seus endereços se alteram.
Erros tipográficos também são uma questão de limpeza. Palavras são digitadas incorretamente.

2.10.1.2 DADOS AUSENTES

É comum que num arquivo contendo registros de fatos ocorridos durante um longo período de tempo, alguns campos não estejam preenchidos, por diversos motivos, como indisponibilidade do dado na ocasião do preenchimento e falha na revisão. É necessário tratar este problema completando, se possível, os campos não preenchidos, porque quanto mais valores não preenchidos, menor a chance de obter resultados úteis numa MD.

2.10.1.3 DERIVAÇÃO DE DADOS

Quando um dado necessário para um estudo não existe, mas pode ser obtido pela combinação ou transformação de outros disponíveis, diz-se que o dado pode ser derivado. Um exemplo típico é quando se dispõe da data de nascimento de um cliente, mas se necessita sua idade na data de cada compra. A idade pode ser derivada a partir das duas datas citadas.

2.10.1.4 MISTURA DE DADOS

Os dados necessários para uma MD podem estar em várias fontes diferentes, tornando necessário colocá-los numa tabela bidimensional. Geralmente os dados históricos disponíveis estão armazenados em mídia digital, em diversos arquivos, nos formatos em que foram processados pelos aplicativos do dia a dia das empresas. Mesmo se estiverem armazenados em data warehouse, estarão em diversas tabelas normalizadas, tornando necessário misturar dados de diversas tabelas para gerar uma única na forma aceita pelo programa de MD.

2.10.2 DEFINIÇÃO DE UM ESTUDO

Definir um estudo difere para aprendizado supervisionado e não supervisionado
Para aprendizado supervisionado, definir um estudo envolve estabelecer um objetivo, escolhendo uma variável dependente/saída que caracteriza um aspecto daquele objetivo, e especificando os campos que são usados no estudo[GRO98].
Para aprendizado não supervisionado, o objetivo geral é grupar tipos similares de dados ou identificar exceções num conjunto de dados[GRO98].
Existem problemas menos importantes relevantes a todos os estudos. Inicialmente, definir estudos envolve especificar um uso para os conjuntos de dados. Um conjunto de dados pode ser usado para construir um modelo, outro para validar a correção do modelo, e ainda outro para fazer previsões usando o modelo.

Outro problema na definição de um estudo é o tamanho da amostra. Não é sempre necessário minerar um conjunto de dados inteiro. Pode ser escolhido um subconjunto de linhas (registros) por amostragem aleatória.
Determinar o número de linhas necessário para representar com precisão o conjunto inteiro impõe diversos desafios.

2.10.3 LEITURA DOS DADOS E CONSTRUÇÃO DE UM MODELO

Uma vez definido um estudo, um produto de MD lê um conjunto de dados e constrói um modelo. Enquanto todos os modelos variam, o conceito básico é o mesmo. Um modelo sumariza grandes quantidades de dados pela acumulação de indicadores[GRO98]. Alguns dos indicadores que vários modelos acumulam são:
Freqüência: mostram com que freqüência um certo valor ocorre.
Pesos: ou impactos, indicam quão bem algumas entradas indicam a ocorrência de uma saída[GRO98].
Conjunção: algumas vezes entradas tem mais peso juntas do que separadas. Por exemplo, pode não ser verdade que homens sejam clientes fiéis, mas pode ser que homens donos de cachorro e praticantes de ciclismo sejam clientes fiéis[GRO98].

2.11 ALGORITMO APRIORI

A seguir será apresentado o algoritmo **APRIORI** conforme descrito em [AIS93].

2.11.1 ABSTRATO

Seja dada uma base de dados de transações, onde cada transação consiste de itens adquiridos por um cliente numa visita. A seguir é apresentado um algoritmo que gera todas a regras significativas de associação entre itens na base de dados. O algoritmo incorpora técnicas de estimativa e poda.

2.11.2 MODELO FORMAL

Seja $\Gamma = I_1, I_2, I_3, ... I_m$ um conjunto de atributos binários, denominados itens.

Seja T uma base de dados de transações. Cada transação binária t é representada como um vetor binário, com t[k] = 1 se t adquiriu o item I_k , e t[k] = 0 caso contrário. Existe uma tupla na base de dados para cada transação.

Seja X um conjunto de alguns itens em Γ. Diz-se que uma transação t satisfaz X se para todos os itens I_k em X, t[k] = 1.

Uma regra de associação deve ser entendida como uma implicação da forma $X \Rightarrow I_j$ onde X é um conjunto de alguns itens em Γ, e I_j é um item individual em Γ, que não está presente em X. A regra $X \Rightarrow I_j$ é satisfeita no conjunto de transações T com o fator de confidência $0 \leq c \leq 1$ se ao menos c% das transações em T que satisfazem X também satisfazem I_j . Será usada a notação $X \Rightarrow I_j \mid c$ para especificar que a regra $X \Rightarrow I_j$ tem um fator de confiança c .

Dado o conjunto de transações T, há o interesse em gerar todas as regras que satisfaçam certas restrições de duas diferentes formas:

1. Restrições Sintáticas: Estas restrições envolvem restrições sobre itens que podem aparecer numa regra. Por exemplo, pode haver interesse apenas em regras que tenham um item I_x específico aparecendo no conseqüente, ou regras que tenham um item I_y aparecendo no antecedente. Combinações das restrições acima também são possíveis. Podem ser solicitadas todas as regras que tenham itens de algum conjunto X de itens predefinido aparecendo no conseqüente, e itens de algum outro conjunto Y aparecendo no antecedente.
2. Restrições de Suporte: Estas restrições se referem ao número de transações em T que suportam a regra. O suporte para uma regra é definido como a fração de transações em T que satisfazem a união de itens no conseqüente e antecedente da regra. Suporte não deve ser confundido com confiança. Enquanto confiança é uma medida da força da regra, suporte corresponde à significância estatística. Além da significância estatística, outra motivação para restrições de suporte vem

do fato de usualmente haver interesse apenas em regras com suporte acima de algum limiar mínimo por razões do negócio. Se o suporte não é extenso o suficiente, significa que a regra não merece consideração ou que é simplesmente menos preferida.

Nesta formulação, o problema de mineração de regras pode ser decomposto em dois problemas:

1. Gerar todas as combinações de itens que tenham suporte a transação fracionária acima de um certo limiar denominado minsuporte. Denominando estas combinações extenso conjunto de itens, e todas as outras combinações, que não atingem o limiar, de pequeno conjunto de itens. Restrições sintáticas posteriormente restringem as combinações admissíveis. Por exemplo, se apenas regras envolvendo um item I_x no antecedente são de interesse, então é suficiente gerar apenas aquelas combinações que contém I_x.
2. Para um dado extenso conjunto de itens $Y = I_1\ I_2\ I_3\ ...\ I_k$, $k \geq 2$, gerar todas as regras (no máximo k regras) que usam itens do conjunto I_1, I_2, I_3, ..., I_k, .O antecedente de cada uma destas regras será um subconjunto X de Y tal que X tem k – 1 itens, e o conseqüente será o item Y – X. Para gerar uma regra $X \Rightarrow I_j \mid c$, onde $X = I_1\ I_2\ ...\ I_{j-1}\ I_{j+1}\ ...\ I_k$, toma o suporte de Y e divide-o pelo suporte de X. Se a razão é maior que c então a regra é satisfeita com fator de confiança c; caso contrário não. Note que se o conjunto de itens Y é extenso, então todo subconjunto de Y será também extenso, e deve haver disponível seus contadores de suporte como o resultado da solução do primeiro subproblema. Também, todas as regras derivadas de Y devem satisfazer a restrição de suporte porque Y satisfaz a restrição de suporte e Y é a união dos itens no conseqüente e antecedente de toda regra.

Havendo determinado os extensos conjuntos de itens, a solução para o segundo subproblema é ao contrário direta.

2.11.3 DESCOBRINDO EXTENSOS CONJUNTOS DE ITENS

Dado um conjunto de itens Γ, um conjunto de itens X + Y de itens em Γ é dito ser uma extensão do conjunto de itens X se X Y = 0. O parâmetro dbsize é o número total de tuplas na base de dados.

O algoritmo faz múltiplas passadas sobre a base de dados. O conjunto fronteira para uma passada consiste daqueles conjuntos de itens que são estendidos durante a passada. Em cada passada, o suporte para certos conjuntos de itens é medido. Estes conjuntos de itens, chamados conjuntos de itens candidatos, são derivados das tuplas na base de dados e dos conjuntos de itens contidos no conjunto fronteira.

Associado a cada conjunto de itens há um contador que armazena o número de transações no qual o correspondente conjunto de itens apareceu. Este contador é inicializado com zero quando um conjunto de itens é criado.

Inicialmente o conjunto fronteira consiste de apenas um elemento, o qual é um conjunto vazio. Ao fim da passada, o suporte para um conjunto de itens candidato é comparado com minsuporte para determinar se ele é um extenso conjunto de itens. Ao mesmo tempo, é determinado se este conjunto de itens poderia ser acrescentado ao conjunto fronteira para a próxima passada. O algoritmo termina quando o conjunto fronteira se torna vazio. A contagem de suporte para o conjunto de itens é preservada quando um conjunto de itens é acrescentado ao conjunto extenso/fronteira.

A seguir é apresentado o algoritmo **APRIORI** em linguagem estruturada.

Notação:

k	Número da passagem sobre os dados
k-itemset	Um itemset tendo k itens
L_k	Conjunto de extensos k-itemsets (aqueles com suporte mínimo)
C_k	Conjunto de k-itemsets candidatos (potencialmente extensos itemsets). Cada membro deste conjunto tem dois campos: i) itemset e ii)contagem do suporte.

Algoritmo **APRIORI**

1) L_1 = {extensos 1-itemsets};
2) **for** (k = 2; L_{k-1} <> 0; k++) **do begin** // k representa o número da passagem
3) C_k = apriori-gen(L_{k-1}); // Novos candidatos de tamanho k gerados de L_{k-1}
4) **forall** transações t ∈ D **do begin**
5) C_t = subset (C_k , t); // Adicionar a t todos os ancestrais de cada
// item em t, removendo qualquer duplicata
6) **forall** candidatos c ∈ C_t **do** **//** Incrementar o contador de todos os
7) c.count++; // candidatos em C_k que estão contidos em t
8) ...**end**
9) L_k = { c ∈ C_k | c.count >= minsup } // Todos os candidatos em C_k com suporte mínimo
10) **end**
11) Resposta = $\Upsilon_k L_k$

Geração de candidato apriori:

Function apriori-gen(L_{k-1});

1) **insert into** C_k
2) **select** p[1],p[2],...,p[k-1],q[k-1]
3) **from** L_{k-1} p, L_{k-1} q
4) **where** p[1] = q[1],...,p[k-2] = q[k-2],p[k-1] < q[k-1];

2.11.3.1 NÚMERO DE PASSADAS VERSUS MEDIDAS DE DESPERDÍCIO

Na versão mais direta do algoritmo, todo conjunto de itens presente em alguma das tuplas será medido em uma passada, terminando o algoritmo em uma passada. No pior caso, este enfoque requererá a inicialização de 2^m contadores correspondendo a todos os subconjuntos do conjunto de itens Γ, onde m é o número de itens em Γ. Isto é, de fato, não apenas impraticável (m pode ser facilmente mais de 1000 num supermercado) mas também desnecessário. De fato, haverá conjuntos de itens menores contendo mais que *I* itens, onde *I* é

pequeno. Além disso, diversas dessas 2^m combinações cairão fora tornando-se pequena de alguma forma.

Um melhor enfoque é medir na k-ésima passada apenas aqueles conjuntos de itens que contém exatamente k itens. Tendo medido alguns conjuntos de itens na k-ésima passada, é necessário medir na (k+1)-ésima passada apenas aqueles conjuntos de itens que são extensão-um (um conjunto de itens estendidos de exatamente um item) dos extensos conjuntos de itens encontrados na k-ésima passada. Se um conjunto de itens é pequeno, sua extensão-um também será pequena. Portanto, o conjunto fronteira para a próxima passada se torna o conjunto de itens candidato identificados como extensos na atual passada, e apenas os extensão-um de um conjunto fronteira são gerados e medidos durante uma passada. Esta alternativa representa outro extremo – serão feitas muitas passadas sobre a base de dados.

Estes dois enfoques extremos ilustram o balanço entre número de passadas e esforço desperdiçado devido a medida de conjunto de itens que se tornam pequenos. Alguma medida de desperdício é inevitável – se o conjunto de itens A é extenso, AB deve ser medido para determinar se é extenso ou pequeno. Entretanto, tendo determinado que AB é pequeno, é desnecessário medir ABC, ABD, ABCD, e assim em diante. Portanto, além da praticidade, se for medido um grande número de conjuntos de itens candidato numa passada, vários deles podem se tornar pequenos de qualquer maneira – esforço desperdiçado. Por outro lado, se for medido um pequeno número de candidatos e vários deles se tornarem extensos, então será necessária uma outra passada, que poderia não ter sido necessária. Por isso, é necessário alguma cuidadosa estimativa antes de decidir se um conjunto de itens candidato deve ser medido numa dada passada.

2.11.3.2 DETERMINAÇÃO DE CONJUNTOS DE ITENS CANDIDATOS

Pode-se pensar que se possa medir na passada atual apenas aquelas extensões do conjunto de itens fronteira que se espera sejam extensos. Entretanto se fosse o caso e os dados estivessem de acordo com as expectativas e os conjuntos de

itens que se esperava serem extensos de fato se tornam extensos, então pode-se ainda necessitar outra passada sobre a base de dados para determinar o suporte das extensões daqueles extensos conjuntos de itens. Para eliminar esta situação, em adição aquelas extensões do conjunto de itens fronteira que são esperados ser extensos, também são medidas as extensões X + l_j que são esperadas ser pequenas mas tais que X é esperado ser extenso e X contém um conjunto de itens fronteira. Entretanto, não são medidas quaisquer posteriores extensões de tais conjuntos de itens. O racional para esta escolha é que se as previsões são corretas e X + l_j de fato vem a ser pequeno, então nenhum super conjunto de X + l_j precisa ser medido. A passada adicional é então necessária apenas se os dados não estiverem de acordo com as expectativas e X + l_j vier a ser extenso. Esta é a razão porque não medindo os X + l_j que são esperados ser pequenos seria um engano – mesmo que os dados concordem com as previsões, uma passada extra sobre a base de dados seria necessária.

2.11.3.3 SUPORTE ESPERADO PARA UM CONJUNTO DE ITENS

Usa-se a suposição da independência estatística para estimar o suporte para um conjunto de itens. Suponhamos que um conjunto de itens candidato X + Y é uma extensão-k do conjunto de itens fronteira X e que Y = l_1 l_2 l_3 ... l_k. Suponhamos que o conjunto de itens X apareça num total de x tuplas. Sabe-se o valor de x porque X foi medido na passada prévia (x é considerado ser dbsize para o conjunto vazio de itens fronteira). Suponhamos que X + Y está sendo considerado como um conjunto de itens candidato para a primeira vez após c tuplas contendo X terem já sido processadas na passada atual. Denotando por f(l_j) a freqüência relativa do item l_j na base de dados, o suporte esperado s??? para o conjunto de itens X + Y é dado por

$$s = f(l_1) * f(l_2) * ... * f(l_k) * (x - c)/dbsize$$

Note-se que (x – c)/dbsize é o suporte atual para X na porção restante da base de dados. Sob suposição da independência estatística, o suporte esperado para X +

Y é um produto do suporte para X e freqüências relativas individuais de itens em Y.
Se s é menor que minsuporte então se diz que X + Y é esperado ser pequeno; caso contrário, é esperado ser extenso.

2.11.3.4 PROCEDIMENTO DE GERAÇÃO DE CONJUNTOS DE ITENS CANDIDATOS

Um conjunto de itens que não esteja presente em nenhuma tupla na base de dados nunca se torna um candidato para medição. É lida uma tupla por vez na base de dados e verificado quais conjuntos fronteira estão contidos na tupla lida. Conjuntos de itens candidato são gerados a partir destes conjuntos de itens fronteira estendendo-os recursivamente com outros itens presentes na tupla. Um conjunto de itens que é esperado ser pequeno não é estendido posteriormente.

Para não replicar diferentes maneiras de construir o mesmo conjunto de itens, os itens são ordenados e um conjunto de itens X é tentado para extensão apenas por itens que são posteriores na ordenação em relação a qualquer dos membros de X. Por exemplo, seja Γ = {A,B,C,D,E,F} e assumamos que os itens estão ordenados em ordem alfabética. Além disso assumamos que o conjunto fronteira contém apenas um conjunto de itens, AB. Para a tupla t = ABCDF da base de dados, os seguintes conjuntos de itens candidatos são gerados:

ABC esperado extenso: continua estendendo
ABCD esperado pequeno: não estende posteriormente
ABCF esperado extenso: não pode ser estendido posteriormente
ABD esperado pequeno: não estende posteriormente
ABF esperado extenso: não pode ser estendido posteriormente

A extensão ABCDF não foi considerada porque ABCD foi esperado ser pequeno. Similarmente ABDF não foi considerada porque ABD foi esperado ser pequeno.

