Mineração de Exceções Negativas e Positivas

Eduardo Corrêa Gonçalves

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre. Área de concentração: Otimização Combinatória e Inteligência Artificial.

Orientador: Alexandre Plastino

Niterói, Dezembro de 2004

Mineração de Exceções Negativas e Positivas

Eduardo Corrêa Gonçalves

Dissertação de Mestrado submetida ao Programa de Pós-Graduação em Computação da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Mestre. Área de concentração: Otimização Combinatória e Inteligência Artificial.

Aprovada por:

Prof. Alexandre Plastino / IC-UFF (Presidente)

Prof. Nelson Francisco Favilla Ebecken / UFRJ

Profa. Ana Cristina Bicharra Garcia / IC-UFF

Niterói, Dezembro de 2004.

À minha família.

Agradecimentos

Inicialmente, agradeço a todos os professores das diversas matérias que cursei na Universidade Federal Fluminense, tanto na graduação como na pós-graduação. Um abraço amigo para os professores Dante, Teresa, Carlos Ribeiro, Luciana, Anna Dolejsi , Cristina Boeres, Martinhon, Simone, Satoru, Regina Toledo e Lúcia Drummond. Para o professor Alexandre Plastino, é claro, um agradecimento muito especial, pois foi com este que eu mais aprendi. Agradeço pela sorte que tive em ter trabalhado com ele nesta dissertação.

Na Fundação Getulio Vargas gostaria de prestar meus sinceros agradecimentos aos senhores Mário Rocha e Miguel Lacerda que me permitiram alterar o horário de trabalho para que eu pudesse cursar, sem problemas, as disciplinas de mestrado. Agradeço também aos colegas Claudenize, André Brazil, Valéria e aos demais funcionários da DITI/FGV pelos diversos favores prestados.

Felizmente tenho convivido com uma série de amigos e amigas muito gentis e inteligentes na UFF: Luis Valente, Adriana, Ilza, Marleusa, Marcos, Haroldo, Bruno, Viviane, Carlos Rodrigo, Renata, Stênio, Luciana, Sanderson, Daniela, Áthila, Cristiane, Sandoval, Geiza, Ivairton, Jacques, Keity e tantos outros. Vocês são ótimos, obrigado!

Agradeço aos meus pais, Ana e Joaquim, pelo apoio, suporte e carinho de sempre. E obrigado Amanda e Inês, por serem sempre extremamente amáveis comigo.

Acima de tudo, agradeço a Deus por estar sempre ao meu lado, me proporcionando saúde e felicidade.

Resumo da dissertação apresentada ao Programa de Pós-Graduação em Computação do Instituto de Computação da UFF como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação (M.Sc.).

Mineração de Exceções Negativas e Positivas

Eduardo Corrêa Gonçalves

Dezembro/2004

Orientador: Alexandre Plastino
Programa de Pós-Graduação em Computação - IC/UFF

As regras de associação multidimensionais representam um tipo importante de conhecimento que pode ser minerado a partir de bancos de dados relacionais ou data warehouses. Estas regras descrevem combinações de valores de atributos que freqüentemente ocorrem juntos na base de dados, podendo revelar padrões escondidos e úteis.

A contribuição principal desta dissertação é a proposta de um método para a mineração de exceções negativas e positivas em bases de dados multidimensionais. O objetivo deste método é encontrar associações que tornam-se mais fracas (exceções negativas) ou mais fortes (exceções positivas) em subconjuntos da base de dados que satisfazem condições específicas sobre atributos selecionados. As exceções candidatas são geradas através da combinação de regras de associação previamente descobertas com um conjunto de atributos especificados pelo usuário. Uma exceção é, de fato, minerada apenas quando o valor de suporte real de uma exceção candidata é muito diferente do seu valor esperado de suporte. Um método para estimar esta expectativa é proposto, assim como medidas de interesse para avaliar as exceções. Um algoritmo para minerar estas exceções e resultados experimentais também são apresentados.

Abstract of Thesis presented to UFF as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.).

Mining Negative and Positive Exceptions

Eduardo Corrêa Gonçalves

December/2004

Advisor: Alexandre Plastino
Department: Computer Science

Multidimensional association rules represent an important type of knowledge that can be mined from large relational databases or data warehouses. These rules describe combinations of attribute values that often occur together in a database and can reveal hidden and useful patterns.

The main contribution of this work is to propose an approach to mine negative and positive exceptions from multidimensional databases. The goal of this approach is to find association rules that become weaker (negative exceptions) or stronger (positive exceptions) in some subsets of the database, which satisfy specific conditions over selected attributes. The candidates for exceptions are generated combining previously discovered multidimensional association rules with a set of significant attributes specified by the user. An exception is mined only when the actual support value of the candidate is much different from its expected support value. A method to estimate this expectation is proposed, as well as interest measures to evaluate the exceptions. An algorithm to mine such exceptions and experimental results are also presented.

Palavras-chave

1. Mineração de Exceções
2. Exceções Positivas e Negativas
3. Regras de Associação Negativas
4. Regras de Associação
5. Mineração de Dados

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Nas últimas décadas, os sistemas para gerenciamento de bancos de dados (SGBD's) tiveram a sua utilização bastante difundida entre as organizações. Atualmente a maioria das operações efetuadas por uma empresa costuma produzir registros em bancos de dados. Como conseqüência, estas empresas vêm coletando e armazenando de forma contínua uma enorme quantidade de dados a respeito de seus clientes, fornecedores, produtos e serviços. Esta grande massa dados pode ser examinada por especialistas, para que novas informações sejam descobertas e utilizadas em benefício da organização. Não se trata de uma tarefa trivial, já que, em muitos casos, um banco de dados contém milhões de registros e existem centenas de atributos independentes que precisam ser simultaneamente considerados durante a análise. Desta maneira, métodos tradicionais como investigação manual, consultas SQL e planilhas de cálculo tornam-se inviáveis [11].

Para atender a esta necessidade, no início da década de 90, pesquisadores começaram a apresentar as idéias que dariam origem a uma linha de pesquisa que foi denominada **mineração de dados** (*data mining*). A mineração de dados é realizada por meio de estratégias automatizadas para a análise de grandes bases de dados, procurando extrair das mesmas informações que estejam implícitas, que sejam previamente desconhecidas e potencialmente úteis. Em geral, o conhecimento descoberto através de processos de mineração de dados é expresso na forma de **regras** e **padrões**. Dentre os diferentes tipos de informação que podem ser

minerados em bases de dados encontram-se as regras de associação, regras de classificação, *clusters* de dados, padrões seqüenciais e os padrões em séries temporais. Detalhes sobre as técnicas e tarefas de mineração de dados podem ser encontrados em [4, 9, 11, 14, 13, 17].

As regras de associação são o objeto de estudo deste trabalho. Estas regras representam combinações de itens que ocorrem com determinada freqüência em uma base de dados. As regras de associação possuem como aplicação típica a análise de transações de compras (*market basket analysis*). Este processo examina padrões de compras de consumidores para determinar produtos que costumam ser adquiridos em conjunto. A partir de uma base de dados que armazena os produtos comprados por clientes de, por exemplo, um supermercado ou uma loja de departamentos, uma estratégia para a mineração de regras de associação poderia gerar o seguinte padrão: $\{salaminho\} \Rightarrow \{cerveja\}$. Esta regra indica que os clientes que compram $\{salaminho\}$ tendem a também comprar $\{cerveja\}$.

As regras de associação foram introduzidas em [1] da seguinte forma. Sejam $\mathcal{I} = \{i_1, i_2, ...i_m\}$ um conjunto de m itens distintos e $\mathcal{D}$ uma base de dados formada por um conjunto de transações, onde cada transação T é composta por um conjunto de itens, tal que $T \subseteq \mathcal{I}$. Uma regra de associação é uma expressão na forma $A \Rightarrow B$, onde $A \subset \mathcal{I}$, $B \subset \mathcal{I}$, $A \neq \emptyset$, $B \neq \emptyset$ e $A \cap B = \emptyset$. A é denominado antecedente e B denominado conseqüente da regra. O **suporte** de um conjunto de itens Z, $Sup(Z)$, representa a porcentagem de transações da base de dados que contêm os itens de Z. O suporte de uma regra de associação $A \Rightarrow B$, $Sup(A \Rightarrow B)$, é dado por $Sup(A \cup B)$. Já a **confiança** desta regra, $Conf(A \Rightarrow B)$, representa, dentre as transações que contêm A, a porcentagem de transações que também contêm B. Utilizando a notação da Teoria da Probabilidade [19], tem-se que $Sup(A \Rightarrow B) = P(A, B)$ e $Conf(A \Rightarrow B) = P(B|A)$.

O suporte é utilizado para indicar a relevância de uma regra de associação, enquanto a confiança é utilizada para indicar a sua validade. Tipicamente, o problema da mineração de regras de associação consiste em gerar todas as regras que possuam suporte e confiança maiores ou iguais, respectivamente, a um suporte

mínimo ($SupMin$) e uma confiança mínima ($ConfMin$), especificados pelo usuário. Por esta razão é dito que o suporte e a confiança atuam como **medidas de interesse** no processo de mineração de regras de associação.

As regras de associação obtidas a partir de bases de dados de transações são conhecidas na literatura como **regras de associação transacionais**. No entanto, também é possível minerar regras de associação a partir de bases de dados que contêm atributos quantitativos e categóricos, como *data warehouses* e bancos de dados relacionais. Neste caso, as regras de associação extraídas envolvem múltiplos atributos (ou **dimensões** - terminologia empregada na área de *data warehouses*). Este tipo de regra é denominado **regra de associação multidimensional** [17, 18]. Considere uma base de dados de um supermercado que possui, além dos produtos comprados por seus clientes, outros atributos que informam os dados pessoais destes. Um exemplo de regra multidimensional que poderia ser minerada a partir desta base é dado por: *(Sexo = "F")* $\wedge$ *(30* $\leq$ *Idade* $\leq$ *35)* $\Rightarrow$ *(Forma de Pagamento = "cartão de crédito").* Esta regra hipotética indica que clientes do sexo feminino, com idade entre 30 e 35 anos, costumam pagar pelas suas compras utilizando cartão de crédito. Note que esta regra envolve três atributos (dimensões), sendo um deles quantitativo (*Idade*) e dois deles categóricos (*Sexo* e *Forma de Pagamento*).

Usuários também podem estar interessados em minerar **regras de associação negativas** em bases de dados. O problema da mineração deste tipo de regra foi introduzido em [27]. Segundo esta proposta, uma regra de associação negativa é extraída da base de dados quando não alcançar um **suporte esperado**. Este suporte esperado é computado baseado na existência de uma taxonomia que classifica hierarquicamente os itens que pertencem ao domínio de conhecimento da aplicação. É esperado que itens que pertençam a uma mesma classe, como por exemplo {*pepsi*} e {*coca-cola*}, possuam associações similares com outros itens. Um exemplo apresentado em [27] ilustra que a regra de associação negativa {*ruffles*} $\not\Rightarrow$ {*pepsi*} pode ser minerada caso o suporte da regra convencional (positiva) {*ruffles*} $\Rightarrow$ {*pepsi*} seja significativamente inferior ao suporte da regra positiva {*ruffles*} $\Rightarrow$ {*coca-cola*} (levando em consideração a proporção entre os valores de suporte dos produtos {*coca-cola*} e {*pepsi*}).

Em [38] é apresentado um algoritmo que minera, ao mesmo tempo, regras positivas e negativas em uma base de transações sem a necessidade da utilização de taxonomias. A técnica é capaz de minerar três tipos de regras de associação negativas: $A \Rightarrow \neg B$, $\neg A \Rightarrow B$ e $\neg A \Rightarrow \neg B$. Para evitar o problema da geração de regras redundantes e contraditórias e para podar o espaço de busca, as regras de associação negativas são obtidas a partir dos chamados **conjuntos de itens infreqüentes interessantes**. Um conjunto de itens $A \cup B$ é considerado infreqüente interessante se $Sup(A \cup B) < SupMin$, $Sup(A) \geq SupMin$ e $Sup(B) \geq SupMin$.

Em [21] foi introduzido um novo tipo de regra de associação negativa, denominado **regra de associação relacional negativa**. Estas regras podem ser mineradas a partir de bancos de dados relacionais e são extraídas se possuírem o suporte ou confiança significativamente inferior a uma determinada expectativa. Este conceito será ilustrado através de um exemplo. Considere uma base de dados real que armazena informações sobre pessoas portadoras do vírus da AIDS no Brasil, que foram diagnosticadas entre 1980 e 2001 [24]. Esta base contém registros de mais de 170.000 pacientes. Existem atributos que descrevem o grupo de risco, sexo, raça, idade e outras características de cada uma das pessoas. Uma das regras de associação mais fortes nesta base de dados é: *(Transmissão Sexual = "Não")* $\Rightarrow$ *(Drogas = "Sim")*. Esta regra possui suporte de 13,05% e confiança de 88,09%. Ela indica que a maior parte dos pacientes que não contraíram AIDS através de relações sexuais são usuários de drogas injetáveis.

Suponha que um analista esteja interessado em descobrir se esta associação continua relevante e válida para pacientes de qualquer faixa etária. Neste caso, ele poderia utilizar uma estratégia para a mineração de regras de associação relacionais negativas, que descobriria o seguinte padrão: *(Transmissão Sexual = "Não")* $\wedge$ *(Idade = "$\geq$ 50 anos")* $\not\Rightarrow$ *(Drogas = "Sim")*. Esta regra negativa indica que não é verdade que os pacientes que não contraíram AIDS através de relações sexuais e que possuem idade acima de 50 anos, são usuários de drogas. Considerando-se que a condição *(Idade = "$\geq$ 50 anos")* é freqüente na base de dados (possui suporte mínimo) e é sabido que a regra *(Transmissão Sexual = "Não")* $\Rightarrow$ *(Drogas = "Sim")* possui valores altos para suporte e confiança, era de se esperar que a regra *(Trans-*

missão Sexual = "Não") ∧ (Idade = "≥ 50 anos") ⇒ (Drogas = "Sim") possuísse valores proporcionais para estas medidas. No entanto, a mesma apresentou suporte significativamente inferior a expectativa, tornando-se, por isto, uma regra negativa interessante.

Ainda em [21], é apresentada uma técnica para a extração de regras negativas em bases transacionais, sem a necessidade da utilização de taxonomias. De maneira similar às regras de associação relacionais negativas, uma **regra de associação transacional negativa** é minerada em uma base de dados quando possui valor de suporte ou confiança inferior a uma determinada expectativa. Um exemplo deste tipo de padrão negativo é dado por: $\{couve\} \wedge \{lingüiça\} \not\Rightarrow \{brócolis\}$. Esta regra negativa hipotética indica que clientes que compram couve e lingüiça não costumam comprar brócolis. A regra foi inferida a partir de uma avaliação da regra positiva $\{couve\} \Rightarrow \{brócolis\}$ contra o item $\{lingüiça\}$, pelo fato do suporte ou confiança da regra convencional $\{couve\} \wedge \{lingüiça\} \Rightarrow \{brócolis\}$ ser inferior a uma determinada expectativa. Uma regra negativa deste tipo pode ser útil para identificar clientes com diferentes perfis de compra. Um especialista pode concluir, por exemplo, que a regra $\{couve\} \Rightarrow \{brócolis\}$ é válida entre os clientes adeptos de refeições que priorizam o consumo de verduras e legumes, mas que a mesma torna-se inválida entre os clientes que consomem carne de boi ou de porco.