Os conjuntos de itens ABCF e ABF, embora esperado serem extensos, não poderiam ser estendidos porque não há item em t que seja maior que F. As extensões ABCE e ABE não foram consideradas porque o item E não está em t.

2.11.3.5 DETERMINAÇÃO DO CONJUNTO FRONTEIRA

Decidir quais conjuntos de itens por no próximo conjunto fronteira se mostra ser algo engenhoso. Pode-se pensar que é suficiente selecionar apenas máximos (em termos de inclusão no conjunto) extensos conjuntos de itens. Esta escolha, entretanto é incorreta – pode resultar que o algoritmo desapareça com alguns extensos conjuntos de itens como o seguinte exemplo ilustra:

Suponhamos que o conjunto fronteira AB seja estendido como mostrado na subseção anterior. Entretanto, ambos ABD e ABCD se tornaram extensos ao fim da passada. Então ABD como um extenso conjunto de itens não máximo, não poderia se tornar fronteira – um engano, dado que ABDF não seria considerado, o qual poderia ser extenso, e ficaria incompleto.

São incluídos no conjunto fronteira para a próxima passada aqueles conjuntos de itens candidatos que eram esperados ser pequenos mas se tornaram extensos na atual passada. Para ver que estes são os únicos conjuntos de itens que é necessário incluir no próximo conjunto fronteira, primeiro é preciso estabelecer o seguinte lema:

Lema. Se o conjunto de itens candidato X é esperado ser pequeno na atual passada sobre a base de dados, então nenhuma extensão $X + l_j$ de X, onde $l_j > l_k$ para qualquer l_k em X é um conjunto de itens candidato nesta passada.

O lema funciona devido ao procedimento de geração de conjunto de itens candidato.

Consequentemente, sabe-se que nenhuma extensão do conjunto de itens que está sendo incluído no próximo conjunto fronteira foi considerado na passada atual. Mas dado que estes conjuntos de itens são extensos atualmente, eles podem ainda produzir extensões que sejam extensas. Portanto, estes conjuntos de itens devem ser incluídos no conjunto fronteira para a próxima passada. Eles não levam a qualquer redundância porque nenhuma de suas extensões foi medida tão longe. Adicionalmente, também fica completo. Além disto, se um conjunto de itens candidato era extenso mas não era esperado ser pequeno então não poderia estar no conjunto fronteira para a próxima passada porque, pelo modo que o algoritmo é definido, todas as extensões deste conjunto de itens já tinham sido consideradas nesta passada. Um conjunto de itens candidato que é pequeno não deve ser incluído no próximo conjunto fronteira porque o suporte para uma extensão de um conjunto de itens não pode ser maior que o suporte para o conjunto de itens.

2.11.4 GERENCIAMENTO DE MEMÓRIA

O computador disponível pode não dispor de memória suficiente para armazenar todos os conjuntos de itens fronteira e candidato numa passada. Os conjuntos de itens extensos não precisam estar na memória durante uma passada sobre a base de dados e podem ser residentes em disco. Podem ser desenvolvidas rotinas específicas para administrar o uso da memória disponível, para viabilizar a mineração que de outra forma não poderia ser realizada. Foge ao objetivo deste trabalho tratar este assunto. Ver [AGR93].

2.11.5 PODA BASEADA NA CONTAGEM DE TUPLAS RESTANTES NA PASSADA

É possível durante uma passada determinar que um conjunto de itens candidato eventualmente não virá a ser extenso, e portanto descartá-lo cedo. Esta poda

poupa ambos memória e esforço de medição. Esta poda pode ser referida como *otimização por tuplas restantes*.

Suponhamos que um conjunto de itens candidato X + Y é uma extensão do conjunto de itens fronteira X e que o conjunto de itens X aparece num total de x tuplas (como discutido na seção 4.3.2, x é sempre conhecido). Suponhamos que X + Y está presente na c-ésima tupla contendo X. No momento de processar esta tupla, faça-se o contador de tuplas (incluindo esta tupla) contendo X + Y ser s.

Isto significa que restam no máximo x – c tuplas nas quais X + Y podem aparecer. Então compara-se maxconta = (x – c + s) com minsuporte * dbsize. Se maxconta for menor, então X + Y está limitado a ser pequeno e pode ser podado fora.

A *otimização por tuplas restantes* é aplicada assim que um novo conjunto de itens candidato é gerado, e pode resultar em imediata poda de alguns destes conjuntos de itens. É possível que um conjunto de itens candidato não seja inicialmente podado, mas ele pode satisfazer as condições de poda após mais algumas tuplas serem processadas. Para podar tais conjuntos de itens candidato “velhos”, aplica-se o teste de poda sempre que uma tupla contendo tal conjunto de itens é processada e estivermos para incrementar o contador de suporte para este conjunto de itens.

2.11.6 PODA BASEADA EM FUNÇÕES SINTETIZADAS DE PODA

Outra técnica para podar um conjunto de itens candidato logo que ele é gerado pode ser referida como *poda por função de otimização.*

A poda por função de otimização é motivada por funções de poda como *preço total da transação*. Preço total da transação é uma função cumulativa que pode ser associada com um conjunto de itens como uma soma de preços de itens individuais no conjunto. Se for conhecido que há menos do que a fração minsuporte de transações que compram mais que ë d ólares em mercadoria de itens, podemos imediatamente eliminar todos o conjuntos de itens para os quais

seu preço total excede ë. Tais conjuntos de itens não têm de ser medidos e incluídos no conjunto de itens candidatos.

Em geral, não se sabe o que essas funções de poda são. A partir dos dados disponíveis, funções de poda devem ser sintetizadas, tendo a forma

$$w_1 l_{j1} + w_2 l_{j2} + \ldots + w_m l_{jm} \leq ë$$

onde cada binário valorado $\in \Gamma$. Pesos w_i são selecionados como segue. Primeiro os itens individuais devem ser ordenados em ordem decrescente de suas freqüências de ocorrência na base de dados. Então o peso do i-ésimo item l_{ji} nesta ordem

$$w_i = 2^{i-1} å$$

onde å é um pequeno número real tal como 0.000001. Pode ser mostrado que sob certas suaves suposições uma função de poda com os pesos acima terá ótimo valor de poda – ela podará o maior número de conjunto de itens candidato.

Uma função de poda deve ser sintetizada a parte para cada conjunto de itens fronteira. Estas funções diferem em seus valores para ë. Uma vez que o suporte da transação para o item XY não pode ser maior que o suporte para o conjunto de itens X, a função de poda associada com o conjunto fronteira X pode ser usada para determinar se uma expansão de X poderia ser adicionada ao conjunto de itens candidato ou se poderia ser podada direto. Fazendo z(t) representar o valor da expressão

$$w_1 l_{j1} + w_2 l_{j2} + \ldots + w_m l_{jm}$$

para a tupla t. Dado um conjunto de itens fronteira X, é necessário um procedimento para estabelecer $ë_x$ tal que count(t | tupla t contém X e $z(t) > ë_x$) < minsuporte.

Havendo determinado o conjunto de itens fronteira numa passada, não se quer fazer uma passada separada sobre os dados só para determinar as funções de poda. Pode-se coletar informação para determinar ë para um conjunto de itens X enquanto X é ainda um conjunto de itens candidato e está sendo medido em antecipação que X pode se tornar um conjunto de itens fronteira na próxima passada. Afortunadamente, sabe-se que somente os conjuntos de itens candidato que são esperados ser pequenos são os únicos que podem se tornar um conjunto fronteira. É necessário coletar informação ë apenas para estes conjuntos de itens e não para todos os conjuntos de itens candidato.

Um procedimento direto para determinar ë para um conjunto de itens X será manter o número minsuporte de valores maiores de z para tuplas contendo X.

Esta informação pode ser coletada ao mesmo tempo que o contador de suporte para X está sendo medido numa passada. Este procedimento requererá memória para manter o número minsuporte com cada conjunto de itens candidato que é esperado ser pequeno. É possível salvar memória ao custo de perder alguma precisão (estabelecendo um valor um pouco maior para ë). Finalmente, deve ser lembrado que, como discutido na seção 2.9.4 quando a memória é limitada, um conjunto de itens candidato cujos filhos são eliminados na passada atual também se tornam um conjunto de itens fronteira. Em geral, filhos de um conjunto de itens candidato são eliminados no meio de uma passada, e não se deve colecionar informação ë para tal conjunto de itens. Tais conjuntos de itens herdam o valor ë de seus pais quando eles se tornam fronteira.

2.11.7 EXEMPLO DO ALGORITMO APRIORI

Considerando uma base de dados contendo os itens adquiridos pelos clientes identificados como 100, 200, 300, 400, conforme a tabela a seguir:

Cliente	Itens
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

2.11.7.1 IDENTIFICAÇÃO DOS CONJUNTOS MAIS FREQÜENTES

Teremos formalmente, os conjuntos de itemsets ocorridos em cada transação, representados como na tabela a seguir:

Cliente	Conjunto de Itemsets
100	{ {1}, {3}, {4} }
200	{ {2}, {3}, {5} }
300	{ {1}, {2}, {3}, {5} }
400	{ {2}, {5} }

Pressupondo um suporte mínimo de duas transações, obtém-se a tabela a seguir contendo os ítemsets a serem considerados, cada um com seu suporte: (1)

Itemsets	Suporte
1	2
2	3
3	3
5	3

Combinando os itemsets dois a dois, obtém-se os itemsets cadidatos com seu suporte:

Itemset	Suporte
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

Considerando que apenas os itemsets com suporte mínimo 2 interessam, obtém-se para cada transação quais associações ocorreram:

Cliente	Itemset
100	{1 3}
200	{2 3}, {2 5}, {3,5}
300	{1 2}, {1 3}, {1 5}, {2 3}, {2 5}, {3 5}
400	{2 5}

Os itemsets de interesse tornam-se então: (2)

Itemset	Suporte
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

Em mais uma iteração obtém-se as combinações três a três ocorridas, com o suporte:

Itemset	Suporte
{2 3 5}	2

Considerando que apenas os itemsets com suporte mínimo 2 interessam, obtém-se para cada transação quais associações ocorreram:

Cliente	Conjunto de Itemsets
200	{ {2 3 5} }
300	{ {2 3 5} }

Os itemsets de interesse tornam-se então: (3)

Itemset	Suporte
{2 3 5}	2

Uma tentativa de executar mais uma iteração resultaria num conjunto candidato vazio, o que leva ao encerramento das iterações.

2.12 GERAÇÃO DE REGRAS DE ASSOCIAÇÃO

As regras de associação consideradas a seguir são um pouco mais gerais que em [AIS93] no aspecto de permitirem o conseqüente ter mais de um item.

Para todo conjunto de itens extenso l, resultam todas as regras a => (l – a), onde a é um subconjunto de l, tal que a razão suporte(l)/suporte(a) é no mínimo minconf.

O suporte de qualquer subconjunto ã de a deve ser tão grande quanto o suporte de a. Então a confiança da regra ã => (l - ã) não pode ser maior que a confiança de a => (l - a). Portanto, se a não produz uma regra envolvendo todos os itens em l com a como antecedente, ã também não produzirá. Segue que para uma regra (l - a) => a valer, todas as regras da forma (l - ã) => ã devem valer também, onde ã é um subconjunto não vazio de a . Por exemplo, se a regra AB => CD vale, então as regras ABC => D e ABD => C também devem valer.

Esta característica é similar à propriedade que se um conjunto de itens é extenso então também são todos seus subconjuntos. A partir de um conjunto de itens

extensos l ,portanto, inicialmente são geradas todas as regras com um item no conseqüente. Então são usados os conseqüentes destas regras e a função apriori-gen da seção 2.11.3 para gerar todos os possíveis conseqüentes com dois itens que podem aparecer numa regra gerada apartir de l, e assim em diante. Um algoritmo usando esta idéia é apresentado a seguir.

1) **forall** extensos k-itemsets l_k,k >= 2 **do begin**
2) H_1 = { conseqüentes das regras de l_k com um item no conseqüente }
3) **call** ap-genrules(l_k, H_1);
4) **end**
5) **procedure** ap-genrules(l_k : extenso k-itemset, H_1 : conjunto de m-item conseqüentes)
6) **if** (k > m + 1) **then begin**
7) H_{m+1}= apriori-gen (H_m);
8) **forall** $h_{m+1} \in H_{m+1}$ **do begin**
9) conf = support (l_k) / support (l_k - h_{m+1});
10) **if** (conf >= minconf **then**
11) **output** regra(l_k - h_{m+1})=>h_{m+1} com confiança=conf e suporte=support(l_k);
12) **else**
13) **delete** h_{m+1} from H_{m+1} ;
14) **end**
15) **call** ap-genrules(l_k, H_{m+1});
16) **end**

Geração de candidato apriori:
Function apriori-gen(L_{k-1});
5) **insert into** C_k
6) **select** p[1],p[2],...,p[k-1],q[k-1]
7) **from** L_{k-1} p, L_{k-1} q
8) **where** p[1] = q[1],...,p[k-2] = q[k-2],p[k-1] < q[k-1];

2.12.1 EXEMPLO DE GERAÇÃO DE REGRAS DE ASSOCIAÇÃO

O quadro a seguir apresenta as regras de associação obtidas para o exemplo de carrinhos de compra num supermercado, pressupondo um suporte mínimo igual a 2:

1. Prod2=p 3 ==> Prod5=p 3 (1)
2. Prod5=p 3 ==> Prod2=p 3 (1)
3. Prod2=p Prod3=p 2 ==> Prod5=p 2 (1)
4. Prod3=p Prod5=p 2 ==> Prod2=p 2 (1)
5. Prod1=p 2 ==> Prod3=p 2 (1)

2.12.2 CONCLUSÕES

Os criadores do algoritmo APRIORI, afirmam em [AIS93] sua eficácia, baseados na aplicação do mesmo a dados de uma grande empresa de vendas. Para o arquivo usado nos testes os autores afirmam que o algoritmo exibiu excelente performance.

3 PLANO DE SAÚDE

3.1 CONTEXTO

São consideradas operadoras de seguros privados de assistência à saúde: as pessoas jurídicas constituídas e reguladas em conformidade com a legislação específica para a atividade de comercialização de seguros e que garantam a cobertura de riscos de assistência à saúde, mediante livre escolha pelo segurado do prestador do respectivo serviço e reembolso de despesas.

A assistência compreende todas as ações necessárias à prevenção da doença e à recuperação, à manutenção e à reabilitação da saúde, observados os termos da lei e do contrato firmado entre as partes.

As operadoras de planos privados de assistência à saúde só podem comercializar ou operar planos que tenham sido previamente protocolados na SUSEP, de acordo com as normas técnicas e gerais definidas pelo CNSP, Conselho Nacional de Saúde Pública.

Os segurados são reembolsados, nos limites das obrigações contratuais, das despesas efetuadas pelo beneficiário, titular ou dependente, com assistência à

saúde, em casos de urgência ou emergência, quando não for possível a utilização de serviços próprios, contratados ou credenciados pelas operadoras.

O sistema de saúde suplementar viveu nos últimos meses aquilo que podemos chamar de verdadeira revolução. Pela primeira vez na história, os planos de saúde foram objeto de intensas discussões envolvendo todos os segmentos (governo, parlamentares, empresas e órgãos de defesa do consumidor), que resultaram na primeira lei e conseqüente regulamentação do setor.

Por conta disso, desde o ano passado, diversas ações têm sido empreendidas no sentido de enquadrar os planos de saúde oferecidos às regras estabelecidas pela nova legislação. Se, por um lado, o cenário provocou – e ainda provoca – turbulências, por outro criou condições para que aflorassem novas visões e oportunidades dentro do universo da saúde suplementar.

Um importante fato, foi o reconhecimento por parte de todos os segmentos, da importância social e econômica da autogestão. O governo, por exemplo, a reconhece como modelo para planos coletivos de assistência à saúde, por apresentar a melhor relação custo x benefício.

3.2 ELOSAUDE

3.2.1 APRESENTAÇÃO

O ELOSAÚDE é um programa de conteúdo social destinado aos empregados e Diretores da ELETROSUL, da GERASUL e da ELOS, aposentados e pensionistas da ELOS, extensivo aos seus dependentes.

3.2.2 CARACTERÍSTICAS DO ELOSAÚDE

O ELOSAÚDE caracteriza-se por:

- Ser destituído de fins lucrativos.
- Ser constituído financeiramente através de contribuição mensal dos usuários.
- Ser constituído de planos específicos, abrangendo Assistência Médica, Odontológica e Farmacêutica.

3.2.3 OBJETIVOS DO ELOSAÚDE

Oferecer aos usuários do ELOSAÚDE a cobertura das despesas com saúde, conforme estabelecido nos planos.