Nesta dissertação, propõe-se uma extensão para o método apresentado em [21]: a mineração de **exceções negativas** e **positivas** em bases de dados. Nesta abordagem, usuários podem explorar um conjunto de regras de associação, descobrindo o quanto a força de cada uma destas regras desvia-se de seus valores médios em diferentes subconjuntos selecionados da base de dados. As exceções ficam caracterizadas quando estes desvios apresentam valores significativos. Desta forma, informações úteis e inesperadas poderão ser evidenciadas.

Uma regra de associação pode tornar-se muito mais fraca ou muito mais forte em um determinado extrato da base de dados. No exemplo da base de dados da AIDS foi possível descobrir uma regra muito forte, que torna-se fraca (negativa) num subconjunto da base de dados. As pessoas que não contraíram AIDS através

de relações sexuais costumam ser usuárias de drogas injetáveis, mas isto não é válido para o grupo de pacientes com idade igual o superior a 50 anos. De acordo com a proposta para mineração de exceções, que será apresentada neste trabalho, esta regra de associação relacional negativa passa a ser caracterizada como uma **exceção negativa**. Esta exceção seria representada por:

$$\textit{(Transmissão Sexual = "Não")} \overset{-s}{\Longrightarrow} \textit{(Drogas = "Sim")}\ [\textit{(Idade = "}\geq \textit{50 anos")}].$$

O símbolo "$\overset{-s}{\Longrightarrow}$" é utilizado para caracterizar uma exceção negativa, indicando que o suporte da regra em questão encontra-se abaixo da expectativa, quando a base de dados é restringida pela condição entre colchetes.

Será agora apresentado um exemplo de regra de associação fraca (com valores baixos para suporte e confiança), que torna-se mais forte num subconjunto da base de dados da AIDS, caracterizando uma **exceção positiva**. A regra *(Sexo = "F")* $\Rightarrow$ *(Transfusão = "Sim")* possui suporte de 0,59% e confiança de 2,27%. Ela indica que apenas uma pequena fração dos pacientes do sexo feminino contraíram AIDS através de transfusão sanguínea.

Suponha que um analista esteja interessado em descobrir se esta associação torna-se forte em alguma das regiões demográficas do Brasil. Uma estratégia para a mineração de exceções positivas descobriria o seguinte padrão:

$$\textit{(Sexo = "F")} \overset{+s}{\Longrightarrow} \textit{(Transfusão = "Sim")}\ [\textit{(Região = "Norte")}]\ .$$

O significado intuitivo desta notação é o de indicar que no subconjunto da base de dados que contém apenas os pacientes que residem na região norte, o suporte da regra de associação *(Sexo = "F")* $\Rightarrow$ *(Transfusão = "Sim")* possui valor superior a uma determinada expectativa. O cálculo deste valor esperado é realizado avaliando-se o suporte da condição *(Região = "Norte")* e o suporte da regra de associação *(Sexo = "F")* $\Rightarrow$ *(Transfusão = "Sim")*. O símbolo "$\overset{+s}{\Longrightarrow}$", apresentado neste exemplo, é

utilizado para caracterizar uma exceção positiva.

A técnica para mineração de exceções proposta nesta dissertação possui o objetivo principal de evidenciar informações inesperadas e interessantes para usuários de uma ferramenta de mineração de dados. Esta técnica representa uma extensão da proposta para mineração de regras relacionais negativas apresentada em [21]. Nesta dissertação, propõe-se um procedimento para extração de exceções negativas e positivas em bases de dados. Com o objetivo de validar este procedimento, serão descritos os resultados obtidos através da implementação e teste do mesmo sobre diferentes bases de dados. Ainda nesta dissertação, são propostas novas medidas de interesse que avaliam não apenas a força das exceções, mas também o grau de inesperabilidade das mesmas.

Uma técnica para a mineração de exceções também é apresentada em [32]. Nesta abordagem, as exceções são mineradas em forma de **triplas** (*rule triplets*). Uma tripla é composta por duas regras de associação positivas e uma negativa. É importante considerar que, nesta proposta, uma regra negativa é aquela que possui suporte mínimo, mas confiança abaixo de um determinado limiar estabelecido pelo usuário. A interpretação das informações fornecidas pelas triplas caracterizam as exceções encontradas na base de dados.

Considere y e z condições definidas sobre dois atributos distintos de uma base de dados relacional. Considere x e x' duas condições distintas definidas sobre um mesmo atributo desta base. Uma tripla possui o seguinte formato: $(y \rightarrow x,\ yz \rightarrow x',\ z \not\rightarrow x')$. O significado intuitivo desta representação é o seguinte: "y implica em x; porém $y \cup z$ implica em x', apesar de z não implicar em x'". A regra $yz \rightarrow x$ é denominada **regra de exceção**, pois contradiz as regras $y \rightarrow x$ e $z \not\rightarrow x'$. Em [32], é proposto um algoritmo para a mineração automática de triplas, onde o usuário fornece como entrada valores mínimos de suporte e confiança.

O método que será apresentado nesta dissertação difere daquele introduzido em [32] em diversos aspectos. O método definido em [32] emprega um algoritmo para extrair automaticamente as triplas a partir das bases de dados, sem que o usuário escolha os atributos e regras que deseja explorar. Por esta razão, este algoritmo é

considerado como **não-cooperativo**. Embora esta característica seja atraente em alguns aspectos, ela possui a desvantagem de apresentar como saída ao usuário um número extremamente volumoso de regras. Esta dissertação apresentará um método **cooperativo** para obtenção de exceções, onde o usuário pode guiar o foco do processo de mineração para um conjunto de atributos especificados. O método definido em [32] se preocupa em caracterizar uma exceção como uma regra que contradiz a combinação de outras duas regras. Nenhum índice é empregado para classificar o grau de interesse das exceções. Já nesta dissertação, uma exceção representa uma regra cuja força desvia-se de seus valores médios em uma determinada fatia da base de dados. Medidas de interesse são utilizadas para avaliar a relevância e o grau de inesperabilidade das exceções. Por fim, o método a ser apresentado nesta dissertação introduz os conceitos de exceção positiva e negativa.

Esta dissertação é composta pelos seguintes capítulos:

- Capítulo 1: Introdução. Neste capítulo, foi apresentado o tema central da dissertação e os principais trabalhos relacionados foram revistos.

- Capítulo 2: Medidas de Interesse. Este capítulo define e discute o papel das medidas de interesse na mineração de regras de associação. Os conceitos de medida de interesse objetiva e subjetiva são revistos. São apresentados exemplos da utilização destas medidas na mineração de regras de associação.

- Capítulo 3: Mineração de Exceções em Bases Multidimensionais. Neste capítulo, é apresentada a contribuição central desta dissertação: a mineração de exceções em bases de dados multidimensionais. Propõem-se medidas de interesse objetivas para avaliar o grau de interesse das exceções, e também apresenta-se um algoritmo para a mineração de exceções em bases de dados multidimensionais.

- Capítulo 4: Resultados Experimentais. Este capítulo apresenta uma série de resultados obtidos com a mineração de exceções em bases de dados reais de diferentes áreas, como Biologia, Medicina e Ciências Sociais.

- Capítulo 5: Conclusões e Trabalhos Futuros. Finalmente, neste capítulo são apresentadas as conclusões e idéias para futuros trabalhos.

Capítulo 2

Medidas de Interesse

Um dos problemas centrais na área de mineração de dados é o desenvolvimento de boas medidas de interesse para determinar as informações que são, de fato, relevantes e úteis, dentre as muitas que podem ser mineradas. Neste capítulo, o papel das medidas de interesse na mineração de regras de associação é discutido. O texto está organizado da seguinte forma. As Seções 2.1 e 2.2 definem as regras de associação transacionais e multidimensionais, respectivamente. A Seção 2.3 é dedicada a uma avaliação prática das medidas de interesse suporte e confiança, que são comumente utilizadas pelos processos de mineração de regras de associação. Nesta seção, também é apresentado o conceito de dependência entre itens de dados. A Seção 2.4 revisa algumas das medidas de interesse objetivas (índices estatísticos) que foram desenvolvidas com o intuito de avaliar a dependência entre itens de dados. São apresentados exemplos da utilização destas medidas na mineração de regras de associação a partir de uma base de dados real. Finalmente, a Seção 2.5 aborda a importância das medidas de interesse subjetivas.

2.1 Regras de Associação Transacionais

As regras de associação transacionais [1] representam combinações de itens que ocorrem com determinada freqüência em uma base de dados. Uma aplicação

típica para estas regras é a análise de transações de compra (*market basket analysis*). A partir de uma base de dados que armazena produtos comprados por clientes, uma estratégia para a mineração de regras de associação poderia gerar o seguinte exemplo: $\{feijão\} \Rightarrow \{arroz\}$. Esta regra indica que parte significativa dos clientes que compram feijão, também compra arroz. O exemplo ilustra umas das características mais atrativas das regras de associação: elas são expressas em uma forma muito fácil de ser compreendida.

A Tabela 2.1 ilustra uma base de dados de transações de compras de um supermercado hipotético. Cada transação é constituída por um identificador único (TID) e pela relação de produtos adquiridos por um cliente. Dependendo da aplicação, uma transação pode representar a relação de páginas visitadas por um usuário de um portal Internet ou as doenças apresentadas por um paciente de um hospital, entre outros exemplos.

TID	**Lista de Itens**
1	arroz, biscoito, chá, feijão
2	arroz, pão, salaminho
3	café, pão
4	chá, pão
5	arroz, café, feijão, pão
6	café, kiwi, pão

Tabela 2.1: Base de dados de transações.

Definição 2.1.1 *(Regra de Associação Transacional). Seja* $\mathcal{I} = \{i_1, i_2, ...i_m\}$ *um conjunto de* m *itens distintos e* $\mathcal{D}$ *um conjunto de transações, onde cada transação* T *é composta por um conjunto de itens, tal que* $T \subseteq \mathcal{I}$*. Associado a cada transação está um código que a identifica unicamente, denominado* TID*. Uma regra de associação é uma expressão na forma* $A \Rightarrow B$*, onde* $A \subset \mathcal{I}$*,* $B \subset \mathcal{I}$*,* $A \neq \emptyset$*,* $B \neq \emptyset$ *e* $A \cap B = \emptyset$*. A é denominado antecedente e B denominado conseqüente da regra.*

Tanto o antecedente, quanto o conseqüente de uma regra de associação po-

dem ser formados por conjuntos contendo um ou mais itens (*itemsets*). A quantidade de itens pertencentes a um conjunto de itens é chamada de **comprimento** do conjunto. Um conjunto de itens de comprimento k costuma ser referenciado como um *k-itemset*. Associado a cada conjunto de itens, está uma medida de significância chamada **suporte** (Sup), definida a seguir.

Definição 2.1.2 *(Suporte de um Conjunto de Itens). Seja T_n o número total de transações de uma base de dados e T_Z o número de transações que contêm o conjunto de itens Z. O suporte de Z é dado por:*

$$Sup(Z) = \frac{T_Z}{T_n} .$$

Na base de dados representada pela Tabela 2.1, o suporte do itemset $\{arroz\}$ é igual a $3 \div 6 = 50\%$. Isto indica que 50% das transações da base de dados incluem o produto arroz.

O suporte de uma regra de associação $A \Rightarrow B$ representa a porcentagem de transações da base de dados que contêm os conjuntos de itens A e B. Representa a fração das transações que satisfazem tanto o antecedente quanto o conseqüente da regra. Esta medida é utilizada para indicar a relevância de uma regra de associação.

Definição 2.1.3 *(Suporte de uma Regra de Associação). Seja T_n o número total de transações de uma base de dados e $T_{A \cup B}$ o número de transações que contêm os itens de $A \cup B$. O suporte da regra $A \Rightarrow B$, é dado por:*

$$Sup(A \Rightarrow B) = \frac{T_{A \cup B}}{T_n} .$$

Na base de dados representada pela Tabela 2.1, o suporte da regra $\{arroz\} \Rightarrow \{feijão\}$ é igual a $2 \div 6 = 33,33\%$. Isto indica que 33,33% das transações da base de dados incluem o conjunto de itens $\{arroz, feijão\}$.

A confiança ($Conf$) de uma regra de associação representa, dentre as transações que contêm o antecedente da regra, a porcentagem de transações que também contêm o conseqüente desta regra. Esta medida é utilizada para indicar a validade da regra.

Definição 2.1.4 *(Confiança de uma Regra de Associação). Seja T_A o número de transações que incluem os itens de um conjunto A e $T_{A \cup B}$ o número de transações que contêm os itens de $A \cup B$. A confiança da regra $A \Rightarrow B$, é dada por:*

$$Conf(A \Rightarrow B) = \frac{T_{A \cup B}}{T_A} .$$

Como exemplo, na base de dados representada pela Tabela 2.1, existem duas transações que contêm os produtos arroz e feijão juntos ($T_{A \cup B}$) e três transações que contêm o produto arroz (T_A). A confiança da regra $\{arroz\} \Rightarrow \{feijão\}$ é igual a $2 \div 3 = 66.67\%$.

2.2 Regras de Associação Multidimensionais

O crescente interesse pela utilização das regras de associação em ferramentas de mineração de dados levou a necessidade de se definir e oferecer aos usuários variações deste tipo de regra. Um exemplo são as regras de associação multidimensionais [17, 18]. Estas regras são mineradas a partir de bases de dados que contêm atributos quantitativos e categóricos, como *data warehouses* e bancos relacionais, e envolvem mais de um atributo (ou **dimensão**, seguindo a terminologia empregada pela área de *data warehouses*).

Definição 2.2.1 *(Regra de Associação Multidimensional). Seja $\mathcal{D}$ uma relação pertencente a uma base de dados multidimensional. Uma regra de associação multidimensional obtida a partir de $\mathcal{D}$ é uma expressão que possui a seguinte forma:*

$$A_1 = a_1, ..., A_n = a_n \Rightarrow B_1 = b_1, ..., B_m = b_m,$$

onde A_i $(1 \leq i \leq n)$ e B_j $(1 \leq j \leq m)$ representam atributos distintos de $\mathcal{D}$ e a_i e b_j representam, respectivamente, valores dos domínios de A_i e B_j.

Com o objetivo de simplificar a notação, no restante deste trabalho uma regra de associação multidimensional genérica também será representada por $A \Rightarrow B$, onde A e B são conjuntos de condições sobre diferentes atributos.

Os índices estatísticos suporte e confiança também podem ser utilizados para determinar, respectivamente, a relevância e a validade das regras de associação multidimensionais. O suporte de um conjunto de condições Z, definidas sobre uma relação $\mathcal{D}$, representa a porcentagem de tuplas de $\mathcal{D}$ que satisfazem todas as condições em Z. O suporte de uma regra $A \Rightarrow B$ em um banco de dados relacional corresponde à probabilidade de uma tupla satisfazer todas as condições em $A \cup B$. Já a confiança de $A \Rightarrow B$ corresponde à probabilidade de uma tupla satisfazer B, dado que ela satisfaz A.

As técnicas para mineração de regras de associação multidimensionais diferem de acordo com o tratamento dado aos atributos quantitativos da base de dados [17]. A técnica mais simples consiste em discretizar estes atributos antes do processo de mineração. De maneira oposta, uma abordagem mais complexa, introduzida em [31] e denominada mineração de **regras de associação quantitativas**, consiste em discretizar dinamicamente os atributos quantitativos durante o processo de mineração.

Freqüentemente, o conseqüente de uma regra multidimensional é composto por apenas um atributo. Neste caso, o usuário deseja saber como os outros atributos da base de dados se relacionam com este **atributo alvo**.

2.3 O Modelo Suporte/Confiança

O modelo clássico para mineração de regras de associação em bases de dados, introduzido em [1], consiste em encontrar todas as regras que possuam suporte e confiança maiores ou iguais, respectivamente, a um suporte mínimo ($SupMin$) e uma confiança mínima ($ConfMin$), especificados pelo usuário. Por este motivo, o modelo costuma ser referenciado na literatura como **Modelo Suporte/Confiança**. Nesta abordagem, o processo de mineração é dividido em duas etapas:

1. Determinar todos os conjuntos de itens que possuem suporte maior ou igual a $SupMin$. Estes conjuntos são chamados de **conjuntos freqüentes** (*frequent*

itemsets).

2. Para cada conjunto freqüente encontrado na Etapa 1, gerar as regras de associação que possuem confiança maior ou igual a $ConfMin$.

Considere a base de dados ilustrada na Tabela 2.1. Suponha suporte e confiança mínimos iguais a 30% e 65%, respectivamente. Uma estratégia de mineração de regras de associação baseada no Modelo Suporte/Confiança, geraria como resultado as quatro regras apresentadas na Tabela 2.2. Os valores mostrados na segunda (Sup_A) e terceira (Sup_B) colunas representam, respectivamente, os valores de suporte do antecedente e do conseqüente de cada regra. Os valores apresentados na quarta (Sup) e quinta ($Conf$) colunas representam o suporte e a confiança de cada regra de associação, respectivamente.

Regra de Associação	**Sup$_A$**	**Sup$_B$**	**Sup**	**Conf**
{arroz} ⇒ {feijão}	0,5000	0,3333	0,3333	0,6667
{feijão} ⇒ {arroz}	0,3333	0,5000	0,3333	1,0000
{café} ⇒ {pão}	0,5000	0,8333	0,5000	1,0000
{arroz} ⇒ {pão}	0,5000	0,8333	0,3333	0,6667

Tabela 2.2: Exemplo dos índices suporte e confiança.

Observe que no Modelo Suporte/Confiança, para que uma regra seja considerada forte, contendo informação interessante, é necessário que ela apresente bons valores de suporte (Etapa 1) e confiança (Etapa 2). A decisão sobre quais regras devem ser mantidas e quais deverão ser descartadas durante o processo de mineração é baseada nos valores destes dois índices. Isto significa que o suporte e a confiança atuam como **medidas de interesse** no processo de mineração de regras de associação.

2.3.1 Avaliação do Modelo Suporte/Confiança

O Modelo Suporte/Confiança tem recebido muitas críticas ao longo dos últimos anos. O número de regras geradas pelo modelo geralmente é muito grande, dificultando o processo de análise por parte do usuário. Experimentos apresentados em [39] demonstraram que a mineração de bases de dados reais pode levar à geração de centenas de milhares de regras de associação. Além disso, grande parte destes resultados minerados costuma ser composto por regras óbvias, redundantes ou, até mesmo, contraditórias, conforme argumentado em [7, 22].

Para ilustrar estes problemas, serão apresentados os resultados de uma avaliação prática do Modelo Suporte/Confiança. Esta avaliação foi realizada através da mineração de uma base de dados real: a **Pesquisa Sobre Orçamentos Familiares** (POF) da Fundação Getulio Vargas [12]. A POF é uma pesquisa realizada desde 1947 que tem como objetivo produzir informações sobre consumo, através da identificação dos hábitos de compra de famílias residentes em várias capitais do Brasil. No **Caderno B** da POF existe uma relação contendo diversos gêneros alimentícios e bebidas que podem ser adquiridos em supermercados. Este caderno é distribuído para famílias de várias classes sociais residentes em algumas capitais do Brasil. As famílias são orientadas a marcar os itens que constam no Caderno B e que foram adquiridos nas suas últimas compras mensais.

Nesta avaliação, foi utilizada a base que registra as compras realizadas por famílias residentes na cidade do Rio de Janeiro, em Junho de 1998 (denotada como POF-RJ-06-98). São, ao todo, 422 famílias que adquirem uma média de 55 itens (gêneros alimentícios e bebidas) em suas compras mensais. Um programa para mineração de regras de associação transacionais baseado no Modelo Suporte/Confiança, escrito utilizando a linguagem C++ e o compilador g++, foi utilizado para minerar a base de dados POF-RJ-06-98. Neste programa, a Etapa 1 do processo de mineração de regras de associação (geração dos conjuntos freqüentes) é executada através de uma implementação do clássico algoritmo Apriori [2].

Por questões de simplicidade, foram mineradas apenas as regras de asso-

ciação envolvendo dois itens (ou seja, um item no antecedente e um item no conseqüente). O suporte mínimo foi estabelecido em 3% e a confiança mínima em 60%. Como resultado, foram mineradas 8.469 regras de associação a partir dos conjuntos freqüentes de tamanho dois (*2-itemsets*). A Tabela 2.3 apresenta algumas das regras de associação obtidas, ordenadas de maneira decrescente de acordo com o valor da medida de confiança.

Id	**Regra de Associação**	$\mathbf{Sup_A}$	$\mathbf{Sup_B}$	**Sup**	**Conf**
R_1	{cenoura} ⇒ {batata inglesa}	0,7701	0,8175	0,7038	0,9138
R_2	{polpa de acerola} ⇒ {ovos de galinha}	0,0427	0,8886	0,0379	0,8889
R_3	{filé de viola} ⇒ {açúcar refinado}	0,0877	0,8649	0,0758	0,8649
R_4	{milho verde em conserva} ⇒ {ervilhas em conserva}	0,3294	0,3791	0,2701	0,8201
R_5	{fruta de conde} ⇒ {melancia}	0,0450	0,1422	0,0308	0,6842
R_6	{banana nanica} ⇒ {banana prata}	0,1209	0,7607	0,0735	0,6078

Tabela 2.3: Regras de associação mineradas da base de dados da POF.

A distribuição de freqüências dos itens na POF-RJ-06-98 não é balanceada. Alguns poucos produtos, como ovos de galinha, açúcar refinado, batata inglesa, cenoura e banana prata, são muito populares, possuindo valor de suporte acima de 70%. No entanto, a grande maioria dos itens da base de dados da POF possui suporte baixo, inferior a 10%. Alguns exemplos destes produtos menos populares são ilustrados na Tabela 2.3: filé de viola, fruta de conde, banana nanica e polpa de acerola.

Cerca de 80% das regras de associação mineradas a partir da base da POF, envolvem itens com diferentes níveis de suporte (um item com suporte baixo no antecedente e um item com suporte alto no conseqüente). Os produtos muito populares acabaram por compor o conseqüente da maioria das regras obtidas.

Em grande parte dos casos, as regras mineradas não parecem expressar relacionamentos válidos, mesmo quando os valores da medida de confiança são muito altos. Isto acontece especialmente quando os níveis de suporte do antecedente e

do conseqüente são muito diferentes. Observe, por exemplo, a regra R_3. Seria verdadeiro considerar que a compra de filé de viola levou os consumidores a também comprar açúcar refinado ou, na realidade, esta regra foi gerada apenas pelo fato de que muitos consumidores normalmente já iriam adquirir açúcar refinado em suas compras mensais? O Exemplo 2.3.1 aborda esta questão.

Exemplo 2.3.1 *(Independência entre Itens). Considere a regra R_3 (Tabela 2.3). A confiança desta regra representa a probabilidade do cliente comprar {açúcar refinado} dado que compra {filé de viola}. Esta confiança é igual a 86,49%. Entretanto, observando a coluna Sup_B, é possível notar que a probabilidade de qualquer cliente comprar {açúcar refinado} é igual a 86,49%. Portanto a compra de {filé de viola} não aumenta e nem diminui a probabilidade da compra de {açúcar refinado}. A compra de {açúcar refinado}* ***independe*** *da compra de {filé de viola}.*

O Exemplo 2.3.1 identificou um caso de regra de associação minerada através do Modelo Suporte/Confiança, que apresenta um relacionamento **ilusório** entre itens da base de dados. Os produtos filé de viola e açúcar refinado são independentes. Porém a regra $\{$*filé de viola*$\} \Rightarrow \{$*açúcar refinado*$\}$ foi gerada pelo fato de a fórmula da medida de confiança não levar em consideração o valor isolado de suporte do conseqüente da regra de associação.

Ainda na Tabela 2.3, outro exemplo de regra de associação minerada envolvendo itens independentes é ilustrado pela regra R_2 : $\{$*polpa de acerola*$\} \Rightarrow \{$*ovos de galinha*$\}$. Observando os valores da confiança e do suporte do conseqüente (Sup_B) desta regra, fica evidenciado que a compra de polpa de acerola não influencia a compra de ovos de galinha.

De acordo com a análise feita no Exemplo 2.3.1, dois itens de dados A e B são independentes se $Conf(A \Rightarrow B) = Sup(B)$. Conforme a Definição 2.1.4, $Conf(A \Rightarrow B) = Sup(A \cup B) \div Sup(A)$. Desta forma, tem-se que A e B são independentes se $Sup(A \cup B) = Sup(A) \times Sup(B)$.

Definição 2.3.1 *(Independência entre Itens). Seja $\mathcal{D}$ uma base de dados de transações definida sobre um conjunto de itens $\mathcal{I}$. Sejam $A \subset \mathcal{I}$ e $B \subset \mathcal{I}$ dois conjuntos*

não vazios de itens, onde $A \cap B = \emptyset$*. Os conjuntos de itens* A *e* B *são independentes se:*

$$Sup(A \cup B) = Sup(A) \times Sup(B) \ .$$

Na Definição 2.3.1, $Sup(A \cup B)$ representa o suporte real do conjunto de itens $A \cup B$, enquanto $Sup(A) \times Sup(B)$ é o **suporte esperado** do conjunto $A \cup B$, considerando o suporte de A e o suporte de B.

Definição 2.3.2 *(Suporte Esperado de um Conjunto de Itens). Seja* $\mathcal{D}$ *uma base de dados de transações definida sobre um conjunto de itens* $\mathcal{I}$*. Sejam* $A \subset \mathcal{I}$ *e* $B \subset \mathcal{I}$ *dois conjuntos não vazios de itens, onde* $A \cap B = \emptyset$*. O suporte esperado (SupEsp) do conjunto* $A \cup B$ *é obtido por:*

$$SupEsp(A \cup B) = Sup(A) \times Sup(B) \ .$$

O fato de a fórmula da medida de confiança não levar em consideração o valor isolado do suporte do conseqüente das regras de associação, também pode levar a mineração de associações envolvendo produtos que possuem dependência negativa, conforme ilustra o exemplo a seguir:

Exemplo 2.3.2 *(Dependência Negativa entre Itens). Considere a regra* R_6 *(Tabela 2.3). A confiança desta regra é igual a 60,78%. Entretanto, observando a coluna* Sup_B*, é possível notar que a probabilidade de qualquer cliente comprar {banana prata} é igual a 76,07%. Então, na realidade, a compra de {banana nanica} diminui a chance de um cliente comprar {banana prata}. Neste caso, é dito que os produtos possuem uma* ***dependência negativa****.*

Definição 2.3.3 *(Dependência Negativa entre Itens). Seja* $\mathcal{D}$ *uma base de dados de transações definida sobre um conjunto de itens* $\mathcal{I}$*. Sejam* $A \subset \mathcal{I}$ *e* $B \subset \mathcal{I}$ *dois conjuntos não vazios de itens, onde* $A \cap B = \emptyset$*. Os conjuntos de itens* A *e* B *possuem dependência negativa se:*

$$Sup(A \cup B) < SupEsp(A \cup B) \ .$$

Observe, agora, a regra R_4 (Tabela 2.3). Desta vez seria verdadeiro considerar que a compra de milho verde levou os consumidores a também comprar ervilha? O Exemplo 2.3.3 aborda esta questão.

Exemplo 2.3.3 *(Dependência Positiva entre Itens). Considere a regra R_4 (Tabela 2.3). A probabilidade de qualquer cliente comprar {ervilhas em conserva} é de 37.91%, enquanto a probabilidade de um cliente comprar este produto dado que compra {ervilhas em conserva} sobe para 82.01%. Portanto os clientes que compram {milho verde em conserva} têm maior probabilidade de comprar {ervilhas em conserva}. Estes produtos possuem uma* ***dependência positiva****.*

Ainda na Tabela 2.3, as regras R_1 e R_5 ilustram casos de dependência positiva entre produtos. Na primeira regra, tanto o antecedente quanto o conseqüente são produtos muito populares. Já a segunda regra demonstra a dependência positiva entre produtos menos populares.

Definição 2.3.4 *(Dependência Positiva entre Itens). Seja $\mathcal{D}$ uma base de dados de transações definida sobre um conjunto de itens $\mathcal{I}$. Sejam $A \subset \mathcal{I}$ e $B \subset \mathcal{I}$ dois conjuntos não vazios de itens, onde $A \cap B = \emptyset$. Os conjuntos de itens A e B possuem dependência positiva se:*

$$Sup(A \cup B) > SupEsp(A \cup B) \ .$$

A experiência com a base de dados da POF demonstrou alguns problemas do Modelo Suporte/Confiança. Observe, por exemplo, que na Tabela 2.3, a confiança da regra R_2 é bem maior do que a confiança da regra R_5. No entanto, a regra R_5 é aquela que, de fato, contém informação interessante, pois representa uma forte dependência positiva entre dois produtos. A medida de confiança, por não considerar a dependência entre os itens de dados, pode gerar um número muito grande de regras que apresentam relacionamentos falsos e ilusórios.