3.2.4 USUÁRIOS DO ELOSAÚDE

São considerados como usuários as seguintes categorias:

- Participante Titular/Responsável
- Empregado e ex-empregado da ELOS, da ELETROSUL ou da GERASUL, desde que participante do Plano de Benefícios da ELOS;
- Aposentado e Pensionista da ELOS, da ELETROSUL e da GERASUL;
- Diretor da ELETROSUL e da GERASUL.
- Dependente Direto:
  . Cônjuge;
  . Companheira (o) reconhecida (o) como dependente previdenciária (o);
  . Filhos (as), filhos (as) adotivos (as) e enteados (as);
  . Menores sob guarda judicial ou tutela, até o seu término.
- Dependente Agregado:
  . Irmãos (ãs);
  . Pais;

. Sogros (as);

. Netos;

. Sobrinhos;

. Tio (a), primo (a) em primeiro grau;

. Genro, nora;

. Ex-cônjuge ou companheiro (a);

. Pessoas cujo laço de parentesco se assemelhe ao de dependentes aceitos, tais como: padrasto, madrasta do (a) participante Titular/Responsável ou de seu cônjuge ou companheiro (a) de qualquer idade.

3.2.5 ABRANGÊNCIA DO ELOSAÚDE

O quadro a seguir delimita a abrangência do ELOSAUDE caracterizando por grupo, o participante do plano e seus dependentes .

GRUPOS	VINCULAÇÃO
GRUPO I	Participante Titular/Responsável: empregados da ELOS, da GERASUL e da ELETROSUL, que participe do Plano de Benefícios da ELOS e Diretores da ELETROSUL e da GERASUL. Dependente: todos os dependentes diretos reconhecidos pela ELOS.
GRUPO II	Participante Titular/Responsável: aposentado pelo Plano de Benefícios da ELOS. Dependente: todos os dependentes diretos reconhecidos pela ELOS.
GRUPO III	Participante Titular/Responsável: pensionista pelo Plano de Benefícios da ELOS. Dependente: todos os dependentes diretos reconhecidos pela ELOS.
GRUPO IV	Participante Titular/Responsável: ex-empregado da ELOS, GERASUL e da ELETROSUL, desde que participante do Plano de Benefícios da ELOS. Dependente: todos os dependentes diretos reconhecidos pela ELOS.
GRUPO V	São os dependentes agregados dos participantes titulares/Responsável ou responsáveis dos grupos I, II, III e IV.

Figura 3: Abrangência do ELOSAUDE

CONSIDERAÇÕES DA CONDIÇÃO DE USUÁRIO DO ELOSAÚDE

A condição de usuário do ELOSAÚDE está associada às seguintes condições:

- No caso de aposentadoria de usuário empregado da ELOS ou da ELETROSUL e GERASUL, a passagem para o GRUPO II será automática.
- No caso de falecimento do Participante Titular/Responsável, o pensionista vinculado ao Plano de Benefícios da ELOS passará, automaticamente, para o GRUPO III.
- A permanência na condição de usuário, inclusive em período de licença sem remuneração, dar-se-á enquanto mantiver suas contribuições em dia.
- Os dependentes agregados serão considerados usuários somente enquanto permanecer vivo o participante Titular/Responsável ou o pensionista.

3.2.6 COMPOSIÇÃO DO ELOSAÚDE

O ELOSAÚDE é composto por Planos distintos, podendo haver combinação entre eles conforme o interesse do usuário. O ELOSAÚDE será operacionalizado através da utilização dos seguintes sistemas para a cobertura das despesas:

Sistema de Livre Escolha Dirigida (Credenciamento): utilização da rede de profissionais e instituições credenciadas, estabelecidos os preços, de comum acordo, nos termos de Prestação de Serviços. Neste sistema o usuário terá a co-participação conforme definida em tabela no item 11.

Sistema de Livre Escolha (Reembolso): utilização de profissionais e instituições de livre escolha do usuário, para realização de despesas com saúde que serão posteriormente reembolsadas, nos limites estabelecidos para cada Plano.

3.2.6.1 PLANO A

Assistência Médico-Hospitalar: cobertura de despesas com consultas, exames complementares, internações hospitalares (com acomodação em apartamento Simples/sem acomodação para acompanhante e com banheiro privativo). A

Assistência Médico-Hospitalar será concedida pelos sistemas de Credenciamento ou Reembolso. A cobertura no sistema de reembolso será de 1 (uma) vez a tabela da Associação Médica Brasileira - AMB.

3.2.6.2 PLANO B

Assistência Odontológica Básica: cobertura para assistência Odontológica básica, concedida através dos sistemas de Credenciamento ou Reembolso. A cobertura no sistema de Reembolso será de acordo com a Tabela de Serviços Odontológicos própria do Plano.

3.2.6.3 PLANO C

Assistência Odontológica Especializada: cobertura para assistência Odontológica básica acrescida de assistência odontológica especializada, através dos sistemas de Credenciamento ou Reembolso. A cobertura no sistema de Reembolso será de acordo com a tabela de Serviços Odontológicos própria do Plano.

3.2.6.4 PLANO D

Assistência Farmacêutica: cobertura das despesas com aquisição de medicamentos, através do sistema de Reembolso, com um valor mensal limitado.

3.2.6.5 PLANO E

- **Assistência Médico-Hospitalar**: cobertura de despesas com consultas, exames complementares, internações hospitalares (com acomodação em apartamento tipo standard/com acomodação para acompanhante e banheiro privativo), através dos sistemas de Credenciamento ou Reembolso. A

cobertura no sistema de Reembolso será de 3 (três) vezes a Tabela da Associação Médica Brasileira - AMB.

- **Assistência Psiquiátrica e Psicológica**: cobertura das despesas com psiquiatria e psicologia, através dos sistemas de Credenciamento ou Reembolso, com limite mensal de sessões. A cobertura no sistema de Reembolso será de 3 (três) vezes a Tabela própria de Psicologia e Psiquiatria, adotada pelo Plano.
- **Assistência Fonoaudiológica e Foniátrica**: cobertura das despesas com fonoaudiologia e foniatria, através dos sistemas de Credenciamento ou Reembolso, com limite mensal de sessões. A cobertura no sistema de Reembolso será de 3 (três) vezes a Tabela de Fonoaudiologia e Foniatria, adotada pelo Plano.
- **Assistência aos Portadores de Deficiência Física e/ou Mental Incapacitante**: cobertura das despesas exclusivas de deficiências físicas e/ou mentais incapacitantes, pelo sistema de Credenciamento ou Reembolso. A cobertura no sistema de Reembolso será de 3 (três) vezes a Tabela própria do Plano.

3.2.7 SERVIÇOS COBERTOS PELO ELOSAÚDE

Os serviços de saúde cobertos pelo ELOSAÚDE são específicos de cada Plano contratado.

As despesas de um serviço em saúde só poderão ser cobertas pelo sistema de Credenciamento ou Reembolso, não podendo haver dupla cobertura de uma mesma despesa.

3.2.7.1 PLANO A

Assistência Médico-Hospitalar.

Consultas, visita médica domiciliar, medicina física e reabilitação, cardiologia, endocrinologia, geriatria e gerontologia, hepatologia, hematologia e hemoterapia,

infectologia, neurologia, pediatria, pneumologia, anestesiologia, nutrição parenteral e enteral, alergologia, anatomia patológica e citopatológica, eletrencefalografia e neurosisiologia clínica, endoscopia digestiva e peroral, patologia clínica, fisiopneumologia, radiodiagnóstico, ultrassonografia, angiologia cirúrgica, vascular e linfática, cirurgia cardíaca, hemodinâmica, cirurgia de cabeça e pescoço, dermatologia clínico-cirúrgica, cirurgia do aparelho digestivo, órgãos anexos e parede abdominal, cirurgia endocrinológica, ginecologia e obstetrícia, cirurgia da mama, cirurgia da mão, neurocirurgia, oftalmologia, otorrinolaringologia, ortopedia e urologia, artefatos oftalmológicos (óculos e lentes), com internação em apartamento Simples (sem acomodação para o acompanhante e com banheiro privativo).

- 30 (trinta) dias de internação, por ano, em hospital psiquiátrico ou em unidade ou enfermaria psiquiátrica em hospital geral, para portadores de transtornos psiquiátricos em situação de crise.
- 15 (quinze) dias de internação, por ano, em hospital geral, para pacientes portadores de quadros de intoxicação ou abstinência provocados por alcoolismo, ou outras formas de dependência química que necessitem de hospitalização.

3.2.7.2 PLANO B

Assistência Odontológica Básica:

Consulta, profilaxia/controle de placa, restauração em resina composta (classes I, III, IV e V), restauração de amálgama, exodontia, aplicação tópica de flúor, urgência, exodontia de dentes decíduos, raspagem e polimento coronário, tratamento endodôntico de incisivos e caninos, tratamento endodôntico de dentes decíduos, radiografia periapical e radiografia interproximal.

3.2.7.3 PLANO C

Assistência Odontológica Especializada:

Todos os serviços prestados no Plano de Assistência Odontológica Básico e, frenectomia labial e lingual, remoção de dentes retidos, ulectomia, tratamento de lesão cística - enucleação, restauração em resina composta (classe II e restauração), apicetomia (com obturação retrógrada), clareamento ou recromia, retratamento endodôntico em pré-molares, retratamento endodôntico em molares, retratamento endodôntico em incisivos e caninos, retratamento endodôntico em pré-molares e molares, coroa de aço ou policarbonato para dentes decíduos, cirurgia periodontal a retalho com ou sem osteotomia, gengivectomia ou gengivoplastia, raspagem de cálculo, polimento e curetagem de bolsas supra e infra-ósseas, apicificação (dentes com risogênese incompleta a longo termo), restauração metálica fundida, núcleo metálico fundido, prótese unitária em cerâmica (incisivos e caninos), prótese unitária em cerâmica (pré-molares), prótese parcial removível com grampos (bilateral) e prótese total removível. Aparelho Ortodôntico fixo, placa de Hawley. Aparelho de contenção final e aparelho extra-bucal.

3.2.7.4 PLANO D

Assistência Farmacêutica:
Medicamentos adquiridos em farmácias e desde que acompanhados por respectiva receita médica.

3.2.7.5 PLANO E

Assistência Médico-Hospitalar:
Todos os serviços do Plano A e cirurgia plástica reparadora, desde que decorrente de acidente ou de má formação congênita, com internações hospitalares em Apartamento Standard (com acomodações para acompanhante e com banheiro privativo).

Assistência Psiquiátrica/Psicológica:

Psiquiatria: diagnose e terapia.
Psicologia: diagnose, avaliação e terapia.

Assistência Fonoaudiológica e Foniátrica:
Fonoaudiologia: diagnose e terapia.
Foniatria: diagnose e terapia.

Assistência aos Portadores de Deficiência Física e/ou Mental Incapacitante: Cobertura das despesas específicas com recuperação/atendimento de portadores de deficiência física e/ou mental incapacitante, tais como: atendimento psico-pedagógico especializado, fraldões e outros previstos em tabela própria do Plano, até o limite mensal estipulado.
Considera-se como:

- Deficiência Mental: o déficit intelectivo e/ou integrativo moderado, severo e profundo, inato ou adquirido, que se manifesta em qualquer época da vida, tornando o afetado dependente de ajuda de outras pessoas e/ou entidades especializadas para sua sobrevivência e convivência psico-social.
- Dependência Física: Algum tipo de deficiência física permanente, com perda da função voluntária, inata ou adquirida, tornando a pessoa afetada dependente de outras pessoas e/ou de equipamentos especiais para sua locomoção e/ou integração sócio-educacional.

3.2.8 SERVIÇOS NÃO COBERTOS PELO ELOSAÚDE

Não serão consideradas como despesa com saúde os seguintes procedimentos:

- Atos proibidos pelo Código de Ética Médica;
- Tratamento de despesas profissionais práticos;
- Terapia ocupacional;
- Tratamento convalescente após alta médica;
- Tratamento estético, clínico ou cirúrgico;
- Tratamento hospitalar de moléstias incuráveis;

- Transplante, implantes e reimplantes, a exceção de rins e córnea;
- Tratamento domiciliar;
- Check-up;
- Monitoragem fetal;
- Efeito mórbido provocado por atividades esportivas de risco voluntário, como: asa-delta, motociclismo, caça, boxe, esqui aquático, esqui na neve, jet-ski, etc.;
- Enfermagem em caráter particular;
- Exames pré-admissionais e demissionais;
- Atendimento por acidente de trabalho;
- Cirurgia plástica com finalidade estética;
- Serviços prestados por Profissionais / Instituições (Clínicas, Laboratórios, Hospitais, Outros), que tenham sido descredenciados pelo ELOSAÚDE;
- Inseminação artificial;
- Internação para tratamento de estresse;
- DIU;
-
- Sonoterapia;
- Tratamento em estâncias minerais ou de repouso;
- Despesas extras ao tratamento quando da hospitalização (revistas, cigarros, lavanderia, telefones, frigobar, refeições para acompanhante, etc);
- Lesões traumáticas ou deformidades e suas conseqüências, existentes antes do início da cobertura do Plano;
- Tratamentos experimentais e medicamentosos, ainda não reconhecidos pelo Serviço Nacional de Fiscalização de Medicina e Farmácia;

Não serão consideradas como despesas com saúde para cada Plano específico, os seguintes procedimentos:

3.2.8.1 PLANO A

- Todos os serviços de Assistência Odontológica;

- Todos os serviços de Assistência Psiquiátrica e Psicológica;
- Todos os serviços de Assistência Foniátrica ou Fonoaudiológica;
- Todos os serviços pertinentes às deficiências físicas e/ou mentais incapacitantes;
- Todos os medicamentos;
- Aparelhos em geral (surdez, ortopédicos, etc);
- As próteses em geral (mamárias, ocular, auditiva, etc);
- Diárias, refeições e demais despesas de acompanhantes.

3.2.8.2 PLANO B

- Todos os serviços de Assistência Médico-Hospitalar;
- Todos os serviços de Assistência Psiquiátrica e Psicológica;
- Todos os serviços de Assistência Foniátrica ou Fonoaudiológica;
- Todos os serviços pertinentes às deficiências físicas e/ou mentais incapacitantes;
- Todos os medicamentos, inclusive os referentes à Assistência Odontológica;
- Aparelhos em geral (surdez, ortopédicos, etc);
- As próteses em geral (mamárias, ocular, auditiva, etc);
- Os serviços de odontologia específicos, não previstos no Plano de Odontologia Básico.

3.2.8.3 PLANO C

- Todos os serviços de Assistência Médico-Hospitalar;
- Todos os serviços de Assistência Psiquiátrica e Psicológica;
- Todos os serviços de Assistência foniátrica ou fonoaudiológica;
- Todos os serviços pertinentes às deficiências físicas e/ou mentais incapacitantes;
- Todos os medicamentos, inclusive os referentes à Assistência
- Odontológica;

- Aparelhos em geral (surdez, ortopédicos, etc);
- As próteses em geral (mamárias, ocular, auditiva, etc).

3.2.8.4 PLANO D

- Todos os serviços de Assistência Médico-Hospitalar;
- Todos os serviços de Assistência Odontológica;
- Todos os serviços de Assistência Psiquiátrica e Psicológica;
- Todos os serviços de Assistência Fonoaudiológica e Foniátrica;
- Aparelhos em geral (surdez, ortopédicos, etc);
- As próteses em geral (mamárias, oculares, auditivas, etc).

3.2.8.5 PLANO E

- Os aparelhos em geral (surdez, ortopédicos, etc);
- As próteses em geral (mamárias, oculares, auditivas, etc);
- Todos os serviços de Assistência Odontológica;
- Os medicamentos adquiridos.

3.2.9 CARÊNCIAS DO ELOSAÚDE

Serão observadas as seguintes carências, contadas a partir da data de inscrição:

- 300 (trezentos) dias para partos a termo.
- prazo de 180 (cento e oitenta) dias para os demais casos, exceto prótese dentária e ortodontia.
- 720 (setecentos e vinte) dias para serviços de próteses dentárias e ortodontia.
- Prazo de 24 (vinte e quatro) horas para as coberturas dos casos de urgência e emergência.

Observação: Quando o atendimento de Urgência ou Emergência evoluir para internação ou necessitar de outros procedimentos dentro do período de carência a cobertura cessará, sendo que a responsabilidade financeira passará a ser do usuário.

- Na transferência para novo Plano, que ofereça serviços ou coberturas superiores ao Plano anterior, serão exigidos os períodos de carência para os serviços ou coberturas superiores. Será concedida a permanência na utilização dos serviços do Plano anterior, comum ao novo Plano, sem carências.
- Na transferência para novo Plano, que seja inferior em serviços ou coberturas do seu Plano anterior, não será necessário o cumprimento dos períodos de carência desde que já tenham sido cumpridos no Plano anterior.

3.2.10 CO-PARTICIPAÇÃO

Visando a regulação de demanda na utilização dos serviços de saúde, serão adotadas as taxas de co-participação, conforme tabela abaixo:

3.2.10.1 TABELA DE CO-PARTICIPAÇÃO

a) A aquisição de medicamentos respeitará o limite mensal de CH (Coeficiente de Honorários), por usuário do Plano.

b) As despesas com deficiências físicas e/ou mentais incapacitantes respeitarão o limite mensal em CH (Coeficiente de Honorários), por usuário do Plano.

c) Na transferência do usuário para outro Plano (superior ou inferior), serão considerados os serviços já utilizados no plano anterior para aplicação da tabela de co-participação.

d) No caso de tomografia computadorizada e ressonância magnética que tenham o mesmo código na tabela de Honorários Médicos e que sejam realizadas em

diferentes membros do corpo humano, a co-participação será cobrada a partir da terceira, isto é, não haverá co-participação para o primeiro procedimento em cada membro.