2.4 Medidas de Interesse Objetivas

As medidas de interesse objetivas são índices estatísticos utilizados para selecionar regras interessantes dentre as muitas que podem ser descobertas por um algoritmo de mineração de regras de associação. O suporte e a confiança são exemplos de medidas de interesse objetivas. Nesta seção, serão apresentadas outras medidas de interesse objetivas, desenvolvidas com o objetivo de medir a dependência entre itens de dados. Estas medidas consideram que uma regra de associação é interessante apenas quando o valor de seu suporte real é maior do que o valor de seu **suporte esperado**. O suporte esperado é computado baseado no suporte dos itens que compõem a regra.

Definição 2.4.1 *(Suporte Esperado de uma Regra de Associação). Seja $\mathcal{D}$ uma base de dados de transações. Seja $A \Rightarrow B$ uma regra de associação obtida a partir de $\mathcal{D}$. O suporte esperado de $A \Rightarrow B$ é obtido por:*

$$SupEsp(A \Rightarrow B) = SupEsp(A \cup B) \ .$$

2.4.1 Lift

A medida de interesse *lift* [7], também conhecida como *interest*, é uma das mais utilizadas para avaliar dependências. Dada uma regra de associação $A \Rightarrow B$, esta medida indica o quanto mais freqüente torna-se B quando A ocorre.

Definição 2.4.2 *(Lift). Seja $\mathcal{D}$ uma base de dados de transações. Seja $A \Rightarrow B$ uma regra de associação obtida a partir de $\mathcal{D}$. O valor do* lift *para $A \Rightarrow B$ é computado por:*

$$Lift(A \Rightarrow B) = \frac{Conf(A \Rightarrow B)}{Sup(B)} = \frac{Sup(A \cup B)}{Sup(A) \times Sup(B)} = \frac{Sup(A \cup B)}{SupEsp(A \cup B)}$$

Se $Lift(A \Rightarrow B) = 1$, então A e B são independentes. Se $Lift(A \Rightarrow B) > 1$, então A e B são positivamente dependentes. Se $Lift(A \Rightarrow B) < 1$, A e B são negativamente dependentes. Esta medida varia entre 0 e ∞.

Exemplo 2.4.1 *(Lift). O valor do índice lift para a regra R_3 : {açúcar refinado} ⇒ {filé de viola} (Tabela 2.3) é calculado por:* $0,8649 \div 0,8649 = 1$*, indicando que os itens {açúcar refinado} e {filé de viola} são independentes.*

O valor do índice lift para a regra R_6 : {banana nanica} ⇒ {banana prata} (Tabela 2.3) é calculado por: $0,6078 \div 0,7607 = 0.80$*, indicando que os itens {banana nanica} e {banana prata} possuem dependência negativa (o suporte real da regra é 0.80 vezes o valor de seu suporte esperado).*

O valor do índice lift para a regra R_4 : {milho verde em conserva} ⇒ {ervilhas em conserva} (Tabela 2.3) é calculado por: $0,8201 \div 0,3701 = 2.21$*, indicando que os itens {milho verde em conserva} e {ervilhas em conserva} possuem dependência positiva (o suporte real da regra é 2.21 vezes maior do que seu suporte esperado).*

2.4.2 RI

O índice *Rule Interest* (*RI*) [23], também pode ser utilizado para avaliar dependências. Esta medida indica o valor da diferença entre suporte real e o suporte esperado de uma regra de associação.

Definição 2.4.3 *(RI). Seja $\mathcal{D}$ uma base de dados de transações. Seja $A \Rightarrow B$ uma regra de associação obtida a partir de $\mathcal{D}$. O valor do* RI *para $A \Rightarrow B$ é computado por:*

$$RI(A \Rightarrow B) = Sup(A \cup B) - Sup(A) \times Sup(B) = Sup(A \cup B) - SupEsp(A \cup B).$$

Se $RI(A \Rightarrow B) = 0$, então A e B são independentes. Se $RI(A \Rightarrow B) > 0$, então A e B são positivamente dependentes. Se $RI(A \Rightarrow B) < 0$, A e B são negativamente dependentes. Esta medida varia entre -0.25 e 0.25.

Exemplo 2.4.2 *(RI). O valor do índice RI para a regra R_3 : {açúcar refinado} ⇒ {filé de viola} (Tabela 2.3) é calculado por:* $0,0877 - (0,8649 \times 0,0758) = 0$*, indicando que os itens {açúcar refinado} e {filé de viola} são independentes.*

O valor do índice RI para a regra R_6 : {banana nanica} ⇒ {banana prata} (Tabela 2.3) é calculado por: $0,1209 - (0,7607 \times 0,0735) = -0.06$*, indicando que os itens {banana nanica} e {banana prata} possuem dependência negativa (a diferença entre o valor de suporte real e o valor de suporte esperado da regra é de -6%).*

O valor do índice RI para a regra R_4 : {milho verde em conserva} ⇒ {ervilhas em conserva} (Tabela 2.3) é calculado por: $0,3294 - (0,3791 \times 0,2701) = 0.14$*, indicando que os itens {milho verde em conserva} e {ervilhas em conserva} possuem dependência positiva (a diferença entre o valor de suporte real e o valor de suporte esperado da regra é de 14%).*

É importante observar que tanto o *lift* quanto o *RI* são medidas **simétricas** (ou seja, $RI(A \Rightarrow B) = RI(B \Rightarrow A)$). Isto ocorre porque estes índices possuem o objetivo de mensurar dependência entre os itens, ao invés de medir implicação (o sentido da seta "⇒").

Observa-se que o *lift* consegue destacar com maior facilidade a dependência positiva entre conjuntos de itens que possuem suporte baixo. Já a medida *RI* é especialmente útil para destacar a dependência positiva entre conjuntos de itens que possuem suporte médio ou alto. O exemplo a seguir ilustra estas características das duas medidas.

Exemplo 2.4.3 *(RI × Lift). O valor do índice RI para a regra R_1 : {cenoura} ⇒ {batata inglesa} (Tabela 2.3), que apresenta suporte igual a 70,38% e confiança igual a 91,38%, é calculado por:* $0,7038 - (0,7701 \times 0,8175) = 0,0742$*. Já o valor do índice lift para esta regra é obtido por:* $0,9138 \div 0,8175 = 1,118$*.*

O valor do índice RI para a regra R_5 : {fruta de conde} ⇒ {melancia} (Tabela 2.3), que apresenta suporte igual a 4,50% e confiança igual a 68,42%, é calculado por: $0,0308 - (0,0450 \times 0,1422) = 0,0244$*. Já o valor do índice lift para esta regra é obtido por:* $0,6842 \div 0,1422 = 4,811$*.*

As duas regras são interessantes. Note porém que o valor da medida *RI* para a regra R_1 é bem maior do que o valor de *RI* para a regra R_5. E, no entanto, o

lift da primeira regra é bem menor do que o *lift* da segunda regra. Este exemplo e os outros que foram apresentados ao longo deste capítulo, evidenciam que a utilização conjunta das medidas de interesse suporte, confiança, *lift* e RI permite que usuários possam realizar análises alternativas sobre uma mesma regra, enriquecendo o poder de entendimento a respeito das associações.

Além dos índices aqui apresentados, diversos trabalhos encontrados na literatura apresentam outras medidas de interesse objetivas importantes. Uma técnica que utiliza o teste *chi-square* para identificar a dependência entre itens é apresentado em [6]. Em [7], é introduzida a **medida de convicção**, com o objetivo de avaliar a implicação (o sentido da seta "⇒") em uma regra de associação. Por fim, uma abrangente comparação entre diversas medidas de interesse pode ser encontrada em [33].

2.5 Medidas de Interesse Subjetivas

As medidas de interesse objetivas identificam, estatisticamente, a força das regras de associação. No entanto, uma regra pode possuir valores altos para determinadas medidas objetivas e não ser **subjetivamente** interessante para o analista que as examina. Em muitos casos, uma regra de associação minerada é interessante para determinado usuário, mas não para outro.

Em [28], são identificados os dois principais fatores que podem tornar uma regra de associação subjetivamente interessante para o usuário: utilidade e inesperabilidade. A medida de utilidade considera que uma regra é interessante se o usuário pode fazer algo a partir dela, ou seja, pode tirar proveito do padrão minerado. Já a medida de inesperabilidade considera que uma regra tem grande chance de ser interessante quando contradiz as expectativas do usuário, o que depende de suas convicções, ou seja, do que ele imagina que esteja armazenado na base.

Um exemplo conhecido de regra útil e inesperada, que pôde ser descoberta através de técnicas de mineração de dados, é a associação entre as vendas de fraldas e de cerveja em uma grande loja de departamentos, quando os consumidores são

casais jovens e as compras são realizadas nas noites de quinta-feira [4]. A regra é inesperada porque os analistas imaginavam que as vendas de cerveja estivessem associadas apenas a produtos como salgados, carne para churrasco e outras bebidas alcoólicas, mas nunca a produtos de higiene infantil. Ela também é útil porque os gerentes da loja de departamentos puderam tomar ações capazes de aumentar as vendas de cerveja (este produto foi colocado numa prateleira próxima à prateleira das fraldas).

Ainda em [28], argumenta-se que, embora as medidas de utilidade e inesperabilidade sejam independentes, na prática as regras úteis são na maioria das vezes inesperadas e a maioria das regras inesperadas também costuma ser útil. Por esta razão a medida de inesperabilidade torna-se uma boa aproximação para a medida de utilidade. Seguindo este princípio, alguns trabalhos [20, 22, 35] consideram que os modelos para mineração de regras de associação devem encontrar uma forma para representar as expectativas do analista e incorporá-las ao algoritmo de mineração, para que regras inesperadas possam ser mineradas.

Em [20] uma regra é considerada inesperada se difere sintaticamente de uma convicção (*belief*) informada pelo usuário. Nesta abordagem, uma regra é diferente de uma convicção se os antecedentes são similares e os conseqüentes são distintos, ou vice-versa. A diferença e a similaridade entre os antecedentes e os conseqüentes são calculadas através de uma técnica *fuzzy*.

Em [22], uma regra $A \Rightarrow B$ é considerada inesperada em relação a uma convicção informada pelo usuário na forma $X \Rightarrow Y$, se:

1. $B = \neg Y$, e
2. A regra $X \wedge A \Rightarrow B$ possui os valores mínimos de suporte e confiança.

Esta idéia foi aplicada a um experimento que é relatado em [22], para minerar regras de associação inesperadas em relação a expectativas de analistas, em uma base de dados de compras de consumidores. Os analistas elaboraram um conjunto com 15 crenças (*beliefs*). Dentre elas, encontrava-se a crença "pessoas empregadas

costumam fazer compras no fim de semana". Um algoritmo desenvolvido para mineração de regras inesperadas foi aplicado sobre a base e encontrou a seguinte regra inesperada (em relação à crença citada): "No mês de dezembro, pessoas empregadas costumam fazer compras em dias de semana".

Já em [35], apresenta-se um modelo mais sofisticado para a incorporar o conhecimento do usuário. Um exemplo apresentado neste trabalho descreve a mineração de regras inesperadas a partir de uma base de dados real que mantém informação relativa a doações para a campanha eleitoral de um partido político nos Estados Unidos. Considere que os analistas, baseados na sua experiência, possuam as seguintes crenças: "moradores de Beverly Hills ganham bem" e "pessoas que ganham bem costumam doar mais de US$ 200". O algoritmo de mineração de dados passa a incorporar este conhecimento e pode obter a seguinte regra inesperada em relação a estas duas crenças: "moradores de Beverly Hills tendem a doar menos de US$ 50".

Capítulo 3

Mineração de Exceções em Bases Multidimensionais

Neste capítulo, apresenta-se a contribuição central desta dissertação: a mineração de exceções em bases de dados multidimensionais. A Seção 3.1 descreve a motivação para a mineração deste tipo de padrão. A Seção 3.2 aborda a questão da dependência entre conjuntos de condições para que alguns conceitos preliminares sejam definidos. A Seção 3.3 introduz a abordagem proposta para a mineração de exceções negativas e positivas. O problema da mineração deste tipo de padrão é formalizado e são propostas medidas objetivas para avaliar a força e o grau de inesperabilidade destas. Na Seção 3.4, é apresentado um algoritmo que minera, ao mesmo tempo, exceções negativas e positivas em bases de dados multidimensionais. A Seção 3.5 trata de uma avaliação da utilização da medida de interesse confiança na busca por exceções. Por fim, a Seção 3.6 resume os diferentes tipos propostos de exceção que podem ser minerados em bases de dados.

3.1 Motivação

O principal objetivo de uma tarefa de mineração de dados é transformar dados acumulados em conhecimento útil. Este novo conhecimento minerado pode ser

então utilizado para auxiliar o processo de tomada de decisões em uma organização. As informações úteis são, geralmente, aquelas que surpreendem as expectativas do analista, ou seja, que diferem do que este analista imagina estar armazenado na base de dados. Neste capítulo, será apresentado um método cooperativo para a mineração de exceções em bases de dados. Nesta abordagem, usuários podem explorar um conjunto de regras de associação, descobrindo o quanto a força de cada uma destas regras afasta-se de seus valores médios em diferentes subconjuntos selecionados da base de dados. De maneira análoga aos princípios apresentados em [21] e [27], as exceções são mineradas somente quando estes desvios apresentam valores significativos. Desta forma, o método pode ser capaz de evidenciar informações previamente ocultas e inesperadas.

3.2 Dependência entre Conjuntos de Condições

Conforme apresentado na Seção 2.3.1, dois conjuntos de itens pertencentes a uma base de transações podem ser independentes ou dependentes. Neste último caso, os conjuntos de itens podem possuir dependência negativa ou positiva. O Exemplo 2.3.2, obtido a partir da base de dados da POF, demonstrou um caso de dependência negativa entre dois itens. Este exemplo ilustrou que, na POF-RJ-06-98, a probabilidade de um cliente comprar o produto {*banana prata*} é muito menor entre os clientes que compram o produto {*banana nanica*}. Já o exemplo 2.3.3 ilustrou um caso de dependência positiva nesta mesma base: a probabilidade de um cliente comprar {*ervilhas em conserva*} é muito maior entre os clientes que compram {*milho verde em conserva*}.

De maneira análoga, é também possível descobrir se dois conjuntos de condições, definidas sobre atributos (dimensões) de uma relação de uma base de dados multidimensional, possuem dependência negativa ou positiva. Nesta seção, serão apresentadas duas medidas baseadas no índice *lift*, que são utilizadas para medir estas dependências.

Definição 3.2.1 *(Índice de Dependência Negativa). Seja $\mathcal{D}$ uma relação perten-*

cente a uma base de dados multidimensional. Sejam $X = \{X_1 = x_1, ..., X_n = x_n\}$ *e* $Y = \{Y_1 = y_1, ..., Y_m = y_m\}$ *dois conjuntos de condições definidas sobre atributos de* $\mathcal{D}$*, onde* $X \cap Y = \emptyset$*. O* índice de dependência negativa *entre* X *e* Y*, denotado por DN(X,Y), é obtido por:*

$$DN(X,Y) = 1 - \left(\frac{Sup(X \cup Y)}{SupEsp(X \cup Y)} \right) .$$

O resultado do cálculo do índice de dependência negativa de um conjunto de condições $X \cup Y$ indica, em percentual, o quanto o valor do suporte real deste conjunto é inferior ao valor de seu suporte esperado. Quanto mais próximo o valor de $DN(X,Y)$ está de 1 (que é o valor máximo para esta medida), maior a dependência negativa entre X e Y. Se $DN(X,Y) = 0$, então X e Y são independentes; e se $DN(X,Y) < 0$, então na verdade existe dependência positiva entre X e Y, pois o suporte real do conjunto de condições será maior do que seu suporte esperado.