PROCEDIMENTO	PARTICIPAÇÃO DE 50% PELO USUÁRIO
Consulta de qaulquer natureza	Acima de 06 ao ano
Visitas médicas domiciliares	Acima de 06 ao ano
Hospitalização/Internação	Acima de 40 dias ao ano
Hospitalização em UTI	Acima de 15 dias ao ano
Psicologia, psiquiatria, fonoaudiologia e foniatria	Acima de uma sessão semanal
Exames complementares de qualquer natureza, exceto os *	Acima de 03 por procedimento ao ano
*Ressonância magnética	Acima de 01 por procedimento ao ano
*Tomografia computadorizada	Acima de 01 por procedimento ao ano
Radioterapia	Acima de 60 aplicações ao ano
Quimioterapia	Acima de 20 aplicações ao ano
Fisioterapia	Acima de 20 sessões por procedimento ao ano
Visitas hospitalares de qualquer especialidade médica	Acima de 40 ao ano
Serviços odontológicos realizados através do sistema de credenciamento	Participação de 10%

Figura 4: **Tabela de co-participação**

3.2.10.2 PAGAMENTO DA CO-PARTICIPAÇÃO

a)A co-participação será descontada através da folha de benefícios da ELOS e da folha de pagamento das Patrocinadoras.

b)Caso não haja margem consignável, a cobrança será efetuada através de boleto bancário, podendo haver parcelamento a pedido do participante titular.

3.2.11 CONDIÇÕES DA UTILIZAÇÃO DO SISTEMA DE REEMBOLSO

Os usuários que utilizarem o sistema de reembolso terão o crédito em conta-corrente do participante Titular/Responsável, das despesas apresentadas, até os tetos previstos nas tabelas adotadas pelo ELOSAÚDE.

- Usuários que estejam inadimplentes com o Plano, terão seus benefícios suspensos, não tendo retroatividade dos mesmos após a quitação.
- Todos os serviços hospitalares, tratamentos odontológicos e reembolsos diversos, estarão sujeitos a perícia ou outro tipo de comprovação.
- Tratamento no exterior: será reembolsado de acordo com a tabela da Associação Médica Brasileira - AMB, cujo comprovante de pagamento deverá ser encaminhado pelo participante devidamente traduzido, constituindo responsabilidade do titular do plano a prestação das competentes informações.

3.2.12 MENSALIDADE

A mensalidade de cada Plano será constituída em concordância com os serviços cobertos e o Grupo a que o usuário estiver vinculado e acrescida da taxa de administração.

3.2.13 REAJUSTE DAS MENSALIDADES

- 3.2.14.1 Os valores das mensalidades dos Planos, serão reajustados de acordo com a variação do Coeficiente de Honorários (CH), divulgado pela Associação Médica Brasileira - AMB.
- 3.2.14.2 As mensalidades em atraso terão seus valores corrigidos pela variação do Coeficiente de Honorários – CH, além da atualização monetária.
- 3.2.14.3 Caso ocorrer déficit no GRUPO V , os valores das mensalidades deste GRUPO serão reajustadas tão somente para recompor o equilíbrio econômico e financeiro do GRUPO.

- 3.2.14.4 Periodicamente a Fundação ELOS fará uma reavaliação atuarial do ELOSAÚDE, podendo ocorrer reajustes nos preços, independente do previsto no item 3.2.14.1.

3.2.14 CUSTEIO DO ELOSAÚDE

- O custeio do ELOSAÚDE será formado através das mensalidades de seus usuários, das taxas para formação de Fundo de Reserva, Taxas Administrativas, resultado das aplicações financeiras e possíveis doações.

- Todo usuário pagará uma taxa de 10% (dez por cento), adicional ao valor da mensalidade, durante o período de 30 (trinta) meses, a partir de sua inscrição, para a constituição do Fundo de Reserva.

– Constitui-se também numa forma de custeio a co-participação do usuário definida no item 11.

3.2.15 ADESÃO, DESLIGAMENTO E REINCLUSÃO

A adesão, desligamento e reinclusão seguem critérios descritos a seguir.

3.2.15.1 A ADESÃO

- A adesão ao ELOSAÚDE deverá ser solicitada através do Pedido de Filiação, mediante declaração no ato, da aceitação integral deste Regulamento.
- O pedido de filiação de dependente deverá ser acompanhado de comprovação de vínculo familiar jurídico de cada um com o participante Titular ou Responsável.
- Declaração de saúde - As informações prestadas no formulário de Declaração de Saúde, são de inteira responsabilidade do participante Titular/Responsável.

– Nos casos de doenças e lesões pré-existentes, o participante optará por uma das condições abaixo:

a) a cobertura parcial temporária por 24 (vinte e quatro) meses, a contar da data de adesão ao plano.

b)Agravo, cujo acréscimo de valor será definido de acordo com a complexidade da patologia apresentada. Esta alternativa somente será oferecida após a regulamentação pelo CONSU – Conselho de Saúde Suplementar do Ministério da Saúde.

DOENÇA PRÉEXISTENTE - aquela cujos sinais ou sintomas manifestarem-se antes da adesão ao ELOSAÚDE.

- A critério da Fundação ELOS, quando esta julgar necessário, os usuários a serem incluídos na vigência deste Regulamento deverão ser submetidos previamente a exame médico/odontológico, para fins de avaliação da sua integridade física e mental.

3.2.15.2 O DESLIGAMENTO

- O desligamento de um usuário do ELOSAÚDE, independente do GRUPO a que estiver vinculado, será efetivado mediante preenchimento de formulário próprio para o desligamento, não cabendo restituição das contribuições efetuadas.
- Quando da comunicação de desligamento, o usuário deverá obrigatoriamente, devolver a "Carteira de Usuário" do ELOSAÚDE, e será imediatamente suspensa a utilização dos serviços prestados pelo Plano, enquadrando-se seu uso, após o desligamento, como crime de falsidade ideológica, sujeito às penalidades cabíveis.
- O não pagamento de 2 (duas) mensalidades consecutivas, acarretará o desligamento automático do participante titular, bem como dos dependentes diretos e/ou agregados a ele vinculados.
- O titular que se desligar do Plano de Benefícios da ELOS, poderá permanecer no ELOSAÚDE no máximo por 24 (vinte e quatro) meses, assumindo o

pagamento das mensalidades e participações relativas a todos os usuários a ele vinculados.

- O titular/responsável que não regularizar o débito, observado o disposto no item 11.2, referente a co-participação no prazo de 60 (sessenta) dias, terá seus benefícios suspensos.
- O usuário ao se desligar do plano, deverá quitar os débitos existentes.
- O usuário que cometer fraude decorrente de falsidade ideológica ou de outra natureza, será desligado imediatamente do Plano, juntamente com seus dependentes de qualquer categoria, independente do ressarcimento dos prejuízos causados e das penalidades da lei.
- O desligamento automático do titular e de seus dependentes, não implicará na anistia das mensalidades em atraso, permanecendo a dívida até a sua quitação.

3.2.15.3 A REINCLUSÃO

A reinclusão de usuário que tenha ou não cumprido os períodos de carência, dar-se-á mediante o cumprimento integral das carências previstas no item 10 (dez) deste Regulamento.

3.2.15.4 FORMALIZAÇÃO DE PEDIDO

Para inclusão, alteração, desligamento e reinclusão, os formulários ou correspondências de solicitação, deverão chegar ao ELOSAÚDE até o 5° (quinto) dia útil de cada mês, para processamento em tempo hábil.

3.2.16 PAGAMENTO DAS MENSALIDADES

- A responsabilidade pelo pagamento das mensalidades dos usuários (participante titular, dependente direto ou agregado), será do participante titular ou responsável.

3.2.17 RESPONSABILIDADES PELO ELOSAÚDE

- A Fundação ELOS será a responsável pelas aplicações das receitas do ELOSAÚDE no mercado financeiro, pelo registro e contabilização em separado, de acordo com a legislação que norteia a atividade previdenciária, das atividades do ELOSAÚDE, devendo extrair relatórios periódicos da situação econômico-financeira.
- É de responsabilidade da ELOS manter todos os usuários informados das decisões referentes ao ELOSAÚDE.
- Será cobrada uma Taxa de Administração de até 15% (quinze por cento) da mensalidade para cobertura dos custos operacionais do ELOSAÚDE.
- Compete ao Conselho de Curadores da Fundação ELOS aprovar alterações neste Regulamento, deliberar sobre alterações de mensalidades, estabelecer normas administrativas e deliberar sobre casos omissos neste Regulamento.

3.2.18 CONSIDERAÇÕES GERAIS

- Os vários estipuladores na Tabela de Reembolso, para honorários médicos e exames complementares, têm como parâmetro a Tabela da Associação Médica Brasileira - AMB, edição 1992.
- Os valores estipulados na Tabela de Serviços Odontológicos têm como parâmetro a Tabela própria do Plano.
- As notas técnicas atuariais integram este Regulamento.

- As orientações contidas no MANUAL DO BENEFICIÁRIO, integram este REGULAMENTO, subordinando-se a este, na hipótese de haver dúvidas de interpretação.

3.2.19 CONCLUSÃO

O ELOSAUDE, por ser um plano de saúde fechado, sustentado exclusivamente pela mensalidade de seus participantes, requer uma administração austera, eficaz e atenta ao comportamento do mercado.

Há um constante monitoramento sobre o custo e o perfil de demanda de serviços de seus beneficiários, para manter viável individualmente cada um dos planos oferecidos. Tem sido usados cálculos estatísticos como auxiliar neste monitoramento.

A mineração de dados pode atuar como coadjuvante, justamente apontando comportamentos sistemáticos dos beneficiários, e também na detecção de indicativos de faturas fraudulentas.

4 MODELO PROPOSTO

4.1 CONSIDERAÇÕES

Para o processamento de dados com a finalidade de obter regras de associação, é necessário dispor de uma infra-estrutura que facilite o processamento de todas as etapas inerentes ao processo:

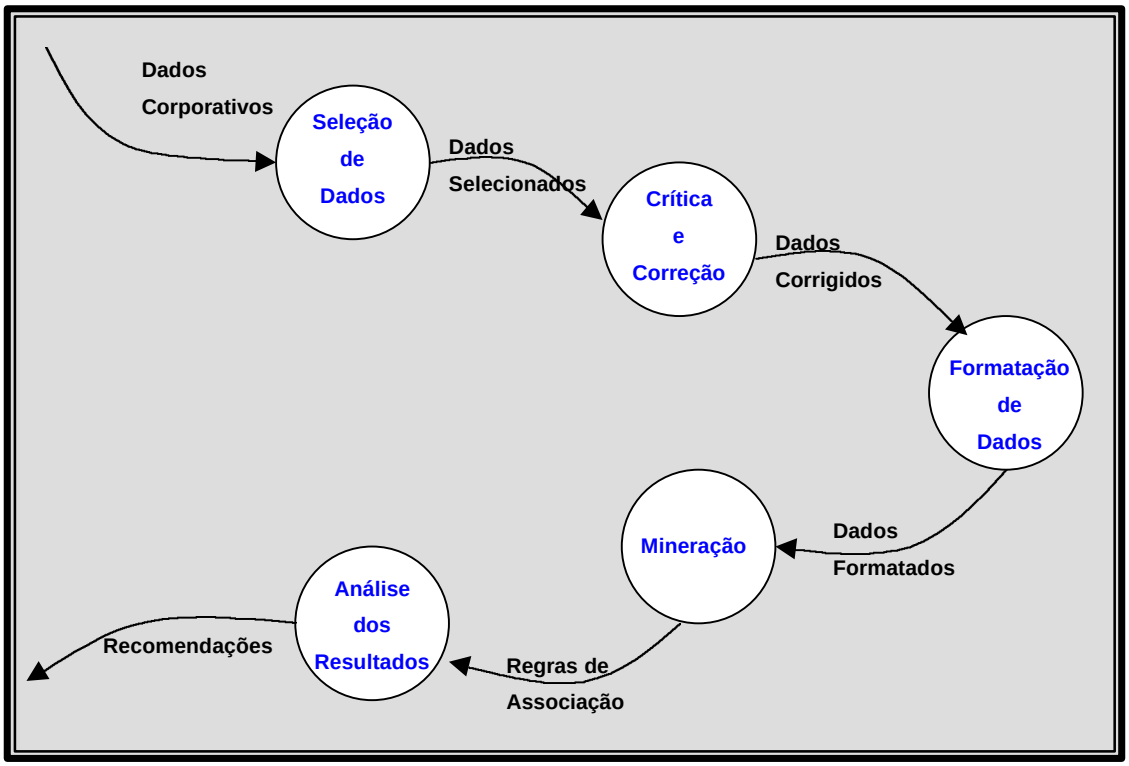

Figura 5: Etapas inerentes ao processo de mineração de dados

- Seleção de dados de interesse ao estudo
- Crítica dos dados e correção
- Geração de um arquivo no formato adequado ao programa de mineração
- Mineração dos dados para obtenção de regras de associação
- Análise dos resultados obtidos

Implementando em computador um sistema composto de módulos correspondentes a cada uma das etapas citadas, fica construída a ponte entre os dados corporativos e a obtenção de resultados pela técnica de mineração de dados.

A seguir serão descritas e justificadas cada uma destas etapas.

4.2 SELEÇÃO DE DADOS DE INTERESSE AO ESTUDO

Uma vez definido o estudo a ser feito, fica fácil arrolar os dados a serem envolvidos no processamento.
Em geral os dados corporativos estão na terceira forma normal, armazenados em arquivos convencionais ou num gerenciador de banco de dados, para uso pelos sistemas legados da empresa, ou em ambos, dependendo do estágio de informatização e das plataformas de software em uso.
Além das séries históricas de consumo de produtos e/ou serviços, podem ser necessários dados pessoais como, sexo, idade, região de moradia. Especificamente idade não costuma ser armazenada, mas sim data de nascimento, o que torna necessário calcular a idade na ocasião de cada operação de compra de cada cliente. Se o estudo envolver região de moradia, mas só existir no cadastro o CEP (código de endereçamento postal), da mesma forma será necessário converter CEP em região.
Considerando que o algoritmo **APRIORI** efetua várias leituras completas do arquivo de dados, convém minimizar não apenas o número de registros a ler, mas

também o número de diferentes itens de dados, objetivando melhorar a performance do **APRIORI**, em termos de tempo total de processamento.

4.3 CRÍTICA DOS DADOS E CORREÇÃO

Embora os dados corporativos estejam armazenados, criticados e supostamente corretos, é necessário verificá-los e corrigi-los quando for o caso. Diferenças na grafia, embora mantendo o significado, são diferenças para um programa de computador. Por exemplo, pepsi é diferente de pepsi-cola, apesar de não haver dúvida de tratar-se do refrigerante.

4.4 GERAÇÃO DE UM ARQUIVO NO FORMATO ADEQUADO AO PROGRAMA DE MINERAÇÃO

Um módulo do sistema deve ler os dados selecionados e corrigidos para gerar um arquivo no formato adequado a cada estudo a ser realizado.
Nesta etapa podem ocorrer várias situações:

- O programa de mineração pode ser um pacote adquirido pronto, cuja entrada de dados seja num formato proprietário, tornando necessário gerar um arquivo neste formato, para cada estudo. É comum ser um arquivo texto, pela generalidade que propicia ao se adaptar facilmente à quantidade variada de itens de dados de cada estudo.
- A equipe interna da empresa desenvolveu um programa de mineração integrado à base de dados corporativa, simplificando ou até mesmo eliminando esta etapa. Isto pode ser feito, com uma entrada de dados dependente da modelagem de dados em uso na empresa. Talvez seja necessário escrever um módulo de entrada de dados para cada estudo de interesse.

4.5 MINERAÇÃO DOS DADOS PARA OBTENÇÃO DE REGRAS DE ASSOCIAÇÃO

Uma vez preparado o arquivo de dados no formato adequado à mineração, pode ser utilizado o programa de computador adotado.

Nesta etapa o usuário executa os seguintes passos repetidas vezes, até obter resultados satisfatórios:

- define os parâmetros básicos do programa,
- executa o programa,
- avalia os resultados obtidos,
- faz ajustes nos parâmetros básicos do programa.

Algumas vezes embora os dados estejam corretos e o processamento ocorra normalmente, podem ocorrer as seguintes situações:

- Nenhuma regra é gerada – porque nenhuma regra atingiu o suporte mínimo padrão do programa. O programa tenta induzir dez regras, ou a quantidade indicada por parâmetro, com suporte 1.0 (100 %). Não obtendo a quantidade de regras desejadas, tenta com suporte 0.95. O programa tentará obter regras com confiança mínima, decrementando de 5 % o suporte mínimo, até o limite inferior de 0.1 (10 %).
- São obtidas menos regras que o especificado – porque mesmo decrementando o suporte mínimo, em sucessivas iterações, até o limite inferior, poucas regras se mostraram representativas no estudo em questão.

Em ambas as situações, sendo a primeira um caso extremado da segunda, cabe ao usuário fazer tentativas, experimentando suporte mínimo menor que 10 % e/ou confiança mínima menor que 0.9 (90 %).

4.6 ANÁLISE DOS RESULTADOS OBTIDOS

Obtidas as regras de associação, com suporte mínimo e confiança mínima aceitas pelo usuário, cabe a este avaliar o significado das mesmas, lembrando a característica básica das regras induzidas: **Não são mutuamente exclusivas, podendo ser coletivamente exaustivas ou não.**

4.7 CONCLUSÕES

O modelo proposto é muito voltado para a realidade da necessidade de preparação e conseqüente mineração de dados, possibilitando a obtenção do resultado final, as regras de associação.

5 APLICAÇÃO DO ALGORITMO APRIORI A DADOS DE UM PLANO DE SAÚDE PARA OBTER REGRAS DE ASSOCIAÇÃO

5.1 INTRODUÇÃO

Para avaliar o comportamento do algoritmo **APRIORI** poderiam ser utilizados dados de três origens diferentes:

- Dados sintéticos – gerados por programa de computador, considerando características estatísticas de cada item de dado, como média, variância e distribuição de probabilidade.
- Dados disponíveis na internet, consagrados pelo uso, referenciados em diversos trabalhos internacionais, geralmente usados para estudos comparativos entre diferentes algoritmos de mineração.
- Dados reais e atuais, pertencentes a uma empresa do mercado brasileiro, ou mais proximamente, catarinense.

Neste trabalho foram utilizados registros de serviços de saúde prestados aos beneficiários do ELOSAUDE.

5.2 BASE DE DADOS HISTÓRICA

Mensalmente o ELOSAUDE efetua o pagamento dos fornecedores de serviços de saúde. Deste processamento resulta um arquivo contendo quais serviços foram utilizados, por quais beneficiários, quando, em que quantidade e o custo.
O arquivo com o histórico de vários anos destes dados disponibilizado para este trabalho, consta de mais de quatrocentos e cinqüenta mil registros, com a estrutura especificada a seguir:

Campo	Nome do Campo	Tipo	Tamanho	Decimais
1	Tipo do Serviço	Caractere	2	
2	Código do Serviço	Caractere	6	
3	Quantidade	Numérico	3	
4	Valor	Numérico	10	2
5	Data	Caractere	8	
6	Matricula	Caractere	7	

5.3 SELEÇÃO DE DADOS DE INTERESSE AO ESTUDO

Para elaborar estudos envolvendo sexo e idade dos beneficiários do plano de saúde, foi necessário utilizar um cadastro contendo o sexo e a data de nascimento, conforme especificado a seguir:

Campo	Nome do Campo	Tipo	Tamanho	Decimais
1	Matricula	Caractere	7	
2	Nascimento	Caractere	8	
3	Sexo	Caractere	1	

Completou-se o arquivo de dados históricos, com o sexo e a idade dos beneficiários. A idade foi calculada na data de cada utilização de serviço, pois como são vários anos de dados, um mesmo beneficiário pode ter idades diferentes com o passar dos anos. O arquivo final completo, contendo 457.832 registros ficou com a seguinte estrutura:

Campo	Nome do Campo	Tipo	Tamanho	Decimais
1	Tipo do Serviço	Caractere	2	
2	Código do Serviço	Caractere	5	
3	Quantidade	Numérico	3	
4	Valor	Numérico	10	2
5	Data	Caractere	8	
6	Matricula	Caractere	7	
7	Sexo	Caractere	1	
8	Idade	Numérico	3	

Cabe observar que partiu-se de dois arquivos normalizados, ou seja, na terceira forma normal, e chegou-se a um arquivo final desnormalizado. É comum isto ocorrer, e mesmo inevitável, porque o arquivo precisa estar completo para facilitar seu processamento seqüencial. Efetivamente é o mesmo critério usado para projetar Data Warehouse ou Data Mart, que são em sua essência, bases de dados de séries históricas, modeladas para otimizar a performance das consultas, o que é conseguido com a desnormalização.

5.4 GERAÇÃO DE UM ARQUIVO NO FORMATO ADEQUADO AO PROGRAMA DE MINERAÇÃO

É conveniente adotar um formato padronizado, para representar conjuntos de dados que consistem de instâncias independentes, desordenadas que não envolvam relacionamentos entre as instâncias.

5.4.1 O FORMATO ARFF

As referências ao formato ARFF pressupõem a utilização do programa **APRIORI** adotado neste trabalho, que utiliza este formato para sua entrada de dados. Sendo utilizado outro programa de computador, o formato poderá ser outro.
O formato ARFF foi criado pela equipe do projeto weka (Waikato Environment for Knowledge Analysis), da University of Waikato, Department of Computer Science, New Zealand.

O quadro a seguir mostra um arquivo ARFF para os dados do exemplo do supermercado.

```
% Exemplo de carrinhos de compra num supermercado
@relation DadosArtigoIBM.Symbolic

@attribute IdentComprador {100,200,300,400}
@attribute Prod1 {p}
@attribute Prod2 {p}
@attribute Prod3 {p}
@attribute Prod4 {p}
@attribute Prod5 {p}

@data
100,p,?,p,p,?
200,?,p,p,?,p
300,p,p,p,?,p
400,?,p,?,?,p
```

Linhas começando com um símbolo % são comentários.
Após os comentários vem o nome da relação, identificado por uma linha começando com @relation.

A seguir vem um bloco definindo os atributos dos dados a minerar, identificados por linhas começando por @attribute. Atributos nominais são seguidos pelo conjunto de valores que podem assumir, entre colchete, enquanto os numéricos são seguidos pela palavra chave numeric.
Uma linha começando com @data indica que a partir da próxima linha estarão os dados, obedecendo a seqüência indicada pela definição dos atributos. Um atributo não existente numa transação deve ser indicado pelo símbolo **?**.

5.4.2 GERAÇÃO DE ARQUIVO NO FORMATO ARFF

Com o arquivo de dados históricos criticado e corrigido, basta gerar um arquivo texto no formato ARFF para cada estudo a ser realizado.

Há várias considerações a fazer:

- Nas tabelas auxiliares contendo a codificação de tipos de serviço e código de serviços, há 56 tipos de serviço e 5771 códigos de serviços
- A forma tabular de entrada, no formato ARFF é muito ineficiente para muitos dos problemas de regras de associação. As regras de associação são, muitas vezes, usadas em situações onde os atributos são unários – presente ou não – e a maioria dos valores de atributos associados com uma instância estão ausentes. Por exemplo, há mais serviços na tabela de serviços que diferentes serviços consumidos por um determinado beneficiário do plano de saúde. Seria mais eficiente representar cada instância como uma lista de atributos cujo valor está presente ao invés de um vetor com um elemento para cada item possível.
- O algoritmo **APRIORI** trabalha naturalmente com atributos nominais, não com quantidades. É necessário transformar atributos numéricos em atributos nominais. Em outras palavras, para um estudo envolvendo faixa etária, é necessário tratar faixas etárias pelo seu nome: meia idade pode corresponder a 40 a 59 anos; se o estudo envolver idade. 40 deverá ser substituído por quarenta, por exemplo.

É recomendável dispor de um programa de computador que leia o arquivo de dados históricos e gere o arquivo ARFF de cada estudo, pois assim este será gerado íntegro, porque o programa **APRIORI** verifica a integridade entre atributos e dados, interrompendo o processamento caso perceba alguma incoerência.

5.5 DESCRIÇÃO DO PROGRAMA APRIORI

O programa utilizado para o processamento dos dados e obtenção de regras de associação, implementa o algoritmo **APRIORI**, conforme descrito em [AGR93] e [AGR94].

O programa **APRIORI** é composto de dois módulos principais:

- Identificação dos conjuntos mais freqüentes
- Indução de regras de associação

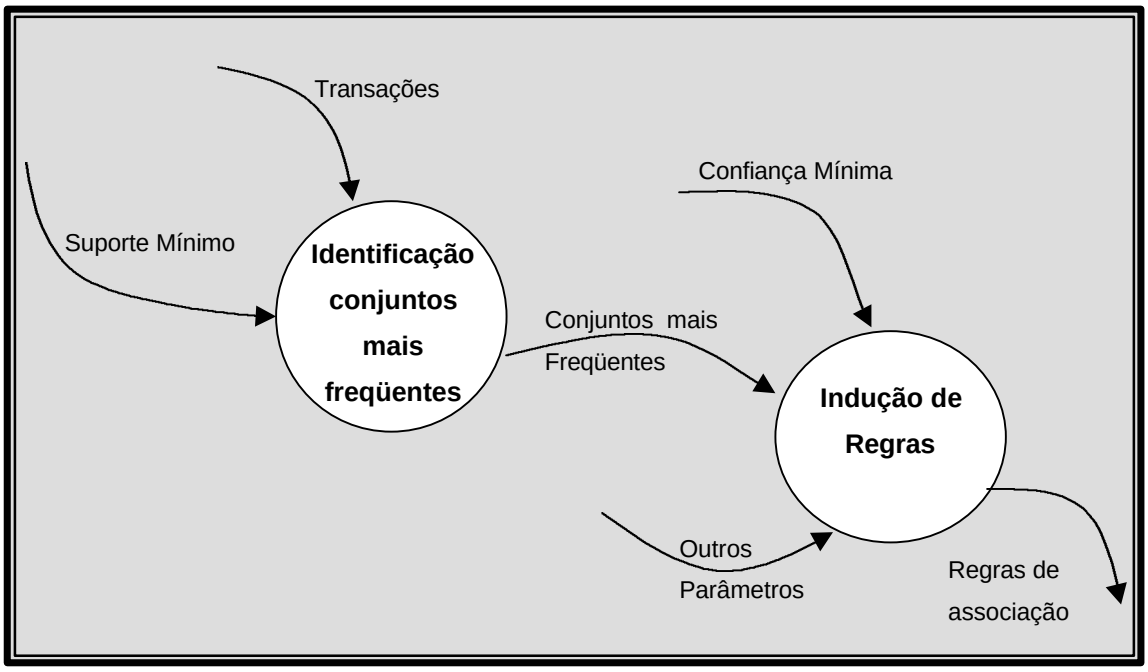

Figura 6: Estrutura do programa APRIORI

Foi utilizada a linguagem java em sua implementação pela equipe do projeto weka. Há duas versões, uma versão interface caractere considerada estabilizada,

e uma versão interface gráfica que, embora considerada em desenvolvimento, funcionou corretamente em todos os casos experimentados.
Por omissão, o **APRIORI** tenta gerar dez regras. Ele começa com um suporte mínimo para as regras, de 100% dos itens dados e decresce este em passos de 5% até que existam pelo menos dez regras com a mínima confiança necessária, ou até que o suporte tenha atingido um limite inferior de 10%, o que ocorrer primeiro. A mínima confiança é estabelecida igual a 0.9, por omissão.

5.5.1.1 ENTRADA DE DADOS

Um arquivo texto no formato ARFF é a entrada de dados para o programa.

5.5.1.2 PARÂMETROS

Alguns parâmetros podem ser fornecidos ao programa **APRIORI** para alterar o seu comportamento:

Opções válidas para o programa:

-N nn
O número requerido de regras, por omissão igual a 10. Exemplo **–N 40**

-C n.nn
A confiança mínima de uma regra, por omissão igual a 0.9. Exemplo **–C 0.5**

-D n.nn
O delta pelo qual o suporte minimo é decrescido em cada iteração, por omissão igual a 0.05. Exemplo: **-D 0.01**

-M n.nn
O limite inferior para o suporte mínimo, por omissão igual a 0.1. Exemplo: **-M 0.01**

-S n.nn

Se usado, as regras são testadas para significância num dado nível, por omissão igual a sem teste de significância.

-I

Se o conjunto de itemsets encontrados deve ser apresentado, por omissão igual a não.

5.5.1.3 RESULTADOS DO PROCESSAMENTO

O quadro a seguir apresenta o resultado do processamento do arquivo ARFF do exemplo de carrinhos de compra num supermercado.

```
Apriori
=======

Minimum support: 0.5
Minimum confidence: 0.9
Number of cycles performed: 11
Generated sets of large itemsets:
Size of set of large itemsets L(1): 9
Size of set of large itemsets L(2): 20
Size of set of large itemsets L(3): 18
Size of set of large itemsets L(4): 7
Size of set of large itemsets L(5): 1
Best rules found:
 1. Prod2=p 3 ==> Prod5=p 3 (1)
 2. Prod5=p 3 ==> Prod2=p 3 (1)
 3. Prod2=p Prod3=p 2 ==> Prod5=p 2 (1)
 4. Prod3=p Prod5=p 2 ==> Prod2=p 2 (1)
 5. Prod1=p 2 ==> Prod3=p 2 (1)
```

```
 6. IdentComprador=300 1 ==> Prod1=p Prod2=p Prod3=p Prod5=p 1 (1)
 7. IdentComprador=300 Prod1=p 1 ==> Prod2=p Prod3=p Prod5=p 1 (1)
 8. IdentComprador=300 Prod2=p 1 ==> Prod1=p Prod3=p Prod5=p 1 (1)
 9. IdentComprador=300 Prod3=p 1 ==> Prod1=p Prod2=p Prod5=p 1 (1)
10. IdentComprador=300 Prod5=p 1 ==> Prod1=p Prod2=p Prod3=p 1 (1)
```

A última parte apresenta as regras de associação que foram induzidas.

O número precedendo o símbolo = = > indica o suporte da regra, isto é, o número de ítens cobertos por sua premissa. Após a regra está o porcentual de ocorrências do antecedente da regra para os quais o conseqüente da regra é válido, ou seja, a sua confiança. O programa **APRIORI** ordena as regras de acordo com sua confiança e usa o suporte para desempate.

Precedendo as regras estão os números de conjuntos de ítens encontrados para cada suporte considerado.

Assim, a regra 1 deve ser lida como:

A presença do produto 2(antecedente da regra) implicou na presença do produto 5(conseqüente da regra) em três carrinhos, o que representou 100% dos carrinhos em que foi encontrado o produto 2 (3 carrinhos). O suporte da regra é ¾= 75%. Aconfiança da regra é 100%.

5.6 MINERAÇÃO DE DADOS PARA OBTER REGRAS DE ASSOCIAÇÃO

A partir do arquivo de dados históricos, procurou-se elaborar diversos estudos, buscando ao mesmo tempo obter resultados verídicos e verificar o comportamento do algoritmo **APRIORI**, em sua versão programada na linguagem JAVA.

Num primeiro estudo, foi utilizado o arquivo completo, para sentir o comportamento do algoritmo **APRIORI** e conhecer a base de dados do plano de saúde.

A seguir, pareceu interessante selecionar um grupo de serviços de saúde mais conhecido, recaindo a escolha sobre serviços odontológicos, que por fazerem

parte do cotidiano das pessoas, pareceu ser adequado como exemplo didático, para facilitar a compreensão dos resultados obtidos.
Os diversos estudos realizados são a seguir descritos através do registro sistemático dos dados e parâmetros de entrada e dos resultados obtidos, acompanhados da avaliação destes resultados.

5.6.1 TIPOS DE SERVIÇO POR SEXO

Neste estudo procurou-se deixar o **APRIORI** induzir regras de associação entre os tipos de serviço utilizados pelos beneficiários do plano de saúde, considerando também o sexo dos beneficiários.

5.6.1.1 LINHA DE COMANDO

No prompt do DOS foi usado o seguinte comando:
java weka.associations.Apriori -t tipser22.nominal.arff

5.6.1.2 NÚMERO DE REGISTROS NO BANCO DE DADOS

A partir dos 457.832 registros, considerando apenas o tipo do serviço e aglutinando as ocorrências dos diferentes códigos de serviço de mesmo tipo, foram obtidos 97760 registros adequados ao estudo.

5.6.1.3 NÚMERO DE TRANSAÇÕES NO ARQUIVO ARFF

A partir dos 97760 registros, o arquivo no formato ARFF resultou com 5038 transações.

5.6.1.4 TEMPO DE PROCESSAMENTO

5 minutos

5.6.1.5 DADOS DE ENTRADA

```
@relation TipoServMatrSexo.Symbolic
@attribute Matricula {0000001,0000002,.....}
@attribute SexoMasc    {M}
@attribute SexoFemi    {F}
@attribute 00 {P}
.
.
@attribute 90 {P}
@data
0000001,M,?,P,?,P,?,?,P,P,?,?,P,?,?,P,P,?,?,?,P,P,P,?,?,?,?,?,?,?,?,?,?,?,?,P,?,?,?,?,P,?,?,?,?,?,?,?,P,?,?,?,?
.
.
9999999,M,?,P,?,?,?,?,P,?,?,?,?,?,?,?,P,?,?,?,P,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?
```

Observação sobre os dados de entrada:

Optou-se por codificar o sexo masculino como um atributo, e o feminino como outro, para obter regras específicas para cada sexo se ocorressem.

Para obter regras de associação entre tipos de serviço é necessário codificar cada tipo de serviço como um atributo diferente, indicando sua presença(P) ou ausência(?).