Será agora apresentado um exemplo de dependência negativa, obtido a partir da base de dados da AIDS, apresentada no Capítulo 1. Nesta base, a condição $X = \{$*(Transmissão Sexual = "Sim")*$\}$ possui suporte igual a 84,91%. Este valor representa o percentual de pessoas que contraíram AIDS através de relações sexuais. Já a condição $Y = \{$*(Idade = "0-12 anos")*$\}$, que possui suporte igual a 4,04%, indica o percentual de pacientes com idade inferior a 13 anos.

De acordo com a Definição 2.3.2, o suporte esperado de $X \cup Y$ é obtido por: $84,91\% \times 4,04\% = 3,43\%$. No entanto, o suporte real de $X \cup Y$ na base de dados da AIDS é igual a 0,0039%. Este valor de suporte real é muito inferior ao suporte esperado. Desta forma, o índice de dependência negativa pode ser utilizado para avaliar a forte dependência negativa entre o fato de uma pessoa ter contraído AIDS através de relações sexuais e o de possuir idade inferior a 13 anos. O índice de dependência negativa entre X e Y é computado por: $1 - (0,0039 \div 3,42) = 0,9988$. Este valor indica que $X \cup Y$ possui um valor de suporte 99,88% inferior ao esperado.

Analogamente, a seguir será apresentada uma outra medida de interesse, desta vez capaz de quantificar a dependência positiva entre conjuntos de condições.

Definição 3.2.2 *(Índice de Dependência Positiva). Seja $\mathcal{D}$ uma relação pertencente a uma base de dados multidimensional. Sejam $X = \{X_1 = x_1, ..., X_n = x_n\}$ e $Y = \{Y_1 = y_1, ..., Y_m = y_m\}$ dois conjuntos de condições definidas sobre atributos de $\mathcal{D}$, onde $X \cap Y = \emptyset$. O* índice de dependência positiva *entre X e Y, denotado por DP(X,Y), é obtido por:*

$$DP(X,Y) = 1 - \left(\frac{SupEsp(X \cup Y)}{Sup(X \cup Y)} \right) .$$

O resultado do cálculo do índice de dependência positiva de um conjunto de condições $X \cup Y$ indica, em percentual, o quanto o valor do suporte real deste conjunto é superior ao valor de seu suporte esperado. Desta vez, o suporte esperado está no numerador da fração e o suporte real está no denominador. Quanto mais próximo o valor de $DP(X,Y)$ está de 1 (que é o valor máximo para esta medida), maior a dependência positiva entre X e Y. Se $DP(X,Y) = 0$, então X e Y são independentes; e se $DP(X,Y) < 0$, então na verdade existe dependência negativa entre X e Y.

Na base de dados da AIDS, o suporte de $X = \{$*(Transfusão = "Sim")*$\}$ é igual a 1,97%, representando o percentual de pacientes que contraíram AIDS através de transfusão sanguínea. O suporte de $Y = \{$*(Idade = "≥50 anos")*$\}$ é igual a 6,65%, que representa o percentual de pacientes com idade igual ou superior a 50 anos.

O suporte esperado de $X \cup Y$ é obtido por: $1,97\% \times 6,65\% = 0,13\%$. No entanto, o suporte real de $X \cup Y$ na base de dados da AIDS é igual a 0,29%. Este valor de suporte real é consideravelmente superior ao suporte esperado, evidenciando uma dependência positiva entre uma pessoa ter contraído AIDS através de transfusão sanguínea e possuir idade igual ou superior a 50 anos. Segundo a Definição 3.2.2, o índice de dependência positiva entre X e Y é obtido por: $1 - (0,13 \div 0,29) = 0,5482$. Este resultado indica que o suporte de $X \cup Y$ possui um valor 54,82% superior ao esperado.

3.3 Exceções

Nesta seção, um método cooperativo para a mineração de exceções negativas e positivas em bases de dados multidimensionais será proposto. As exceções representam regras de associação cuja força afasta-se de seus valores médios em determinados subconjuntos da base de dados, caracterizados por condições específicas. Caso estes desvios sejam significativos, ficarão caracterizados como exceções a regras de associação relevantes.

3.3.1 Exceções Negativas

As exceções negativas representam regras de associação que tornam-se fracas em determinados extratos da base de dados. Estas regras são mineradas se possuírem o suporte significativamente inferior a uma determinada expectativa. Esta expectativa é calculada de acordo com o conceito de dependência negativa entre conjuntos de condições.

Com o objetivo de ilustrar esta proposta, considere a **Base de dados de vendas** apresentada na Tabela 3.1. Esta base contém dados das compras efetuadas por clientes de uma loja de departamentos hipotética, em uma determinada semana. Cada instância da base de dados representa um cliente diferente, identificado pelo atributo **Id**. Os atributos **Sexo**, **Idade** e **Carro** descrevem propriedades dos clientes, enquanto os atributos **Cartão de Crédito** e **Valor da Compra (R$)** descrevem propriedades das compras realizadas por estes.

Um algoritmo para a mineração de regras de associação convencionais a partir da base de dados apresentada na Tabela 3.1 seria capaz de obter o seguinte padrão: "clientes do sexo feminino costumam gastar mais de R$ 100,00 em suas compras". Esta regra possui suporte, confiança e *lift* iguais a, respectivamente, 40%, 66,67% e 1,11.

Entretanto, observando a base de dados, é possível notar que nenhuma das mulheres que efetuaram as suas compras utilizando o cartão de crédito da marca

Id	**Sexo**	**Idade**	**Carro**	**Cartão de Crédito**	**Valor da Compra (R$)**
1	F	<18	Não	Credicard	≥100,00
2	M	26-30	Sim	Visa	≥100,00
3	F	26-30	Sim	Visa	25,00-49,00
4	F	18-25	Sim	Amex	≥100,00
5	F	31-40	Sim	Credicard	≥100,00
6	M	<18	Não	Visa	≥100,00
7	F	26-30	Sim	Amex	≥100,00
8	F	<18	Não	Visa	<25,00
9	M	18-25	Sim	Visa	50,00-99,00
10	M	<18	Não	Credicard	<25,00

Tabela 3.1: Base de dados de vendas.

Visa, gastou mais de R$ 100,00 em suas compras. Então, seria interessante inferir o seguinte padrão negativo: "clientes do sexo feminino, que compram utilizando o cartão da marca VISA, **não** costumam gastar mais de R$ 100,00 em suas compras". Este padrão negativo foi obtido a partir da regra de associação convencional "clientes do sexo feminino costumam gastar mais de R$ 100,00 em suas compras" e foi gerado porque o valor de suporte da regra "clientes do sexo feminino, que compram utilizando o cartão da marca VISA, costumam gastar mais de R$ 100,00 em suas compras" foi significativamente inferior ao esperado. Este exemplo ilustra uma **exceção negativa** associada a uma regra de associação forte. Esta exceção negativa pode ser representada da seguinte maneira:

$$\textit{(Sexo = "F")} \overset{-s}{\Longrightarrow} \textit{(Valor = "≥100,00")}\ [\textit{(Cartão = "Visa")}].$$

O símbolo "$\overset{-s}{\Longrightarrow}$" é utilizado para indicar que o suporte da regra de associação convencional *(Sexo = "F")* ⇒ *(Valor = "≥100,00")* é significativamente inferior ao esperado no subconjunto da base de dados definido pelos clientes que compram utilizando o cartão da marca Visa. A seguir uma definição formal para exceção negativa é apresentada.

Definição 3.3.1 *(Exceção Negativa). Seja $\mathcal{D}$ uma relação pertencente a uma base de dados multidimensional. Seja $R : A \Rightarrow B$ uma regra de associação obtida a partir de $\mathcal{D}$. Seja $Z = \{Z_1 = z_1, ..., Z_k = z_k\}$ um conjunto de condições definidas sobre k atributos distintos de $\mathcal{D}$, onde $\{Z_1 = z_1, ..., Z_k = z_k\} \cap \{A_1 = a_1, ..., A_n = a_n\} \cap \{B_1 = b_1, ..., B_m = b_m\} = \emptyset$. Z é chamado de **conjunto de prova**. Uma exceção negativa relacionada à regra R é uma expressão com a seguinte forma: $A \overset{-s}{\Longrightarrow} B\ [Z]$.*

O objetivo da exceção negativa é representar o quanto a presença de um conjunto de prova pode enfraquecer uma regra de associação que é, originalmente, forte. Estas exceções negativas são obtidas a partir de **exceções candidatas**. Uma exceção candidata é formada através da combinação de uma regra de associação com um conjunto de prova.

Definição 3.3.2 *(Exceção Candidata). Seja $\mathcal{D}$ uma relação pertencente a uma base de dados multidimensional. Seja $R : A \Rightarrow B$ uma regra de associação obtida a partir de $\mathcal{D}$. Seja Z um conjunto de prova, onde $Z \cap A \cap B = \emptyset$. Uma exceção candidata é uma expressão com a seguinte forma: $A \Rightarrow B\ [Z]$.*

Definição 3.3.3 *(Suporte de uma Exceção Candidata). Seja $C : A \Rightarrow B\ [Z]$ uma exceção candidata. O suporte de C, denotado por $Sup(C)$, é definido por:*

$$Sup(C) = Sup(A \cup B \cup Z)\ .$$

Como exemplo, considere a seguinte exceção candidata C_1 : *(Sexo = "F")* $\Rightarrow$ *(Valor = "≥100,00")* [*(Cartão = "Visa")*]. Esta exceção candidata foi gerada através da combinação da regra R_1 : *(Sexo = "F")* $\Rightarrow$ *(Valor = "≥100,00")* com o conjunto de prova $Z_1 =$ {*(Cartão = "Visa")*}. O suporte de C_1 é obtido por $Sup(C_1) = Sup(\{$*(Sexo = "F"),(Valor = "≥100,00"),(Cartão = "Visa")*$\}) = 0\%$.

Uma exceção negativa deve ser minerada a partir de uma base de dados multidimensional somente quando o suporte real de uma exceção candidata não alcança uma determinada expectativa. Esta expectativa é calculada em função do

suporte da regra original $A \Rightarrow B$ e do suporte das condições que compõem o conjunto de prova Z.

Definição 3.3.4 *(Suporte Esperado de uma Exceção Candidata). Seja* $C : A \Rightarrow B\ [Z]$ *uma exceção candidata. O suporte esperado de* C*, denotado por* $SupEsp(C)$*, é computado por:*

$$SupEsp(C) = Sup(A \cup B) \times Sup(Z)\ .$$

Considere, mais uma vez, a exceção candidata C_1 : *(Sexo = "F")* $\Rightarrow$ *(Valor = "≥100,00")* [*(Cartão = "Visa")*]. Levando-se em conta que o suporte da condição (Cartão = "Visa") é igual a 50% e o suporte da regra de associação (Sexo = "F") $\Rightarrow$ (Valor da Compra = "≥100,00") é igual a 40%, era esperado que C_1 possuísse valor de suporte igual a $50\% \times 40\% = 20\%$. No entanto o valor de suporte desta exceção candidata é igual a 0%.

Para avaliar o grau de interesse das exceções negativas, serão utilizadas duas medidas de interesse objetivas. Estas duas medidas são utilizadas para medir, respectivamente, a força (relevância) e o grau de inesperabilidade (validade) das exceções.

Uma exceção negativa $E : A \overset{-s}{\Longrightarrow} B\ [Z]$ pode ser considerada como **potencialmente interessante** se o valor do suporte real da exceção candidata $A \Rightarrow B\ [Z]$ é significativamente inferior ao seu valor esperado de suporte. A **Medida da Força para Exceções Negativas** (MF^-) é utilizada para medir este desvio do valor de suporte em relação ao esperado.

Definição 3.3.5 *(Medida da Força para Exceções Negativas). Seja* $E : A \overset{-s}{\Longrightarrow} B\ [Z]$ *uma exceção negativa. A medida da força para* E*, representada por* $MF^-(E)$*, é computada por:*

$$MF^-(E) = 1 - \left(\frac{Sup(A \Rightarrow B\ [Z])}{SupEsp(A \Rightarrow B\ [Z])} \right)\ .$$

O valor apresentado pela medida MF^- para uma exceção negativa $E : A \overset{-s}{\Longrightarrow} B\ [Z]$, corresponde ao índice de dependência negativa entre um conjunto

de prova Z e uma regra de associação $A \Rightarrow B$. Quanto mais próximo o valor está de 1, mais interessante a exceção negativa. Se $MF^{-}(E) = 0$, então Z e $A \Rightarrow B$ são independentes. Este índice é utilizado para avaliar a relevância de uma exceção negativa.

O suporte real da exceção candidata C_1 : *(Sexo = "F")* $\Rightarrow$ *(Valor = "≥100,00")* [*(Cartão = "Visa")*] é igual a 0%. O suporte do conjunto de prova $Z_1 =$ {*(Cartão = "Visa")*} é igual a 50%. O suporte esperado para C_1 é obtido por $Sup(R_1) \times Sup(Z_1) = 40\% \times 50\% = 20\%$. A exceção E_1 : *(Sexo = "F")* $\overset{-s}{\Longrightarrow}$ *(Valor = "≥100,00")* [*(Cartão = "Visa")*] é potencialmente interessante, pois $MF^{-}(E_1) = 1 - (0 \div 0,20) = 1,00$.

Embora o índice MF^{-} represente a força de uma exceção negativa, um valor alto desta medida não é garantia de que foi descoberta uma informação interessante. O exemplo a seguir ilustrará esta afirmação. Considere a regra de associação "clientes do sexo feminino possuem carro", também extraída da base de dados representada pela Tabela 3.1. Esta regra possui suporte, confiança e *lift* iguais a, respectivamente, 40%, 66,67% e 1,11.

Observando a Tabela 3.1, pode também ser notado que nenhuma das mulheres com idade inferior a 18 anos possui carro. Esta informação poderia levar a conclusão de que a exceção negativa E_2 : *(Sexo = "F")* $\overset{-s}{\Longrightarrow}$ *(Carro = "Sim")* [*(Idade = "<18")*] é interessante, visto que o valor do índice MF^{-} para esta exceção é igual a 1. Entretanto, na realidade, nenhum dos clientes com idade inferior a 18 anos possui carro, independente do fato deles serem homens ou mulheres. Levando-se em consideração que apenas pessoas maiores de 18 anos podem tirar carteira de motorista, a exceção E_2 representa uma informação que é completamente óbvia e sem nenhuma utilidade. Este tipo de informação certamente não deve ser minerada. A medida MF^{-} não foi capaz de detectar a forte dependência negativa entre o fato de um cliente possuir idade inferior a 18 anos e o de ser dono de um automóvel.