5.6.1.6 RESULTADOS OBTIDOS

Minimum support: 0.65

Minimum confidence: 0.9
Number of cycles performed: 8
Generated sets of large itemsets:
Size of set of large itemsets L(1): 4
Size of set of large itemsets L(2): 5
Size of set of large itemsets L(3): 2
Best rules found:
1. 28=P 90=P 3314 ==> 00=P 3306 (1)
2. 28=P 32=P 3415 ==> 00=P 3405 (1)
3. 28=P 4042 ==> 00=P 4008 (0.99)
4. 32=P 3748 ==> 00=P 3706 (0.99)
5. 90=P 3669 ==> 00=P 3614 (0.99)
6. 00=P 32=P 3706 ==> 28=P 3405 (0.92)
7. 00=P 90=P 3614 ==> 28=P 3306 (0.91)
8. 32=P 3748 ==> 28=P 3415 (0.91)
9. 32=P 3748 ==> 00=P 28=P 3405 (0.91)
10. 90=P 3669 ==> 28=P 3314 (0.9)

5.6.1.7 LEGENDA

Legenda de Tipo de Serviço:
00-Consulta Médica
28-Patologia Clínica (exames de laboratório)
32-Radiodiagnóstico
90-Farmácia

5.6.1.8 ANÁLISE DOS RESULTADOS OBTIDOS

Para obter dez regras com confiança 0.9, o suporte mínimo foi 0.65 e foram necessárias oito iterações, pois o **APRIORI** começa tentando obter as dez regras com suporte 1.0 e decrementa em passos de 0.05, até conseguir.

Regra 1: 28=P 90=P 3314 ==> 00=P 3306 (1)

Tradução:

- 100 % das pessoas que usam simultaneamente exames de laboratório(28) e reembolso de medicamentos(90) usam consulta médica(00).

Considerações:

- Em parte é uma regra óbvia porque pelo regulamento do plano de saúde, somente são reembolsados medicamentos acompanhados pela respectiva receita médica, enquanto exames de laboratório podem ser realizados através do plano de saúde apenas por solicitação médica.
- A informação adicional contida nesta regra é a associação do reembolso de medicamentos com exame de laboratório em 3314 das 5038 transações processadas, o que dá o suporte da regra, 3314/5038=0.65 ou 65%. A confiança da regra é obtida dividindo a freqüência do conseqüente da regra pela freqüência do antecedente da regra, 3306/3314=0.9975 ou 99.75%, que foi arredondado para 1 ou 100% pelo programa **APRIORI**.

Regra 2: 28=P 32=P 3415 ==> 00=P 3405 (1)

Tradução:

- 100 % das pessoas que utilizam simultaneamente exames de laboratório(28) e Radiodiagnóstico usam consulta médica(00).

Considerações:

- Regra análoga à regra 1,com radiodiagnóstico no lugar de reembolso de medicamentos

Regra 3: 28=P 4042 ==> 00=P 4008 (0.99)

Tradução:

- 99 % das pessoas que utilizam exames de laboratório(28) usam consulta médica(00).

Considerações:

- É uma regra óbvia porque pelo regulamento do plano de saúde, somente são reembolsados exames de laboratório acompanhados pela requisição receita médica. Por quê 99 % ?
- É um subconjunto da regra 1

Regras 4 a 10:
São variações das regras 1 a 3, porque o algoritmo **APRIORI** faz combinações exaustivas, o que muitas vezes resulta em regras redundantes.

Considerações finais:
- Este estudo serviu para verificar que o **APRIORI** percebe fatos reais, ou seja funciona corretamente, o que fica evidenciado se for considerado que de fato, consultas médicas são o serviço mais freqüentemente prestado por um plano de saúde, pois a maioria dos atendimentos começa com uma consulta médica.
- Não foi gerada nenhuma regra, entre as dez principais, envolvendo o sexo dos beneficiários, porque o uso dos tipos de serviço referenciados ocorre indiferentemente do sexo dos beneficiários. Regras que envolvam o sexo do beneficiário e tipo de serviço terão suporte inferior ao do tipo de serviço.

5.6.2 SERVIÇOS DE ODONTOLOGIA POR SEXO E FAIXA ETÁRIA– NÍVEL 1

Neste estudo procurou-se deixar o **APRIORI** induzir regras de associação entre os serviços odontológicos usados pelos beneficiários do plano de saúde, considerando também o sexo e a faixa etária dos beneficiários.

Para o tratamento de faixa etária, tornou-se necessário gerar o arquivo ARFF com rótulos para as faixas etárias de interesse, porque o algoritmo **APRIORI** trabalha naturalmente com atributos nominais, não com quantidades, além do fato de o arquivo de dados históricos conter idade e não faixa etária.

Foram consideradas as seguintes faixas etárias da tabela a seguir:

Rótulo da Faixa Etária	Faixa Etária em anos
Infância	de 0 a 12
Juventude	de 13 a 20
Adulto	de 21 a 39
Meia Idade	de 40 a 59
Maturidade	de 60 em diante

Para processar apenas serviços odontológicos, filtrou-se o arquivo selecionando apenas os registros com tipo de serviço 87, 88 e 89.

5.6.2.1 LINHA DE COMANDO

No prompt do DOS foi usado o seguinte comando:
java weka.associations.Apriori -t odontologicocompleto.nominal.arff

5.6.2.2 NÚMERO DE REGISTROS NO BANCO DE DADOS

A partir dos 457.832 registros, considerando apenas os códigos de serviços odontológicos, foram obtidos 20170 registros adequados ao estudo

5.6.2.3 NÚMERO DE TRANSAÇÕES NO ARQUIVO ARFF

A partir dos 20170 registros, o arquivo no formato ARFF resultou com 2113 transações.

5.6.2.4 TEMPO DE PROCESSAMENTO

O processamento do programa **APRIORI** durou três minutos

5.6.2.5 DADOS DE ENTRADA

@relation MatrSexoFaixaEtariaCodiServ.Symbolic

@attribute Matricula {0000001,0000002,...}
@attribute Sexo {M,F}
@attribute FaixaEtaria {Adulto,Infancia,Juventude,Maturidade,MeiaIdade}
@attribute 87010011 {P}
@attribute 87010020 {P}
.
.
.
@attribute 89900421 {P}
@attribute 89900430 {P}
@data
0000001,M,MeiaIdade ,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,...
.
.
.
0002113,F,Adulto ,P,?,?,?,?,?,?,?,?,P,?,?,?,?,?,?,?,...

Observação sobre os dados de entrada:
Optou-se por codificar o sexo como um atributo, com dois valores possíveis
Para obter regras de associação entre os diferentes serviços odontológicos, é necessário codificar cada código de serviço como um atributo diferente, indicando sua presença(P) ou ausência(?).

5.6.2.6 RESULTADOS OBTIDOS

Minimum support: 0.15
Minimum confidence: 0.9

Number of cycles performed: 18
Generated sets of large itemsets:
Size of set of large itemsets L(1): 20
Size of set of large itemsets L(2): 33
Size of set of large itemsets L(3): 12
Best rules found:
1. 87010119=P 87010160=P 476 ==> 87010011=P 452 (0.95)
2. Sexo=F 87010119=P 337 ==> 87010011=P 319 (0.95)
3. 87010119=P 639 ==> 87010011=P 600 (0.94)
4. 87010160=P 87011450=P 407 ==> 87010011=P 381 (0.94)
5. FaixaEtaria=MeiaIdade 87010119=P 344 ==> 87010011=P 322 (0.94)
6. 87010119=P 88900010=P 348 ==> 87010011=P 325 (0.93)
7. 87010160=P 88900010=P 497 ==> 87010011=P 460 (0.93)
8. Sexo=F 87010160=P 558 ==> 87010011=P 511 (0.92)
9. 87011450=P 522 ==> 87010011=P 478 (0.92)
10. 87010160=P 87011441=P 392 ==> 87010011=P 358 (0.91)

5.6.2.7 LEGENDA

Legenda de Código de Serviço:
87010011-Consulta inicial
87010119-Radiografia Periapical
87010160-Profilaxia (polimento coronário e tártaro)
87011441-Restauração fotopolimerizável duas faces
87011450-Restauração fotopolimerizável três faces
88900010-Consulta

5.6.2.8 ANÁLISE DOS RESULTADOS OBTIDOS

Para obter dez regras com confiança 0.9, o suporte mínimo foi 0.15 e foram necessárias dezoito iterações, pois o **APRIORI** começa tentando obter as dez

regras com suporte 1.0 e decrementa em passos de 0.05, até conseguir as dez regras ou até atingir o mínimo de 0.10, o que ocorrer primeiro.

Regra 1: 87010119=P 87010160=P 476 ==> 87010011=P 452 (0.95)

Tradução:

- 95 % das ocorrências de uso simultâneo de Radiografia Periapical e Profilaxia (polimento coronário e remoção de tártaro) é acompanhada de Consulta inicial

Considerações:

- Em parte é uma regra óbvia porque é comum o registro e cobrança de uma consulta inicial em tratamento dentário, mesmo quando ocorre apenas uma profilaxia, e/ou uma ou mais radiografias para avaliar a necessidade de restauração.
- O suporte desta regra foi: 476/2113=0.2252 ou 22.52 %
- A confiança da regra foi: 452/476=0.9495 ou 94.95 %

Regra 2: Sexo=F 87010119=P 337 ==> 87010011=P 319 (0.95)

Tradução:

- 95 % das Radiografia Periapical em mulheres estão associadas a Consulta inicial

Considerações:

- É comum na consulta inicial ser feita radiografia periapical.

Regra 3: 87010119=P 639 ==> 87010011=P 600 (0.94)

Tradução: 94 % das Radiografia Periapical implica em Consulta inicial

Considerações:

- Aparentemente é uma regra com pouca informação, além de ser um subconjiunto da regra 1.

Regras 4 a 10:

Algumas são variações das regras 1 a 3, porque o algoritmo **APRIORI** faz combinações exaustivas, o que muitas vezes resulta em regras redundantes. Outras são diferentes, mas com pouca informação adicional.

Considerações finais:

- Também neste estudo o **APRIORI** teve um desempenho consistente, pois gerou regras plausíveis, sempre com a tendência para informar o trivial. Por quê isto? Porque o **APRIORI** se baseia na maior freqüência de ocorrência da associação dos fatos.
- Uma pergunta que surge é: o quê fazer para obter regras de associação referentes a fatos menos freqüentes, que talvez contenham alguma informação nova? Nos estudos que seguem serão feitas prospecções neste sentido e relatados os resultados obtidos. Uma dica simples é a seguinte: para minerarmos apenas serviços odontológicos, abandonamos todos os demais serviços, aceitando apenas tipos de serviço 87, 88 e 89. Isto se deveu ao fato de o **APRIORI** calcular o suporte dividindo uma freqüência pelo número total de registros do arquivo em estudo, que pode ser um subconjunto do arquivo completo.
- Outra pergunta é: quantas regras podem ser obtidas mantendo a confiança 0.9 das mesmas? Para saber, basta processar novamente, especificando o parâmetro N com um valor grande.

5.6.3 SERVIÇOS DE ODONTOLOGIA POR SEXO E FAIXA ETÁRIA – NÍVEL 1 - 50 REGRAS

5.6.3.1 DESCRIÇÃO

Foi utilizado aqui o mesmo arquivo ARFF do estudo 5.6.2, diferindo no objetivo de responder a pergunta: quantas regras podem ser obtidas com confiança mínima 0.9?

Para responder a pergunta, foi submetido um processamento especificando o parâmetro N com valor 50, ou seja solicitando a geração de até cinqüenta regras. Houve uma surpresa no resultado, que faz todo sentido, mas que tornou necessário usar o parâmetro M, para especificar o suporte mínimo.

5.6.3.2 LINHA DE COMANDO

No prompt do DOS foram usados os seguintes comandos:
No primeiro processamento:
java weka.associations.Apriori -t odontologicocompleto.nominal.arff –N 50
O parâmetro –N 50 informa ao **APRIORI** para tentar gerar até 50 regras
No segundo processamento:
java weka.associations.Apriori -t odontologicocompleto.nominal.arff –N 50 –M 0.15
O parâmetro –M 0.15 estabelece 0.15 como limite inferior para o suporte das regras.

5.6.3.3 NÚMERO DE REGISTROS NO BANCO DE DADOS

O banco de dados para este estudo resultou com 20170 registros

5.6.3.4 NÚMERO DE TRANSAÇÕES NO ARQUIVO ARFF

O arquivo ARFF resultou com 2113 transações.

5.6.3.5 TEMPO DE PROCESSAMENTO

Aproximadamente três minutos em ambos os casos.

5.6.3.6 DADOS DE ENTRADA

O mesmo do estudo anterior.

Observação sobre os dados de entrada:

A mesma do estudo anterior.

5.6.3.7 RESULTADOS OBTIDOS

No processamento com apenas o parâmetro –N 50 foram obtidas 44 regras, as quais não serão apresentadas aqui, pela sua extensão e por conter muitas regras redundantes, sem utilidade efetiva.

No processamento com parâmetros –N 50 e –M 0.15 foram obtidas 11 regras:

Minimum support: 0.15

Minimum confidence: 0.9

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 20

Size of set of large itemsets L(2): 33

Size of set of large itemsets L(3): 12

Best rules found:

1. 87010119=P 87010160=P 476 ==> 87010011=P 452 (0.95)
2. Sexo=F 87010119=P 337 ==> 87010011=P 319 (0.95)
3. 87010119=P 639 ==> 87010011=P 600 (0.94)
4. 87010160=P 87011450=P 407 ==> 87010011=P 381 (0.94)
5. FaixaEtaria=MeiaIdade 87010119=P 344 ==> 87010011=P 322 (0.94)
6. 87010119=P 88900010=P 348 ==> 87010011=P 325 (0.93)
7. 87010160=P 88900010=P 497 ==> 87010011=P 460 (0.93)
8. Sexo=F 87010160=P 558 ==> 87010011=P 511 (0.92)
9. 87011450=P 522 ==> 87010011=P 478 (0.92)
10. 87010160=P 87011441=P 392 ==> 87010011=P 358 (0.91)
11. 87010160=P 1036 ==> 87010011=P 933 (0.9)

5.6.3.8 LEGENDA

A mesma do estudo anterior.

5.6.3.9 ANÁLISE DOS RESULTADOS OBTIDOS

Por que um processamento gerou 44 regras e o outro 11 regras?
Para obter até 50 regras com confiança 0.9, no processamento com –N 50 o suporte mínimo não foi especificado, então o algoritmo **APRIORI** decrementou o suporte mínimo em passos de 0.05 até 0.10 (limite padrão do algoritmo),e mesmo assim só conseguiu gerar 44 regras.
No processamento com –N 50 –M 0.15, foram geradas apenas onze regras, sendo as dez primeiras exatamente as dez regras obtidas no estudo anterior (6.4.2), pois com o mesmo suporte mínimo 0.15, ocorreram as mesmas dezoito iterações, enquanto com suporte mínimo 0.10 ocorreram dezenove iterações, não sendo as dez primeiras regras exatamente as mesmas.

Considerações:
O **APRIORI** é muito sensível ao parâmetro suporte mínimo, que pode ser manipulado no caso de escassez de regras. Convém ter em mente que as regras obtidas com suporte muito pequeno, não são muito confiáveis, justamente por representarem associações pouco freqüentes, apesar de serem as mais freqüentes dos dados em estudo.
Considerações finais:
E afinal, não se obtém as 50 regras?
Sim, basta reduzir a confiança mínima exigida, de 0.9 para 0.5, adotando a seguinte linha de comando:
java weka.associations.Apriori -t odontologicocompleto.nominal.arff –N 50 –M 0.15 –C 0.5

Com confiança 0.8 foram obtidas 17 regras, com 0.7 obteve-se 33 regras, com 0.6 obteve-se 39 regras, com 0.5 obteve-se 50 regras,

Outra opção para obter as 50 regras foi adotar suporte mínimo 0.10 e confiança 0.8.

5.6.4 SERVIÇOS DE ODONTOLOGIA POR SEXO E FAIXA ETÁRIA - NÍVEL 2 - CONFIANÇA 0.5

5.6.4.1 DESCRIÇÃO

Apenas serviços de odontologia, excluídos os serviços 87010011 87010119 87010160 87011441 87011450 88900010, que haviam aparecido nas regras obtidas no processamento com todos os serviços de odontologia (itens 5.6.2 e 5.6.3).

5.6.4.2 LINHA DE COMANDO

No prompt do DOS foi usado o seguinte comando:
java weka.associations.Apriori -t odontologiconivel2.nominal.arff –C 0.5

5.6.4.3 NÚMERO DE REGISTROS NO BANCO DE DADOS

A partir dos 457.832 registros, considerando apenas os serviços odontológicos, excluindo os serviços citados acima, foram obtidos 14925 registros adequados ao estudo

5.6.4.4 NÚMERO DE TRANSAÇÕES NO ARQUIVO ARFF

A partir dos 14925 registros, o arquivo no formato ARFF resultou com 2058 transações.

5.6.4.5 TEMPO DE PROCESSAMENTO

Aproximadamente três minutos

5.6.4.6 DADOS DE ENTRADA

O mesmo formato do estudo 5.6.2 e 5.6.3:
@relation MatrSexoFaixaEtariaCodiServ.Symbolic

@attribute Matricula {0000001,0000002,...}
@attribute Sexo {M,F}
@attribute FaixaEtaria {Adulto,Infancia,Juventude,Maturidade,MeiaIdade}
@attribute 87010020 {P}
@attribute 87010038 {P}
.
.
.
@attribute 89900421 {P}
@attribute 89900430 {P}
@data
0000001,M,MeiaIdade ,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,?,...
.
.
.
0002058,F,Adulto ,P,?,?,?,?,?,?,?,?,P,?,?,?,?,?,?,...

Observação sobre os dados de entrada:

Optou-se por codificar o sexo como um atributo, com dois valores possíveis.

Para obter regras de associação entre os diferentes serviços odontológicos, é necessário codificar cada código de serviço como um atributo diferente, indicando sua presença(P) ou ausência(?).

5.6.4.7 RESULTADOS OBTIDOS

Minimum support: 0.1
Minimum confidence: 0.5
Number of cycles performed: 19
Generated sets of large itemsets:
Size of set of large itemsets L(1): 28
Size of set of large itemsets L(2): 22
Best rules found:

1. 88810240=P 292 ==> 88810127=P 213 (0.73)
2. 87010615=P 323 ==> FaixaEtaria=MeiaIdade 206 (0.64)
3. FaixaEtaria=Adulto 416 ==> Sexo=F 246 (0.59)
4. 87011433=P 604 ==> Sexo=F 337 (0.56)
5. 88800032=P 589 ==> FaixaEtaria=MeiaIdade 328 (0.56)
6. 88800032=P 589 ==> Sexo=F 318 (0.54)
7. 88100022=P 421 ==> 88100030=P 220 (0.52)
8. 87011433=P 604 ==> FaixaEtaria=MeiaIdade 310 (0.51)
9. 88100030=P 486 ==> Sexo=M 248 (0.51)
10. 88100022=P 421 ==> Sexo=M 214 (0.51)

5.6.4.8 LEGENDA

Serviços odontológicos:
87010615 – Raspagem supra e sub-gengival – profilaxia e polimento
87011433 – Restauração fotopolimerizável uma face

88100022 – Restauração amálgama uma face
88100030 - Restauração amálgama duas faces
88800032 – Radiografia periapical
88810127 – Faceta em resina
88810240 – Restauração resina fotopolimerizável

Faixa etária:
Adulto – 21 a 39 anos
MeiaIdade – 40 a 59 anos

5.6.4.9 ANÁLISE DOS RESULTADOS OBTIDOS

Regra 1: 88810240=P 292 ==> 88810127=P 213 (0.73)
Tradução:

- 73 % dos casos de utilização de Restauração resina fotopolimerizável também utlizam Faceta em resina. Significa que restauração usando resina fotopolimerizável em geral é aplicada em facetas em resina, geralmente por motivos estéticos.

Regra 2: 87010615=P 323 ==> FaixaEtaria=MeiaIdade 206 (0.64)

- 64 % dos casos de Raspagem supra e sub-gengival ocorre em beneficiários de meia idade(40 a 59 anos). Efetivamente significa remoção de tártaro, o que ocorre com mais freqüência a medida que a idade aumenta.

Regra 3: FaixaEtaria=Adulto 416 ==> Sexo=F 246 (0.59)
Tradução:

- 59 % dos atendimentos da faixa etária adulto são para o sexo feminino. Na faixa etária adulta, as mulheres usam mais serviços odontológicos que os homens. Este fato é conhecido no ramo da odontologia, porque as mulheres se preocupam mais com a estética e com o hálito do que os homens.

Regra 4: 87011433=P 604 ==> Sexo=F 337 (0.56)

Tradução:

- 56 % das Restauração fotopolimerizável uma face são para o sexo feminino. Nesta regra permaneceu o predomínio feminino, independente de idade.

Considerações:

Embora estas regras tenham relativamente menos suporte que as do estudo 6.4.3, ainda contém significado real e útil.

Considerações finais:

Deve ser observado que o arquivo arff deste estudo, resultou da eliminação dos serviços participantes de regras no estudo 6.4.3

O processamento sem nenhum parâmetro, não gerou regras, o que tornou necessário especificar algum parâmetro, sendo escolhido o fator de confiança, experimentalmente com valor 0.5.

O maior fator de confiança (C) obtido foi 0.73; por isto nenhuma regra havia sido gerada com C=0.9

O suporte para o **APRIORI** conseguir gerar dez regras foi 0.1, seu mínimo padrão.

5.6.5 SERVIÇOS DE ODONTOLOGIA POR SEXO E FAIXA ETÁRIA – NÍVEL 3 – CONFIANÇA 0.2

Apenas serviços de odontologia, excluídos além dos serviços 87010011 87010119 87010160 87011441 87011450 88900010, integrantes das regras obtidas no processamento com todos os serviços de odontologia (itens 6.4.2 e 6.4.3), também os serviços 87010615 87011433 88100022 88100030 88800032 88810127 88810240 , integrantes das regras obtidas no estudo do item 6.4.4.

5.6.5.1 LINHA DE COMANDO

No prompt do DOS foi usado o seguinte comando:
java weka.associations.Apriori -t odontologiconivel3.nominal.arff –C 0.2

5.6.5.2 NÚMERO DE REGISTROS NO BANCO DE DADOS

A partir dos 457.832 registros, considerando apenas os serviços odontológicos e excluindo ocorrências dos serviços acima citados, foram obtidos 11638 registros adequados ao estudo

5.6.5.3 NÚMERO DE TRANSAÇÕES NO ARQUIVO ARFF

A partir dos 11638 registros, o arquivo no formato ARFF resultou com 1957 transações.

5.6.5.4 TEMPO DE PROCESSAMENTO

Aproximadamente três minutos

5.6.5.5 DADOS DE ENTRADA

@relation MatrSexoFaixaEtariaCodiServ.Symbolic
@attribute Matricula {0000001,... }
@attribute Sexo {M,F}
@attribute FaixaEtaria {Adulto,Infancia,Juventude,Maturidade,MeiaIdade}
@attribute 87010020 {P}
.
.
@data
0000001,M,MeiaIdade ,?,?,?,?,?,?...

.

.

.

Observação sobre os dados de entrada:

Optou-se por codificar o sexo masculino como um atributo, e o feminino como outro atributo, para obter regras específicas para cada sexo se ocorressem.

Para obter regras de associação entre serviços é necessário codificar cada serviço como um atributo diferente, indicando sua presença(P) ou ausência(?).

5.6.5.6 RESULTADOS OBTIDOS

Minimum support: 0.1

Minimum confidence: 0.2

Number of cycles performed: 19

Generated sets of large itemsets:

Size of set of large itemsets L(1): 21

Size of set of large itemsets L(2): 5

Best rules found:

1. FaixaEtaria=Adulto 374 ==> Sexo=F 227 (0.61)
2. FaixaEtaria=Juventude 368 ==> Sexo=F 199 (0.54)
3. 88120023=P 379 ==> Sexo=F 203 (0.54)
4. Sexo=M 921 ==> FaixaEtaria=MeiaIdade 472 (0.51)
5. FaixaEtaria=MeiaIdade 925 ==> Sexo=M 472 (0.51)
6. FaixaEtaria=MeiaIdade 925 ==> Sexo=F 453 (0.49)
7. Sexo=F 1035 ==> FaixaEtaria=MeiaIdade 453 (0.44)
8. Sexo=F 1035 ==> FaixaEtaria=Adulto 227 (0.22)

5.6.5.7 LEGENDA

Serviços:

88120023- restauração com resina foto-polimerizável de 1 face

Faixa etária:
Juventude: 13 a 20 anos
Adulto: 21 a 39 anos
MeiaIdade: 40 a 59 anos

5.6.5.8 ANÁLISE DOS RESULTADOS OBTIDOS

Para obter oito regras com confiança 0.2, o suporte mínimo foi 0.10 e foram necessárias 19 iterações, pois o **APRIORI** começa tentando obter as dez regras com suporte 1.0 e decrementa em passos de 0.05, até conseguir ou atingir o suporte 0.10.

Considerações:
Neste nível 3 de aprofundamento da mineração de regras de associação entre serviços odontológicos, já não restaram serviços de uso predominante, assumindo maior presença nas associações a faixa etária e o sexo dos beneficiários, aparecendo apenas um serviço numa única regra.
Chamam a atenção as regras 5 e 6, pela coincidência de ambas terem a faixa etária meia idade no antecedente da regra, enquanto a 5 tem o sexo masculino, a 6 tem o sexo feminino no conseqüente da regra. Pode-se ver que 925 = 472+453 e 1.00 = 0.51 + 0.49
O que mostra bem o conceito de confiança de uma regra, além de aferir a correção do programa **APRIORI**.
5. FaixaEtaria=MeiaIdade 925 ==> Sexo=M 472 (0.51)
6. FaixaEtaria=MeiaIdade 925 ==> Sexo=F 453 (0.49)

Considerações finais:
Para um suporte mínimo de 0.10, considerado baixo, após eliminar os serviços participantes das regras obtidas nos estudos 6.4.2, 6.4.3 e 6.4.4, mesmo para uma confiança de 0.2, também considerada baixa, somente uma regra gerada envolveu um serviço.

5.6.6 SERVIÇOS DE ODONTOLOGIA POR SEXO E FAIXA ETÁRIA – NÍVEL 4 – CONFIANÇA 0.1

Neste estudo partiu-se do arquivo usado no nível 3 (item 5.6.5) eliminando o serviço 88120023 que participou de uma das regras.

5.6.6.1 LINHA DE COMANDO

No prompt do DOS foi usado o seguinte comando:
java weka.associations.Apriori -t odontologiconivel4.nominal.arff -C 0.1

5.6.6.2 NÚMERO DE REGISTROS NO BANCO DE DADOS

A partir dos 457.832 registros, considerando apenas os serviços odontológicos e excluindo ocorrências dos serviços acima citados, foram obtidos 11257 registros adequados ao estudo.

5.6.6.3 NÚMERO DE TRANSAÇÕES NO ARQUIVO ARFF

A partir dos 11257 registros, o arquivo no formato ARFF resultou com 1950 transações.

5.6.6.4 TEMPO DE PROCESSAMENTO

Aproximadamente três minutos

5.6.6.5 DADOS DE ENTRADA

@relation MatrSexoFaixaEtariaCodiServ.Symbolic
@attribute Matricula {0000001,... }

@attribute Sexo {M,F}
@attribute FaixaEtaria {Adulto,Infancia,Juventude,Maturidade,MeiaIdade}
@attribute 87010020 {P}
...
@data
0000001,M,MeiaIdade ,?,?,?,?,?,?,?,...
...

Observação sobre os dados de entrada:
Optou-se por codificar o sexo masculino como um atributo, e o feminino como outro atributo, para obter regras específicas para cada sexo se ocorressem.

5.6.6.6 RESULTADOS OBTIDOS

Minimum support: 0.1
Minimum confidence: 0.1
Number of cycles performed: 19
Generated sets of large itemsets:
Size of set of large itemsets L(1): 20
Size of set of large itemsets L(2): 4
Best rules found:
1. FaixaEtaria=Adulto 372 ==> Sexo=F 225 (0.6)
2. FaixaEtaria=Juventude 366 ==> Sexo=F 198 (0.54)
3. Sexo=M 920 ==> FaixaEtaria=MeiaIdade 472 (0.51)
4. FaixaEtaria=MeiaIdade 924 ==> Sexo=M 472 (0.51)
5. FaixaEtaria=MeiaIdade 924 ==> Sexo=F 452 (0.49)
6. Sexo=F 1031 ==> FaixaEtaria=MeiaIdade 452 (0.44)
7. Sexo=F 1031 ==> FaixaEtaria=Adulto 225 (0.22)
8. Sexo=F 1031 ==> FaixaEtaria=Juventude 198 (0.19)

5.6.6.7 LEGENDA

Faixa etária:
Juventude: 13 a 20 anos
Adulto: 21 a 39 anos
MeiaIdade: 40 a 59 anos

5.6.6.8 ANÁLISE DOS RESULTADOS OBTIDOS

Para obter oito regras com confiança 0.1, o suporte mínimo foi 0.10 e foram necessárias 19 iterações, pois o **APRIORI** começa tentando obter as dez regras com suporte 1.0 e decrementa em passos de 0.05, até conseguir ou atingir o suporte 0.10.

Considerações:
Neste nível 4 de aprofundamento da mineração de regras de associação entre serviços odontológicos, já não restaram serviços de uso predominante, assumindo presença nas associações apenas a faixa etária e o sexo dos beneficiários.

Considerações finais:
Para um suporte mínimo de 0.10, considerado baixo, após eliminar os serviços participantes das regras obtidas nos estudos 5.6.2, 5.6.3, 5.6.4 e 5.6.5, mesmo para uma confiança de 0.1, também considerada baixa, nenhuma das regras geradas envolveu um serviço.
A princípio parece o fim desta mineração, mas se reduzirmos um pouco mais o suporte, o que acontece? O próximo estudo mostra, com alguma surpresa.

5.6.7 SERVIÇOS DE ODONTOLOGIA POR SEXO E FAIXA ETÁRIA – NÍVEL 4 – SUPORTE 0.05

Neste estudo reduziu-se o suporte mínimo de 0.10 do estudo anterior para 0.05, para verificar a conseqüência. O objetivo é verificar se ainda é possível obter regras de alguma utilidade envolvendo serviços, reduzindo extremamente o suporte mínimo.

5.6.7.1 LINHA DE COMANDO

No prompt do DOS foi usado o seguinte comando:
java weka.associations.Apriori -t odontologiconivel4.nominal.arff -C 0.1 -M 0.05

5.6.7.2 NÚMERO DE REGISTROS NO BANCO DE DADOS

O mesmo do estudo anterior, 11257 registros adequados ao estudo.

5.6.7.3 NÚMERO DE TRANSAÇÕES NO ARQUIVO ARFF

O mesmo do estudo anterior, 1950 transações.

5.6.7.4 TEMPO DE PROCESSAMENTO

Aproximadamente três minutos.

5.6.7.5 DADOS DE ENTRADA

O mesmo do estudo anterior.

Observação sobre os dados de entrada:

A mesma do estudo anterior.

5.6.7.6 RESULTADOS OBTIDOS

Minimum support: 0.05
Minimum confidence: 0.01
Number of cycles performed: 20
Generated sets of large itemsets:
Size of set of large itemsets L(1): 43
Size of set of large itemsets L(2): 75
Size of set of large itemsets L(3): 2
Best rules found:

1. 87010852=P 87010917=P 117 ==> 87010844=P 103 (0.88)
2. 87010844=P 87010917=P 125 ==> 87010852=P 103 (0.82)
3. 87010852=P 167 ==> 87010844=P 134 (0.8)
4. FaixaEtaria=MeiaIdade 87010852=P 126 ==> 87010844=P 101 (0.8)
5. 87010917=P 156 ==> 87010844=P 125 (0.8)
6. FaixaEtaria=MeiaIdade 87010844=P 129 ==> 87010852=P 101 (0.78)
7. 88600041=P 164 ==> FaixaEtaria=MeiaIdade 128 (0.78)
8. 87010917=P 156 ==> FaixaEtaria=MeiaIdade 120 (0.77)
9. 87010844=P 87010852=P 134 ==> 87010917=P 103 (0.77)
10. 88600106=P 207 ==> FaixaEtaria=MeiaIdade 159 (0.77)

5.6.7.7 LEGENDA

Serviços:
87010844-Coroa metalo cerâmica
87010852-Coroa provisória – por elemento
87010917-Núcleo metálico fundido
88600041-Núcleo metálico fundido
88600106-Coroa de jaqueta metalo-cerâmica

Faixa Etária:

MeiaIdade - 40 a 59 anos

5.6.7.8 ANÁLISE DOS RESULTADOS OBTIDOS

Regra 1: 87010852=P 87010917=P 117 ==> 87010844=P 103 (0.88)

Tradução:

- 88 % das pessoas que colocam Coroa provisória–por elemento, junto com Núcleo metálico fundido, colocam também Coroa metalo cerâmica.

Considerações:

Esta regra corresponde a um fato, porque em tratamento protético de dentes que apresentam sua porção coronal destruída, se faz necessária a colocação de um núcleo metálico fundido. Na mesma consulta é colocada uma coroa provisória para proteção do remanescente dental. Posteriormente, é realizado em várias seções o tratamento para moldagem, verificação de cor e ajuste da forma da coroa definitiva metalo cerâmica.

O suporte desta regra, 117/1950 = 0.06 ou 6% é baixo, o que significa apenas que o antecedente da regra ocorre relativamente pouco nas transações lidas pelo **APRIORI**.

Cabe ainda lembrar que a confiança da regra foi 0.88, ou seja, 88 % das coroas provisórias associadas a núcleo metálico fundido, resultaram em coroa permanente. De fato, algumas pessoas não põem a coroa permanente, porque é relativamente mais cara, embora mais durável.

Regra 2: 87010844=P 87010917=P 125 ==> 87010852=P 103 (0.82)

Tradução:

- 82 % das pessoas que colocam Coroa metalo cerâmica também colocam Coroa provisória–por elemento, junto com Núcleo metálico fundido.