Uma exceção negativa E : $A \overset{-s}{\Longrightarrow} B$ $[Z]$ não é interessante se a dependência negativa entre $A \Rightarrow B$ e o conjunto de prova Z é causada pela dependência negativa entre Z e A ou entre Z e B. Portanto, será considerado que uma exceção negativa

é, de fato, interessante quando o valor do índice de dependência negativa entre Z e $A \Rightarrow B$ é consideravelmente maior do que os valores do índice de dependência negativa entre Z e A e entre Z e B. A **Medida do Grau de Inesperabilidade para Exceções Negativas** (GI^-) é utilizada para avaliar esta questão.

Definição 3.3.6 *(Grau de Inesperabilidade de uma Exceção Negativa). Seja* $E : A \overset{-s}{\Longrightarrow} B\ [Z]$ *uma exceção negativa. Sejam DN(A,Z) e DN(B,Z) os valores para o índice de dependência negativa entre* A *e* Z *e entre* B *e* Z*, respectivamente. O grau de inesperabilidade de* E*, representado por* $GI^-(E)$*, é computado por:*

$$GI^-(E) = MF^-(E) - \max(DN(A,Z), DN(B,Z)) \ .$$

Quanto maior e mais distante de 0 é o valor de GI^-, mais interessante é uma exceção negativa. Se $GI^-(E) \leq 0$, então a exceção negativa é desinteressante.

Considere, mais uma vez, a exceção candidata C_1 : *(Sexo = "F")* $\Rightarrow$ *(Valor = "≥100,00")* [*(Cartão = "Visa")*]. O suporte da condição (Cartão = "Visa") é igual a 50%. Para que a medida do grau de inesperabilidade seja obtida, primeiro deve ser calculado $DN(A,Z) = DN(\{$*(Sexo = "F")*$\},\{$*(Cartão = "Visa")*$\}) = 1-(0,20 \div (0,60 \times 0,50)) = 1-(0,20 \div 0,30) = 0,33$. Em seguida, calcula-se $DN(B,Z) = DN(\{$*(Valor = "≥100,00")*$\},\{$*(Cartão = "Visa")*$\}) = 1-(0,20 \div (0,60 \times 0,50)) = 1-(0,20 \div 0,30) = 0,33$. A exceção E_1 : *(Sexo = "F")* $\overset{-s}{\Longrightarrow}$ *(Valor = "≥100,00")* [*(Cartão = "Visa")*] é, de fato, interessante, pois $GI^-(E_1) = 1,00 - max(0,33\ ,\ 0,33) = 0,67$.

Considere, agora a exceção candidata C_2 : *(Sexo = "F")* $\Rightarrow$ *(Carro = "Sim")* [*(Idade = "<18")*]. O suporte da condição (Idade = "<18") é igual a 40%. Calcula-se $DN(A,Z) = DN(\{$*(Sexo = "F")*$\},\{$*(Idade = "<18)*$\}) = 1-(0,20 \div (0,60 \times 0,40)) = 1-(0,20 \div 0,24) = 0,17$. Em seguida, calcula-se $DN(B,Z) = DN(\{$*(Carro = "Sim")*$\},\{$*(Idade = "<18")*$\}) = 1-(0.00 \div (0.60 \times 0.40)) = 1-(0,00 \div 0,24) = 1,00$. A exceção E_2 : *(Sexo = "F")* $\overset{-s}{\Longrightarrow}$ *(Carro = "Sim")* [*(Idade = "18")*] não é interessante, pois $GI^-(E_2) = 1,00 - max(0,17\ ,\ 1,00) = 0,00$.

3.3.2 Exceções Positivas

As exceções positivas representam regras de associação que se tornam mais fortes em determinados extratos da base de dados. O método de mineração de exceções positivas é similar ao empregado para a mineração de exceções negativas.

O conceito de exceção positiva também será ilustrado através de um exemplo. A base de dados hipotética apresentada na Tabela 3.2 contém dados de uma pesquisa sobre tabagismo. Uma informação que pode ser extraída é a seguinte "50% dos homens entrevistados na pesquisa são fumantes". No entanto, observe que todos os homens que residem na região Sudeste fumam. Este exemplo ilustra uma **exceção positiva** associada a uma regra de associação, que pode ser representada da seguinte maneira:

$$\textit{(Sexo = "M")} \overset{+s}{\Longrightarrow} \textit{(Fumante = "Sim")}\ [\textit{(Região = "Sudeste")}].$$

O símbolo "$\overset{+s}{\Longrightarrow}$" é utilizado para indicar que o suporte da regra de associação convencional *(Sexo = "M")* $\Rightarrow$ *(Fumante = "Sim")* é consideravelmente superior ao esperado no subconjunto da base de dados definido pelas pessoas que residem na região Sudeste. A seguir uma definição formal para exceção positiva é apresentada.

Id	Sexo	Idade	Região	Fumante
1	F	20-29	Nordeste	Sim
2	M	30-39	Sudeste	Sim
3	M	30-39	Nordeste	Não
4	F	30-39	Nordeste	Não
5	F	20-29	Sudeste	Não
6	M	40-49	Nordeste	Não
7	M	20-29	Sul	Não
8	M	30-39	Sul	Sim
9	M	40-49	Sudeste	Sim
10	F	20-29	Nordeste	Não

Tabela 3.2: Base de dados de informações sobre tabagismo.

Definição 3.3.7 *(Exceção Positiva). Seja $\mathcal{D}$ uma relação pertencente a uma base de dados multidimensional. Seja $R : A \Rightarrow B$ uma regra de associação obtida a partir de $\mathcal{D}$. Seja $Z = \{Z_1 = z_1, ..., Z_k = z_k\}$ um conjunto de prova definido sobre atributos de $\mathcal{D}$, onde $\{Z_1 = z_1, ..., Z_k = z_k\} \cap \{A_1 = a_1, ..., A_n = a_n\} \cap \{B_1 = b_1, ..., B_m = b_m\} = \emptyset$. Uma exceção positiva relacionada a regra R é uma expressão com a seguinte forma: $A \overset{+s}{\Longrightarrow} B\ [Z]$.*

A idéia da exceção positiva é a de avaliar o quanto a presença de um conjunto de prova é capaz de fortalecer uma regra de associação que pode, originalmente, ser fraca. Assim como ocorre com as exceções negativas, as exceções positivas também são obtidas a partir de exceções candidatas.

Uma exceção positiva é interessante quando seu suporte real possui um valor muito superior ao esperado. Para efetuar este cálculo, é utilizada a **Medida da Força para Exceções Positivas**, MF^+, definida a seguir.

Definição 3.3.8 *(Medida da Força para Exceções Positivas). Seja $E : A \overset{+s}{\Longrightarrow} B\ [Z]$ uma exceção positiva. A medida da força para E, representada por $MF^+(E)$, é computada por:*

$$MF^+(E) = 1 - \left(\frac{SupEsp(A \Rightarrow B\ [Z])}{Sup(A \Rightarrow B\ [Z])} \right) .$$

Desta vez o suporte esperado da exceção candidata encontra-se no numerador da fórmula. Quanto mais alto o valor desta medida, maior o índice de dependência positiva entre Z e $A \Rightarrow B$. Para que exceções positivas interessantes possam ser geradas, também é necessário computar o grau de inesperabilidade.

Definição 3.3.9 *(Grau de Inesperabilidade de uma Exceção Positiva). Seja $E : A \overset{+s}{\Longrightarrow} B\ [Z]$ uma exceção negativa. Sejam DP(A,Z) e DP(B,Z) os valores para o índice de dependência positiva entre A e Z e entre B e Z, respectivamente. O grau de inesperabilidade de E, representado por $GI^+(E)$, é computado por:*

$$GI^+(E) = MF^+(E) - \max(DP(A, Z), DP(B, Z)) .$$

Quanto maior e mais distante de 0 é o valor de GI^+, mais interessante é uma exceção positiva. Se $GI^+(E) \leq 0$, então a exceção positiva é desinteressante.

Considere a exceção candidata C_1 : *(Sexo = "M")* $\Rightarrow$ *(Fumante = "Sim")* [*(Região = "Sudeste")*]. O suporte da regra C_1 : *(Sexo = "M")* $\Rightarrow$ *(Fumante = "Sim")* é igual a 30%. O suporte da condição (Região = "Sudeste") é igual a 30%. O valor de MF^+ pode ser computador por: $1-((30\%\times30\%)\div20\%) = 0,55$. Para que a medida do grau de inesperabilidade seja obtida, primeiro deve ser calculado $DP(A, Z) =$ $DP(\{$*(Sexo = "M")*$\},\{$*(Região = "Sudeste")*$\}) = 1-((0,60\times0,30)\div0,20) = 0,10$. Em seguida, calcula-se $DP(B, Z) = DP(\{$*(Fumante = "Sim")*$\},\{$*(Região = "Sudeste")*$\}) =$ $1-((0,40\times0,30)\div0,20) = 0,40$. A exceção E_1 : *(Sexo = "M")* $\stackrel{+s}{\Longrightarrow}$ *(Fumante = "Sim")* [*(Região = "Sudeste")*] é, de fato, interessante, pois $GI^+(E_1) = 0,55 - max(0,10\ ,\ 0,40) = 0,15$.

3.3.3 Exceções com Relação a Fatos

Um fato representa uma condição definida sobre um ou mais atributos de uma base de dados. Os fatos, juntamente com a medida de suporte, podem ser utilizados para descrever informações contidas na base. Exemplos de fatos obtidos a partir da base de dados da AIDS são dados por: "84,91% dos pacientes contraíram AIDS através de relações sexuais" e "1,97% dos pacientes contraíram AIDS através de transfusão sanguínea".

Definição 3.3.10 *(Fato). Seja $\mathcal{D}$ uma relação pertencente a uma base de dados multidimensional. Um* ***fato****, obtido a partir de $\mathcal{D}$, é representado por um conjunto de n condições, $n \geq 1$, definidas sobre atributos distintos de $\mathcal{D}$, da seguinte forma:*

$$A_1 = a_1, ..., A_n = a_n,$$

onde A_i, $(1 \leq i \leq n)$, é um atributo de $\mathcal{D}$ e a_i representa um valor do domínio de A_i.

Um fato pode ser considerado forte ou fraco, de acordo com o valor de sua medida de suporte[1]. Quanto mais alto o valor de suporte, mais forte é o fato. O fato B_1 = *(Transmissão Sexual = "Sim")* é um dos mais fortes da base de dados da AIDS. Um total de 84,91% dos pacientes contraíram o vírus da HIV através de relações sexuais. Nesta mesma base, 1,97% dos pacientes contraíram o vírus através de transfusão sanguínea. Este fato, que é muito mais fraco do que o fato B_1, pode ser representado por B_2 = *(Transfusão = "Sim")*.

Também é possível minerar exceções negativas e positivas em relação a fatos. A presença de um conjunto de prova pode diminuir a probabilidade de ocorrência de um fato, tornando-o mais fraco. Ou pode aumentar esta probabilidade de ocorrência, tornando-o mais forte.

Definição 3.3.11 *(Exceção Negativa em Relação a um Fato). Seja $\mathcal{D}$ uma relação pertencente a uma base de dados multidimensional. Seja B um fato obtido a partir de $\mathcal{D}$. Seja $Z = \{Z_1 = z_1, ..., Z_k = z_k\}$ um conjunto de prova definido sobre atributos de $\mathcal{D}$, onde $Z \cap B = \emptyset$. Uma exceção negativa relacionada ao fato B é uma expressão com a seguinte forma: $\overset{-s}{\Longrightarrow} B\ [Z]$.*

Definição 3.3.12 *(Exceção Positiva em Relação a um Fato). Seja $\mathcal{D}$ uma relação pertencente a uma base de dados multidimensional. Seja B um fato obtido a partir de $\mathcal{D}$. Seja $Z = \{Z_1 = z_1, ..., Z_k = z_k\}$ um conjunto de prova definido sobre atributos de $\mathcal{D}$, onde $Z \cap B = \emptyset$. Uma exceção positiva relacionada ao fato B é uma expressão com a seguinte forma: $\overset{+s}{\Longrightarrow} B\ [Z]$.*

A mineração de exceções com relação a fatos, consiste em um caso especial da mineração de exceções com relação a regras de associação, pois o fato pode ser encarado como uma regra de associação cujo antecedente é um conjunto vazio. Os índices MF^- (Definição 3.3.5) e MF^+ (Definição 3.3.8) podem ser empregados para que sejam obtidas, respectivamente, exceções negativas e positivas em relação aos fatos. Por outro lado, não existe sentido em calcular o grau de inesperabilidade das

[1] O suporte de um fato A obtido a partir de uma relação $\mathcal{D}$, representa a porcentagem de tuplas de $\mathcal{D}$ que contêm A.

exceções com relação a fatos. Conforme foi introduzido pelas Definições 3.3.6 e 3.3.9, o grau de inesperabilidade é utilizado para avaliar a contribuição das dependências entre A e Z e entre B e Z, no valor obtido para o índice de dependência (seja negativa ou positiva) entre uma regra $A \Rightarrow B$ e um conjunto de prova Z. Como o fato é uma regra sem antecedente, não é possível calcular a dependência entre A e Z e, conseqüentemente, o cálculo do grau de inesperabilidade não tem sentido e não pode ser obtido.

Considere o conjunto de prova $Z_1 = \{$*(Idade = "0-12 anos")*$\}$, que representa a fração da base de dados da AIDS que contém apenas os pacientes com idade inferior a 13 anos. Na Seção 3.2, apresentou-se um exemplo que ilustrou que a probabilidade de uma pessoa ter contraído AIDS através de transmissão sexual é muito reduzida entre os pacientes nesta faixa etária. Esta situação pode ser representada pela seguinte exceção negativa:

$$E_1 \ : \overset{-s}{\Longrightarrow} \textit{(Transmissão Sexual = "Sim")} \ [\textit{(Idade = "0-12 anos")}].$$

A exceção negativa indica que o fato $B_1 =$ *(Transmissão Sexual = "Sim")* tem a sua força reduzida no subconjunto da base de dados definido pelo conjunto de prova $Z_1 = \{$*(Idade = "0-12 anos")*$\}$. Conforme apresentado na Seção 3.2, o resultado do cálculo do suporte esperado para o conjunto de condições $\{$*(Transmissão Sexual = "Sim"),(Idade = "0-12 anos")*$\}$ é igual a 3,42%. O suporte real deste mesmo conjunto na base de dados é igual a 0,0039%. A medida da força para a exceção negativa E_1, é, então, computada por $MF^{-}(E_1) = 1 - 0,0039 \div 3,42 = 0,9988$.

3.3.4 Definição do Problema

A seguir é apresentada uma definição formal para o problema da mineração de exceções negativas e positivas em bases de dados multidimensionais.

Definição 3.3.13 *(Problema da Mineração de Exceções Negativas e Positivas). Seja $\mathcal{D}$ uma relação pertencente a uma base de dados multidimensional. Seja $\mathcal{R}$ um conjunto de regras de associação da forma $A \Rightarrow B$, (se $A = \emptyset$, trata-se de um*

fato) obtidas a partir de $\mathcal{D}$. Sejam $SupMin \geq 0$, $MF_{min} \geq 0$ e $GI_{min} \geq 0$ parâmetros de entrada especificados pelo usuário. O problema da mineração de exceções negativas e positivas, a partir de $\mathcal{D}$ e $\mathcal{R}$, consiste em:

1. *Para cada regra de associação $A \Rightarrow B$, encontrar todas as exceções negativas da forma $EN : A \overset{-s}{\Longrightarrow} B\ [Z]$ onde:*

   (a) *(i) $Sup(A \cup Z) \geq SupMin$ e (ii) $Sup(B \cup Z) \geq SupMin$;*

   (b) *$MF^{-}(EN) \geq MF_{min}$;*

   (c) *$GI^{-}(EN) \geq GI_{min}$ (se $A \neq \emptyset$).*

2. *Para cada regra de associação $A \Rightarrow B$, encontrar todas as exceções positivas da forma $EP : A \overset{+s}{\Longrightarrow} B\ [Z]$ onde:*

   (a) *(i) $Sup(A \cup Z) \geq SupMin$ e (ii) $Sup(B \cup Z) \geq SupMin$;*

   (b) *$MF^{+}(EP) \geq MF_{min}$;*

   (c) *$GI^{+}(EP) \geq GI_{min}$ (se $A \neq \emptyset$).*

A restrição (a) é utilizada para garantir a significância estatística dos padrões minerados, tanto para as exceções negativas como para as positivas. A restrição (b) é utilizada para assegurar que o desvio entre os valores de suporte real e esperado de uma exceção seja suficientemente alto, garantindo assim a relevância dos padrões minerados. Por fim, a restrição (c) é aplicada com o objetivo de que exceções que representam informações inesperadas possam ser obtidas.

3.4 Procedimento

Nesta seção, é apresentado um procedimento para minerar, ao mesmo tempo, exceções negativas e positivas em bases de dados multidimensionais.

O procedimento utiliza os seguintes parâmetros de entrada:

1. $\mathcal{D}$ - uma relação pertencente a uma base de dados multidimensional;

2. $\mathcal{R}$ - um conjunto de regras de associação e fatos obtidos a partir de $\mathcal{D}$;

3. $\mathcal{A}$ - um conjunto de atributos selecionados de $\mathcal{D}$;

4. $TMax$ - número máximo permitido de condições em um conjunto de prova;

5. $SupMin$ - suporte mínimo;

6. MF_{min} - valor mínimo para as medidas MF^- e MF^+;

7. GI_{min} - valor mínimo para as medidas GI^- e GI^+;

Como saída, são apresentados os seguintes resultados:

1. *ExceçõesNegativas* - um conjunto de exceções negativas;

2. *ExeceçõesPositivas* - um conjunto de exceções positivas.

Este procedimento é composto por três fases, executadas em seqüência: geração das exceções candidatas, contagem de suporte e mineração das exceções.

3.4.1 Fase 1

A Figura 3.1 apresenta os passos da primeira fase do procedimento para mineração de exceções positivas e negativas.

O primeiro passo (linha 1) determina todos os conjuntos de prova que serão utilizados na geração das exceções candidatas, a partir dos atributos especificados pelo usuário através do parâmetro $\mathcal{A}$. O parâmetro $TMax$ é utilizado para restringir o tamanho máximo dos conjuntos gerados. Os valores dos diferentes atributos especificados devem ser combinados para que sejam identificadas todas as possibilidades de conjuntos de prova de tamanho menor ou igual a $TMax$. Os atributos numéricos e contínuos devem ser discretizados antes do processo de mineração. Considere por exemplo a base de dados apresentada na Tabela 3.1. Suponha que o usuário tenha escolhido os atributos **Sexo** e **Carro** para formar o conjunto $\mathcal{A}$ e tenha configurado o parâmetro $TMax$ com o valor 2. O conjunto *ConjuntosDeProva*,

Algoritmo 1 : *Geração das exceções candidatas*

```
1   ConjuntosDeProva ← GerarConjuntosDeProva(A,TMax);
2   Candidatas ← ∅;
3   ÁrvoreDeCondições ← ∅;
4   para cada regra A ⇒ B em R faça
5     para cada conjunto de prova Z em ConjuntosDeProva faça
6       C ← A ⇒ B [Z];
7       Candidatas ← Candidatas ∪ C;
8       X ← {{A},{B},{Z},{A,B},{A,Z},{B,Z},{A,B,Z}};
9       ÁrvoreDeCondições ← ÁrvoreDeCondições ∪ X;
10    fim para
11  fim para
```

Figura 3.1: Geração das exceções candidatas.

que armazena todos os possíveis conjuntos de prova, seria então formado por oito elementos: *ConjuntosDeProva* = {*{(Sexo="F")}, {(Sexo="M")}, {(Carro="Não")}, {(Carro="Sim")}, {(Sexo="F"),(Carro="Não")}, {(Sexo="F"),(Carro="Sim")}, {(Sexo="M"),(Carro="Não")}, {(Sexo="M"),(Carro="Sim")}*}.

As linhas 2 a 11 apresentam a rotina utilizada para a geração das exceções candidatas. Os conjuntos *Candidatas* e *ÁrvoreDeCondições* são inicializados com valor vazio nas linhas 2 e 3, respectivamente. O último conjunto será explicado um pouco mais adiante. Já o primeiro conjunto é utilizado para armazenar todas as exceções candidatas. Estas exceções candidatas são geradas através da combinação de todos os conjuntos de prova contidos em *ConjuntosDeProva* com todas as regras e fatos contidos no parâmetro de entrada $\mathcal{R}$. O laço iniciado na linha 4 é utilizado para percorrer cada elemento de $\mathcal{R}$. O laço iniciado na linha 5 é utilizado para que cada conjunto de prova contido em *ConjuntosDeProva* seja obtido. Cada exceção candidata é então gerada através da combinação de uma regra de associação ou fato com um conjunto de prova (linha 6) e, em seguida, armazenada no conjunto *Candidatas* (linha 7).

Para melhor ilustrar a operação executada pelas linhas 6 e 7 do procedimento, será apresentado um exemplo. Considere, mais uma vez, a base de dados da

Tabela 3.1. Suponha que se deseja minerar exceções a partir desta base, utilizando os seguintes parâmetros de entrada: $\mathcal{R} = \{$*(Sexo = "F") ⇒ (Valor = "≥100,00")*$\}$ e $\mathcal{A} = \{$*Carro*$\}$. A Figura 3.2 apresenta o estado do conjunto *Candidatas*, ao final da execução da primeira fase do procedimento de mineração de exceções. Duas exceções candidatas estariam armazenadas neste conjunto.

- (Sexo = "F") ⇒ (Valor = "≥100,00") [(Carro = "Sim")]
- (Sexo = "F") ⇒ (Valor = "≥100,00") [(Carro = "Não")]

Figura 3.2: Exceções candidatas armazenadas no conjunto *Candidatas*.

As linhas 8 e 9 são utilizadas para atualizar a estrutura de dados *ÁrvoreDeCondições*. Para que as medidas de interesse de uma exceção candidata $A \Rightarrow B\ [Z]$ possam ser calculadas, é necessário que os valores de suporte dos conjuntos {A}, {B}, {Z}, {A,B}, {A,Z}, {B,Z} e {A,B,Z} sejam conhecidos. A estrutura de dados *ÁrvoreDeCondições* é utilizada para armazenar, de maneira eficiente, contadores de suporte para todos estes conjuntos de condições. Cada exceção candidata que for gerada deve ter todas as combinações das condições que a compõem devidamente armazenadas nesta estrutura de dados.

Considere, mais uma vez, o exemplo que envolve os parâmetros de entrada $\mathcal{R} = \{$*(Sexo = "F") ⇒ (Valor = "≥100,00")*$\}$, $\mathcal{A} = \{$*Carro*$\}$ e a base de dados da Tabela 3.1. A Figura 3.3 relaciona todas as combinações de condições que estariam contidas na *ÁrvoreDeCondições* ao final da execução da primeira fase do algoritmo para mineração de exceções negativas e positivas. Note que a estrutura conteria onze conjuntos de condições. Quatro destes conjuntos de condições envolvem apenas um atributo, cinco conjuntos envolvem dois atributos e dois conjuntos envolvem três atributos. A *ÁrvoreDeCondições* deve armazenar contadores de suporte para todos estes conjuntos de condições. O Capítulo 4 apresenta detalhes sobre a implementação desta estrutura.

- {(Sexo = “F”)}
- {(Valor = “≥100,00”)}
- {(Carro = “Sim”)}
- {(Carro = “Não”)}
- {(Sexo = “F”),(Valor = “≥100,00”)}
- {(Sexo = “F”),(Carro = “Sim”)}
- {(Sexo = “F”),(Carro = “Não”)}
- {(Valor = “≥100,00”),(Carro = “Sim”)}
- {(Valor = “≥100,00”),(Carro = “Não”)}
- {(Sexo = “F”),(Valor = “≥100,00”),(Carro = “Sim”)}
- {(Sexo = “F”),(Valor = “≥100,00”),(Carro = “Não”)}

Figura 3.3: Conjuntos de condições armazenados na estrutura *ÁrvoreDeCondições*.

3.4.2 Fase 2

Na segunda fase do procedimento (Figura 3.4), a base de dados é inteiramente varrida, com o objetivo de realizar a contagem de suporte de todos os conjuntos de condições que estão armazenados na *ÁrvoreDeCondições*. É importante observar que apenas uma varredura à base é necessária, visto que os conjuntos de condições que precisam ter os valores de suporte contabilizados já são conhecidos.

Algoritmo 2 : *Contagem de suporte*

```
1  para cada tupla t em D faça
2    atualizar contadores dos conjuntos em ÁrvoredeCondições;
3  fim para
```

Figura 3.4: Contagem de suporte.

3.4.3 Fase 3

A Figura 3.5 ilustra a terceira fase do procedimento de mineração de exceções negativas em positivas.

Algoritmo 3 : *Geração das exceções*

```
1  para cada exceção candidata C : A ⇒ B [Z] em Candidatas faça
2     se Sup(A, Z) ≥ SupMin e Sup(B, Z) ≥ SupMin então
3        se Sup(C) < SupEsp(C) então
4           se MF⁻(C) ≥ MFmin e GI⁻(C) ≥ GImin então
5              ExceçõesNegativas ← ExceçõesNegativas ∪ A ⟹^{-s} B [Z];
6           fim se
7        fim se
8        senão se Sup(C) > SupEsp(C) então
9           se MF⁺(C) ≥ MFmin e GI⁺(C) ≥ GImin então
10             ExceçõesPositivas ← ExceçõesPositivas ∪ A ⟹^{+s} B [Z];
11          fim se
12       fim se
13    fim se
14 fim para
```

Figura 3.5: Geração das exceções negativas e positivas.

Após a geração das exceções candidatas e a contagem de suporte das condições, nesta fase as exceções negativas e positivas finalmente são mineradas. O laço iniciado na linha 1 é utilizado para percorrer todo o conjunto de condições candidatas, geradas na primeira fase do procedimento. Cada uma das exceções candidatas é analisada, da seguinte maneira: primeiro são verificadas as restrições de suporte (linha 2). Em seguida, é realizado um teste para descobrir se o valor do suporte real da exceção candidata é menor do que seu valor de suporte esperado (linha 3). Neste caso, as medidas MF^- e GI^- são computadas (linha 4). Para a realização destes cálculos, o suporte dos conjuntos de condições armazenados em *ÁrvoreDeCondições* devem ser consultados (estes valores de suporte foram obtidos na segunda fase do procedimento). De acordo com o resultado apresentado pelas medidas MF^- e GI^-, uma exceção negativa pode ser minerada e armazenada no conjunto *ExceçõesNega-*

tivas (linha 5). Caso o valor do suporte real da exceção candidata que está sendo correntemente analisada seja maior do que o valor do suporte esperado (linha 8), as medidas MF^+ e GI^+ são computadas (linha 9). De acordo com o resultado deste cálculo, uma exceção positiva pode ser minerada e armazenada no conjunto *ExceçõesPositivas* (linha 10). É importante lembrar que, no caso da mineração de exceções com relação a fatos, não é necessário computar o grau de inesperabilidade (as medidas GI^- ou GI^+).

O procedimento para mineração de exceções aqui apresentado é um exemplo de procedimento cooperativo, pois o foco de todo o processo de mineração é especificado pelo usuário. Ao usuário é dada a possibilidade de escolher um conjunto de regras de associação e fatos cuja variação da força ele está interessado em investigar, em diferentes subconjuntos da base de dados. O usuário também é responsável por especificar estes subconjuntos da base de dados, selecionando os atributos que são utilizados para gerar os conjuntos de prova. O procedimento de mineração de exceções realiza um esforço proporcional às escolhas realizadas pelo usuário, especialmente concentrado na criação da *ÁrvoreDeCondições* e na contagem do suporte dos conjuntos de condições que nela estão armazenados. Apenas os atributos que compõem as exceções candidatas são relevantes e, com isso, um número muito menor de conjuntos de condições necessita ter o suporte contabilizado, ao contrário do que ocorre em um algoritmo convencional de mineração de regras de associação. Além disso, apenas uma varredura na base de dados é suficiente para contabilizar estes valores de suporte.

3.5 Exceções com Relação à Medida de Confiança

A técnica para mineração de exceções apresentada na Seção 3.3 utiliza os valores de suporte de uma regra de associação e de um conjunto de prova, para identificar exceções interessantes. Estas exceções são mineradas quando o suporte de uma regra de associação torna-se muito mais alto ou muito mais baixo numa determinada fatia da base de dados. No entanto, existem situações em que o valor

de suporte de uma regra de associação permanece quase que constante em diferentes subconjuntos da base, mas, ao mesmo tempo, o valor de sua medida de confiança apresenta um largo desvio em relação à expectativa.

Este argumento será ilustrado através de um novo exemplo, obtido a partir dos dados de um estudo descrito em [25]. Este estudo analisou o efeito de características raciais na condenação à pena de morte de indivíduos que cometeram homicídio nos Estados Unidos. Os dados pesquisados correspondem a decisões de juízes relativas a 326 casos ocorridos no estado da Flórida, nos anos de 1976 e 1977. A Tabela 3.3, que está disponibilizada em [3], contém a consolidação dos dados obtidos pela pesquisa de [25].

Raça do Acusado	**Raça da Vítima**	**Pena de Morte**	
		SIM	**NÃO**
Branco	Branco	5,83%	40,50%
Branco	Negro	0,00%	2,76%
Negro	Branco	3,37%	15,95%
Negro	Negro	1,84%	29,75%
	Totais:	11,04%	88,96%

Tabela 3.3: Base de dados de condenações à pena de morte.