Considerações:

É a mesma regra 1, alternando a posição dos serviços.

Regra 3: 87010852=P 167 ==> 87010844=P 134 (0.8)

Tradução:

- 80 % das pessoas que colocam Coroa provisória–por elemento, colocam também Coroa metalo cerâmica.

Considerações:

É um subconjunto das regras 1 ou 2.

Regra 4: FaixaEtaria=MeiaIdade 87010852=P 126 ==> 87010844=P 101 (0.8)

Tradução:

- 80 % das pessoas de meia idade(40 a 59 anos) que utilizam Coroa provisória–por elemento (provisória), adotam Coroa metalo cerâmica(permanente).

Regras 5 a 9:

- São variações das regras 1 a 4

Regra 10: 88600106=P 207 ==> FaixaEtaria=MeiaIdade 159 (0.77)

Tradução:

- 77 % das pessoas que utilizam de Coroa de jaqueta metalo-cerâmica são de meia idade.

Considerações:

É análoga à regra 4.

Considerações finais:

A redução do suporte mínimo, não invalida os resultados obtidos, apenas trata-se de definir se as regras obtidas são ou não de interesse para algum objetivo em questão.

6 CONCLUSÕES E RECOMENDAÇÕES

Regras de associação são uma classe simples e natural de regularidades em bases de dados, úteis em várias análises e tarefas de previsão.

O estudo realizado sobre a literatura relativa ao algoritmo **APRIORI,** e os processamentos realizados com dados reais de um plano de saúde, levam às conclusões apresentadas abaixo.

O algoritmo **APRIORI** apresentou como resultado regras de associação consideradas pertinentes, porque apontaram fatos identificados imediatamente pelos especialistas do plano de saúde, como correspondentes à realidade. Apenas algumas variáveis do plano de saúde foram consideradas nos processamentos feitos. Há muitos estudos que podem ser realizados se forem consideradas outras variáveis como tipo do plano (A, B, C, D, E), região do domicílio, além de considerar cada tipo de serviço separadamente, como foi feito neste trabalho com serviços odontológicos

Durante os experimentos realizados com o **APRIORI**, percebeu-se que não basta aumentar o número de regras, ou reduzir o suporte mínimo das regras visando obter mais informação, ou ainda obter regras que incluam um determinado grupo de códigos. Isto porque o **APRIORI** começa a induzir regras redundantes, que estão corretas, mas não acrescentam informação. Uma estratégia adotada foi

desconsiderar, em estudos sucessivos, os códigos de serviços presentes nas regras de associação geradas, que foram, evidentemente os mais numerosos. Outra estratégia foi fornecer ao **APRIORI** um arquivo contendo apenas os códigos de serviço de um tipo, por exemplo, serviços odontológicos, visando obter regras de associação pertinentes a estes serviços.

O algoritmo **APRIORI** é bastante sensível aos parâmetros suporte mínimo e confiança mínima, o que foi constatado no estudo 5.4.3 e 5.4.4. Caso sejam obtidas poucas regras em um estudo, esses dois parâmetros podem ser variados de forma combinada, até serem obtidas regras na quantidade desejada e com o maior suporte mínimo possível.

Por tudo o que foi exposto acima e pela convivência com o algoritmo **APRIORI**, pode-se concluir que o mesmo funciona, tendo o seu potencial de utilidade, pelos resultados objetivos que apresenta.

Há diversos algoritmos sucessores do **APRIORI**, como **AprioriTid**, **SETM, AprioriHybrid** [AGR94], **Cumulate** e **EstMerge** [SRI95b], **GSP** [SRI95a], **AprioriSome** e **AprioriAll** e **DynamicSome** [AGR95], **DHP** (Direct Hashing and Pruning) [PAR95], todos mais rápidos que o primeiro. É recomendável o estudo destes algoritmos, pela evolução que representam e porque algoritmos mais rápidos são sempre importantes, ainda mais se for considerado que as base de dados estão cada vez maiores.

O **APRIORI** não trata atributos quantitativos, apenas categóricos, o que levou ao surgimento de algoritmos que tratam ambos os tipos de atributos [SRI96]. É recomendável o estudo deste problema, porque simplifica a preparação de dados, se comparado a forçar o APRIORI a tratar atributos quantitativos como categóricos, como foi feito ao longo dos diversos processamentos realizados neste trabalho.

7 ANEXOS

7.1 BIBLIOGRAFIA

[AGR93] Agrawal, Rakesh; Imielinski, Tomasz; Swami; Arun. Mining association rules between sets of items in large databases. In Proceedings. of ACM SIGMOD conference on Management of Data, pages 207-216. Washington, D.C., May 1993.

[AGR94] Agrawal, Rakesh; Swami; Arun. Fast Algorithms for Mining Association Rules. In Proceedings of the 20th VLDB Conference, Santiago, Chile, september 1994.

[AGR95] Agrawal, Rakesh; Srikant; Ramakrishnan. Mining Sequential Patterns. In Proceedings of the 11th Int´l Conference on Data Engineering,Taipei, March 1995.

[AGR96] Agrawal, Rakesh et al. The Quest Data Mining System. Proceedings of the 2nd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, august, 1996.

[AMS96] Agrawal, Rakesh; Manilla, Heikki; Srikant; Ramakrishnan; Toivonen, Hannu; Verkamo, A. Inkeru. Fast Dicovery of Association Rules. In Advances in

Knowledge Discovery and Data Mining, U.M.Fayyad; G.Piatesky-Shapiro, P.Smyth and R.Uthurusamy (editors), MIT Press, 1996 pp 307-328.
[BAY99] Bayardo Jr, Roberto J.; Agrawal, Rakesh. Mining the Most Interesting Rules. In Proceedings of the Fifth ACM SIGKDD Int´l Conf. on Knowledge Discovery and Data Mining, 1999.

[BER97] Berson, Alex. Data Warehousing, datamining, and OLAP. ISBN 0-07-006272-2. USA, McGraw-Hill, 1998.

[CHE95] Chen, M.S; Park,J.S.; Yu, P.S.. Data Mining: An Overview from Database Perspective. IEEE Transactions on Knowledge and Data Eng., 8(6):866-883, December 1996.

[FAU94] Fausett, Laurene V.. Fundamentals of neural networks: architetures, algorithms, and applications. ISBN 0-13-334186-0, Prentice-Hall Inc, 1994.

[FAY96] Fayyad, Usama M.et al. Advances in knowledge discovery and data mining. ISBN 0-262-56097-6, MIT Press, 1996.

[FUL94] Fu, Limin. Neural Networks in Computer Intelligence. ISBN 0-07-911817-8, McGraw-Hill, 1994.

[GIA98] Giarratano, Joseph C.; Riley, Gary. Expert Systems Principles and Programming. ISBN 0-534-95053-1, PWS Publishing Company, 1998

[GRO98] Groth, Robert. Data Mining: a hands-on approach for business professionals. ISBN 0-13-756412-0. New Jersey, Prentice Hall, 1998

[JAC99] Jackson, Peter, Introduction to Expert Systems. ISBN 0-201-87686-8. USA, Addison-Wesley, 1999.

[KIM98] Kimball, Ralph. Data Warehouse Toolkit. ISBN 85-346-0817-2. São Paulo, Makron Books, 1998.

[MAR91] Martin, James; McLure, Carma. Técnicas estruturadas e Case. McGraw-Hill, 1991.

[PAR95] Park,J.S; Chen, M.S.; Yu, P.S.. An Effective Hash-Based Algorithm for Mining Association Rules. In Proceedings of the of ACM SIGMOD Conference on Management of Data, San Jose, California, may 1995.

[PUR00] Purdom,Paul; Gucht, Dirk Van. Average Case Performance of the Apriori Algorithm. In Proceedings of the 6th Int'l Symposium on Artificial Intelligence and Mathematics, Jan. 5-7, 2000 in Ft. Lauderdale, Florida.

[QUI93] Quinlan, John Ross. C4.5 : programs for machine learning. ISBN 1-55860-238-0. San Mateo, CA, Morgan Kaufmann Publishers, 1993.

[RIC94] Rich, Elaine; Knight, Kevin. Inteligência Artificial. Makron books, 1994.

[SRI95a] Srikant, Ramakrishnan; Agrawal, Rakesh. Mining Sequential Patterns: Generalizations and Performance Improvements. Research Report RJ 9994, IBM Almaden Research Center, San Jose, California, december 1995.

[SRI95b] Srikant, Ramakrishnan; Agrawal, Rakesh. Mining Generalized Association Rules. In Proceedings of the 21th VLDB, Zurich, Swizerland, september 1995.

[SRI96] Srikant, Ramakrishnan; Agrawal, Rakesh. Mining Quantitative Association Rules in Large Relational Tables. In Proceedings of the of ACM SIGMOD conference on Management of Data, Montreal, Canada, june 1996.

[THO98] Thomas, Shiby; Sarawagi, Sunita. Mining Generalized Association Rules and Sequential Patterns Using SQL Queries. In American Association for Artificial Intelligence (www.aaai.org), 1998.

[TIW00] Tiwana, Amrit. The Knowledge Management Toolkit. ISBN 0-13-012853-8. USA, Prentice Hall Inc, 2000.

[WIT00] Witten, Ian H.; Frank, Eibe. Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations. ISBN 1-55860-552-5. USA, CA, Morgan Kaufmann Publishers, 2000.

7.2 LISTA DE ENDEREÇOS NA INTERNET

Muitos recursos para descoberta de conhecimento e mineração de dados, incluindo programas, arquivos de dados, e publicações estão disponíveis via internet. Abaixo vão relacionados diversos endereços:

ftp://ftp.stams.strath.ac.uk – diretório /pub/statlog, contém programas e arquivos de dados referentes ao projeto statlog, sobre aprendizado por máquina e redes neurais.

http://lib.stat.cmu.edu/datasets - Coleção de arquivos de dados

http://www.almaden.ibm.com/cs/quest - Projeto Quest da IBM. Excelentes artigos sobre regras de associação e de classificação.

http://www.cs.bham.ac.uk/~anp/software.html – programas gratuitos para mineração de dados e links

http://www.cs.waikato.ac.nz/ml/weka – código fonte java do projeto weka – Waikato Environment for Knowledge Analysis.

http://www.cse.unsw.edu.au/~quinlan – código fonte C do algoritmo C4.5 de árvore de decisão, autoria de Ross Quinlan

http://www.exclusiveore.com/index.html – site comercial e informativo sobre mineração de dados

http://www.ics.uci.edu/AI/ML/Machine-Learning.html – Universidade da California Irvine; contém arquivos de dados comumente usados para testar algoritmos de aprendizado por máquina.

http://www.isl.co.uk/topclem.html - SPSS Clementine Data Mining System

http://wwwipd.ira.uka.de/~prechelt/FAQ/neural-net-faq.html – muito bom para tirar dúvidas (frequently asked questions) sobre redes neurais, com muitos links.

http://www.kdnuggets.com/index.html - Guia para Mineração de dados, mineração na web e descoberta de conhecimento

http://www.mitgmbh.de/mit/sp/index.htm – site comercial, ferramentas para mineração de dados

http://www.mlnet.org/ - MLnet Online Information Service; publicações, dados e software relacionado a MLT (machine learning toolbox).

http://www.recursive-partitioning.com - Diversos links de árvores de decisão, aprendizado por máquina, descoberta de conhecimento

http://www.rulequest.com – site comercial de Ross Quinlan

http://www.sgi.com/Technology/mlc - Biblioteca de códigos fonte C de algoritmos de aprendizado por máquina.

7.3 GLOSSÁRIO DE TERMOS

Algoritmo – um conjunto de declarações organizadas para resolver um problema num número finito de passos.

Base de dados – uma coleção de dados interrelacionados armazenados de acordo com um esquema.

Data Warehouse – uma coleção de bases de dados integradas, orientadas por assunto, projetada para apoiar sistemas de apoio a decisão

Data Mart – um subconjunto de dados altamente sumarizados a partir do Data Warehouse projetado para apoiar as necessidades específicas de uma organização.

Desnormalização – colocar dados normalizados em local duplicado, em geral objetivando melhorar a performance do sistema.

Drill-down – adicionar ou substituir um cabeçalho de linha em um relatório para aumentar o nível de detalhe das linhas do conjunto resposta.

Granularidade – o nível de detalhe contido numa unidade de dado. Quanto mais detalhe há, mais baixo o nível de granularidade.

Heurística – o modo de análise no qual o próximo passo é determinado pelos resultados do passo atual de análise.

Inteligência Artificial – é uma tentativa de reproduzir raciocínio inteligente em computadores. É inspirado pelo desejo de conseguir que computadores façam coisas que tipicamente os humanos fazem melhor.

Machine Learning – a habilidade de uma máquina melhorar sua performance automaticamente, baseada na sua performance passada.

Meta Dados – Dados sobre dados. Meta dado técnico reflete a descrição da estrutura, conteúdo, chaves e índices de dados na sua fonte de origem.

Privacidade – a prevenção de acesso não autorizado e manipulação de dado.

Sistema de Gerenciamento de Base de Dados (SGBD) – um sistema baseado em computador para armazenar e administrar dados.

Sistemas Especialistas – programas de computador usando um enfoque baseado em regras de captura de especialidade numa área bem específica. Muito similar à lógica tipo IF-THEN-ELSE. Pelo uso de regras para combinar conhecimento com informação obtida de um especialista, apresentam conclusões, provêm recomendações, e ajudam a escolher entre alternativas.

Terceira forma normal (3FN) – uma tabela encontra-se na 3FN, quando além de estar na 2FN, não contém dependências transitivas.

Dependência transitiva – uma dependência funcional transitiva ocorre quando uma coluna, além de depender da chave primária da tabela, depende de outra coluna ou conjunto de colunas da tabela.

Segunda forma normal (2FN) – uma tabela encontra-se na 2FN, quando, além de estar na 1FN, não contém dependências parciais.

Dependência parcial – uma dependência (funcional) parcial ocorre quando uma coluna depende apenas de parte de uma chave primária composta.

Chave primária – é uma coluna ou combinação de colunas cujos valores distinguem uma linha das demais dentro de uma tabela.

Primeira forma normal (1FN) – uma tabela está na 1FN, quando ela não contém tabelas aninhadas.

Sistemas legados -tipicamente, chamamos de “legacy” os sistemas em mainframe desenvolvidos em COBOL. Acontece que, quando um novo sistema termina de ser desenvolvido e entra em produção, ele imediatamente se transforma em “legacy”.

7.4 TABELA DE TIPOS DE SERVIÇOS DE SAÚDE

TIPO SERVIÇO	DESCRICAO
00	PERICIA MEDICA
01	PERICIA MEDICA
15	NEFROLOGIA
16	ANESTESIOLOGIA
17	NUTRICAO PARENTERAL E ENTERAL
19	ALERGOLOGIA
20	CARDIOLOGIA
21	ANATOMIA PATOLOGICA E CITOPATOLOGIA
22	ELECTRENCEFALOGRAFIA E NEUROFISIOLOGIA C
23	ENDOSCOPIA DIGESTIVA
24	ENDOSCOPIA PERORAL
25	MEDICINA FISICA E REABILITACAO
26	GENETICA
27	HEMOTERAPIA
28	PATOLOGICA CLINICA
29	TISIOPNEUMOLOGIA
30	QUIMIOTERAPIA DO CANCER

31	MEDICINA NUCLEAR
32	RADIODIAGNOSTICO
33	ULTRA-SONOGRAFIA
34	TOMOGRAFIA COMPUTADORIZADA
35	RADIOTERAPIA
36	RESSONANCIA MAGNETICA
39	ANGIOLOGIA – CIRURGIA VASCULAR E LINFATICA
40	CIRURGIA CARDIACA-HEMODINAMICA
41	CIRURGIA DE CABECA E PESCOCO
42	DERMATOLOGIA CLINICO-CIRURGICA
43	CIRURGIA APARELHO DIGESTIVO
44	CIRURGIA ENDOCRINOLOGICA
45	GINECOLOGIA E OBSTETRICIA
46	MICROCIRURGIA RECONSTRUTIVA
47	CIRURGIA DA MAMA
48	CIRURGIA DA MAO
49	NEUROCIRURGIA
50	OFTALMOLOGIA
51	OTORRINOLARINGOLOGIA
52	ORTOPEDIA E TRAUMATOLOGIA
53	CIRURGIA PEDIATRICA
54	CIRURGIA PLASTICA
55	CIRURGIA TORACICA
56	UROLOGIA
65	TIPO FR SERVICO DO PLANO F
70	PSICOLOGIA
71	PSIQUIATRIA
72	FONOAUDIOLOGIA
73	SESSAO SERVIÇO SOCIAL

74	SERVICOS ESPECIAIS
75	SERVICOS ESPECIAIS
76	SERVICOS ESPECIAIS
80	HOSPITAIS
81	SERVIÇOS HOSPITALARES
85	TAXAS DE EQUIPAMENTOS
87	SERVICO ODONTOLOGICO (TAB2 CRED)
88	ODONTOLOGIA
89	ODONTOLOGIA
90	FARMACIA