O estudo de [25] analisou o efeito de três variáveis: **raça do acusado**, **raça da vítima** e **veredicto de pena de morte**. Observe a terceira coluna da primeira linha da tabela (a linha logo abaixo do cabeçalho). O valor de 5,83% constante nesta coluna, indica que 5,83% dos casos correspondem a acusados da raça branca, que assassinaram indivíduos também da raça branca e acabaram condenados a pena de morte. A última linha da tabela indica que 11,04% de todos os acusados (sejam eles brancos ou negros) foram condenados a pena de morte e que, conseqüentemente, os demais 89,96% não receberam a pena fatal.

Duas regras de associação, que podem ser obtidas a partir desta base, são da-

das por: R_1 : *(Acusado = "Negro")* $\Rightarrow$ *(Pena de Morte = "Sim")* ($Sup = 5,21\%$ e $Conf = 10,24\%$) e R_2 : *(Acusado = "Branco")* $\Rightarrow$ *(Pena de Morte = "Sim")* ($Sup = 5,83\%$ e $Conf = 11,87\%$). Tem-se então que a porcentagem de penas fatais atribuída a acusados negros foi ligeiramente inferior a porcentagem atribuída aos acusados da raça branca.

Um algoritmo para a mineração de exceções pode ser empregado para investigar a força da regra R_1 : *(Acusado = "Negro")* $\Rightarrow$ *(Pena de Morte = "Sim")* em subconjuntos da base de dados definidos de acordo com a raça da vítima. Considere, por exemplo, o conjunto de prova $Z = \{$*(Vítima = "Branca")*$\}$, que possui suporte igual a 65,64%. O suporte esperado para a exceção candidata C_1 : *(Acusado = "Negro")* $\Rightarrow$ *(Pena de Morte = "Sim")* [*(Vítima = "Branca")*] é obtido calculando-se $SupEsp(C_1) = Sup(R_1) \times Sup(Z) = 5,83\% \times 65,64 = 3,43\%$. O valor de suporte real desta exceção candidata é igual a 3,37%. Como $SupEsp(C_1) \approx Sup(C_1)$, então o suporte de R_1 não desviou de seu valor esperado no subconjunto da base de dados formado pelas vítimas brancas. Desta forma, nenhuma exceção é minerada.

No entanto, a medida da confiança da regra R_1 aumenta para 17,43% no subconjunto da base de dados formado pelas vítimas brancas. Isto quer dizer que a chance de um negro ser condenado a pena de morte sobe de 10,24% para 17,43% quando a vítima é da raça branca. O valor de suporte da regra R_1 não desviou do esperado no subconjunto da base de dados formado pelas vítimas da raça branca. Porém o valor da confiança cresceu significativamente (um aumento de 41,24%). Nesta situação, uma nova categoria de exceções passa a ser definida: as **exceções com relação à medida de confiança**. O exemplo da pena de morte pode ser representado pela seguinte notação:

(Acusado = "Negro") $\overset{+c}{\Longrightarrow}$ *(Pena de Morte = "Sim")* [*(Vítima = "Branco")*]

O símbolo "$\overset{+c}{\Longrightarrow}$" é utilizado para indicar uma exceção positiva com relação à medida de confiança.

O conceito de **confiança esperada** foi introduzido em [21] da seguinte

maneira: a confiança esperada de uma regra é definida a partir da relação entre o suporte esperado do conjunto de condições que a compõem e o suporte esperado do conjunto de condições que compõem seu antecedente. Este conceito será aplicado de maneira análoga para a definição da confiança esperada de uma exceção candidata.

Definição 3.5.1 *(Confiança Esperada de uma Exceção Candidata). Seja* $C : A \Rightarrow B\ [Z]$ *uma exceção candidata. A confiança esperada de* C*, denotada por* $ConfEsp(C)$*, é obtida por:*

$$ConfEsp(C) = \frac{SupEsp(A \Rightarrow B\ [Z])}{SupEsp(A \cup Z)} = \frac{Sup(A \Rightarrow B) \times Sup(Z)}{Sup(A) \times Sup(Z)} = Conf(A \Rightarrow B)\ .$$

Observa-se que o valor esperado para a medida de confiança de uma exceção candidata $A \Rightarrow B\ [Z]$ é igual ao valor da medida de confiança da regra $A \Rightarrow B$ (que é a regra que está sendo avaliada contra o conjunto de prova Z).

Caso o valor da confiança de uma regra de associação seja significativamente inferior ao esperado, uma exceção negativa com relação à confiança é caracterizada. O símbolo "$\overset{-c}{\Longrightarrow}$" é utilizado para indicar este tipo de exceção e a **Medida de Desvio da Confiança para Exceções Negativas** é utilizada para avaliar a relevância destas exceções.

Definição 3.5.2 *(Desvio de Confiança para Exceções Negativas). Seja* $C : A \Rightarrow B\ [Z]$ *uma exceção candidata. O desvio negativo de confiança para* C*, representado por* $DC^{-}(C)$*, é computado por:*

$$DC^{-}(C) = 1 - \left(\frac{Conf(A \Rightarrow B\ [Z])}{ConfEsp(A \Rightarrow B\ [Z])} \right)\ .$$

O resultado do cálculo da medida DC^{-} para uma exceção indica, em percentual, o quanto o valor da confiança real de uma exceção candidata é inferior à sua confiança esperada. Quanto mais próximo o valor está de 1, mais interessante é a exceção negativa.

Caso o valor da confiança de uma regra de associação seja significativamente superior ao esperado, uma exceção positiva com relação à confiança é caracterizada.

Este é o caso do exemplo apresentado no início desta seção. A **Medida de Desvio de Confiança para Exceções Positivas** é utilizada para avaliar a relevância destas exceções.

Definição 3.5.3 *(Desvio de Confiança para Exceções Positivas). Seja* $C : A \Rightarrow B\ [Z]$ *uma exceção candidata. O desvio positivo de confiança para* C*, representado por* $DC^+(C)$*, é computado por:*

$$DC^+(C) = 1 - \left(\frac{ConfEsp(A \Rightarrow B\ [Z])}{Conf(A \Rightarrow B\ [Z])} \right) .$$

O resultado do cálculo da medida DC^+ para uma exceção indica, em percentual, o quanto o valor da confiança real de uma exceção candidata é superior à sua confiança esperada. Quanto mais próximo o valor está de 1, mais interessante é a exceção positiva. No exemplo da pena de morte, o valor da medida DC^+ pode ser obtido por $1 - (10,24 \div 17,43) = 0,4124$.

A técnica para mineração de exceções negativas e positivas definida na Seção 3.3, assim como o procedimento descrito na Seção 3.4, podem ser estendidos para que também possibilitem a mineração de exceções com relação à confiança. Não se trata de uma tarefa difícil, visto que o processo utilizado para calcular o desvio entre o valor real e o valor esperado da confiança das exceções candidatas é bastante simples. A estrutura de dados *ÁrvoreDeCondições*, descrita na Seção 3.4, guarda todos os contadores necessários ao cálculo dos índices DC^- e DC^+.

Exceções a fatos não podem ser mineradas através da utilização do desvio da confiança. A medida de suporte é a única utilizada para determinar a força de um fato. Não existe o conceito de confiança para um fato, visto que o fato poderia ser considerado uma regra de associação sem antecedente. Desta forma, não é possível realizar o cálculo dos índices DC^- ou DC^+ para uma exceção a um fato.

3.6 Categorias de Exceções

Na abordagem estendida de mineração de exceções em bases de dados, estas são mineradas quando o valor do suporte ou da confiança de uma regra de associação desviam-se significativamente de uma determinada expectativa. Em algumas situações, o suporte de uma regra pode permanecer constante em diferentes subconjuntos da base de dados, enquanto o valor de sua medida de confiança apresenta grande variação. Em outras situações, a medida de suporte pode variar significativamente, enquanto a confiança permanece constante.

Esta situação motivou a elaboração de uma categorização para os diferentes tipos de exceções que podem ser mineradas em bases de dados multidimensionais. As categorias de exceções são apresentadas na Tabela 3.4.

Exceções Negativas	**Exceções Positivas**	**Exceções Híbridas**
1) $A \overset{-s}{\Longrightarrow} B\ [Z]$	5) $A \overset{+s}{\Longrightarrow} B\ [Z]$	9) $A \overset{-s+c}{\Longrightarrow} B\ [Z]$
2) $A \overset{-c}{\Longrightarrow} B\ [Z]$	6) $A \overset{+c}{\Longrightarrow} B\ [Z]$	10) $A \overset{+s-c}{\Longrightarrow} B\ [Z]$
3) $A \overset{-s-c}{\Longrightarrow} B\ [Z]$	7) $A \overset{+s+c}{\Longrightarrow} B\ [Z]$	
4) $\overset{-s}{\Longrightarrow} B\ [Z]$	8) $\overset{+s}{\Longrightarrow} B\ [Z]$	

Tabela 3.4: Categorias de exceções.

A primeira coluna da Tabela 3.4 exibe a categoria das exceções negativas. As exceções negativas do tipo 1 são relativas apenas à medida de suporte, enquanto as do tipo 2 são relativas apenas à medida de confiança. Já as exceções do tipo 3 são negativas com relação ao suporte e à confiança. As exceções negativas do tipo 4 são referentes a um fato.

De maneira análoga, a segunda coluna apresenta a categoria das exceções positivas. As exceções positivas do tipo 5 são relativas apenas à medida de suporte e as do tipo 6 apenas à medida de confiança. Já as exceções do tipo 7 são positivas com relação ao suporte e à confiança. As exceções positivas do tipo 8 são referentes a um fato.

A terceira coluna introduz as **exceções híbridas**, utilizadas para caracterizar regras que se tornam mais fortes com relação a uma das medidas (suporte ou confiança) e, ao mesmo tempo, mais fracas com relação à outra medida. No caso indicado pelo tipo 9, a exceção representa uma regra cujo suporte tornou-se mais fraco e a confiança tornou-se mais alta. Já na exceção do tipo 10, o suporte tornou-se mais alto e a confiança mais baixa.

Capítulo 4

Resultados Experimentais

Este capítulo apresenta os resultados obtidos através da mineração de exceções negativas e positivas sobre bases de dados de diferentes áreas, como Biologia, Medicina e Ciências Sociais. O capítulo está dividido da seguinte forma. A Seção 4.1 fornece algumas características a respeito da implementação do procedimento para mineração de exceções, enquanto a Seção 4.2 descreve as bases de dados mineradas. Por fim, na Seção 4.3, os resultados são apresentados e comentados.

4.1 Implementação

O procedimento para mineração de exceções negativas e positivas, apresentado no capítulo anterior, foi implementado na linguagem C++, com a utilização do compilador g++.

4.1.1 Formato da Base de Dados

Por questões de praticidade, o padrão ARFF (*Attribute-Relation File Format*), especificado em [37], foi adotado como formato para as bases de dados a serem mineradas. O padrão ARFF vem sendo amplamente utilizado no meio acadêmico ao longo dos últimos anos. Ele também é adotado pelo sistema *Weka* [36], uma

conhecida ferramenta de mineração de dados de código aberto (*open source*).

Uma base de dados ARFF é um arquivo ASCII cuja extensão é .arff e que descreve uma lista de instâncias que compartilham um conjunto de atributos. Este tipo de base de dados possui duas seções distintas: a primeira seção constitui-se das informações de cabeçalho (seção HEADER), que é seguida das informações dos dados (seção DATA). A Figura 4.1 ilustra um exemplo de arquivo ARFF. Este exemplo representa a base de dados de vendas, apresentada na Tabela 3.1.

```
@relation Vendas

@attribute Id {numeric}
@attribute Sexo {categorical}
@attribute Idade {categorical}
@attribute Carro {categorical}
@attribute Cartão de Crédito {categorical}
@attribute Valor da Compra {categorical}

@data
1 , 'F' , '<18' , 'Não' , 'Credicard' , '≥100,00'
2 , 'M' , '26-30' , 'Sim' , 'Visa' , '≥100,00'
3 , 'F' , '26-30' , 'Sim' , 'Visa' , '25,00 - 49,00'
4 , 'F' , '18-25' , 'Sim' , 'Amex' , '≥100,00'
5 , 'F' , '31-40' , 'Sim' , 'Credicard' , '≥100,00'
6 , 'M' , '<18' , 'Não' , 'Visa' , '≥100,00'
7 , 'F' , '26-30' , 'Sim' , 'Amex' , '≥100,00'
8 , 'F' , '<18' , 'Não' , 'Visa' , '≤25,00'
9 , 'M' , '18-25' , 'Sim' , 'Visa' , '50,00-99,00'
10 , 'M' , '<18' , 'Não' , 'Credicard' , '≤25,00'
```

Figura 4.1: Base de dados ARFF.

A seção HEADER de um arquivo ARFF é formada por um comando "@relation", que contém a declaração da relação que o arquivo representa. Esta seção também é composta por uma série de comandos "@attribute", que especificam o

nome e o tipo de todos os atributos que formam a relação. A seção DATA (iniciada pelo comando "@data") contém todas as instâncias da base de dados. Os valores dos atributos para cada instância são separados por vírgulas.

4.1.2 EXCEPMINER

A implementação em C++ do procedimento para mineração de exceções recebeu o nome de EXCEPMINER. O programa necessita, além de uma base de dados no formato ARFF, dos demais parâmetros de entrada apresentados na Seção 3.4: um conjunto de fatos ou regras de associação que serão avaliadas ($\mathcal{R}$), um conjunto de atributos que formarão os conjuntos de prova ($\mathcal{A}$), o tamanho máximo dos conjuntos de prova ($TMax$) e os valores mínimos para as medidas de suporte ($SupMin$), força (MF_{min}) e grau de inesperabilidade (GI_{min}).

4.1.3 ÁrvoreDeCondições

Conforme apresentado na Seção 3.4, o procedimento para mineração de exceções negativas e positivas é dividido em três fases. Na primeira fase do procedimento, a estrutura de dados *ÁrvoreDeCondições* é criada. Esta estrutura é utilizada para armazenar contadores de suporte para uma série de conjuntos de condições que compõem as exceções candidatas. Na Seção 3.4.1 foi apresentado um exemplo que demonstrou o estado da *ÁrvoreDeCondições* ao final da execução da primeira fase do procedimento para mineração de exceções negativas e positivas. Este exemplo foi obtido a partir da base de dados da Tabela 3.1 e considerou os seguintes parâmetros de entrada: $\mathcal{R} = \{$*(Sexo = "F")* $\Rightarrow$ *(Valor = "≥100,00")*$\}$ e $\mathcal{A} = \{$*Carro*$\}$. A Figura 3.3 apresentou todos os conjuntos de condições que estariam armazenados na *ÁrvoreDeCondições*. Os conjuntos de condições envolvem um, dois ou três atributos.

No EXCEPMINER, a *ÁrvoreDeCondições* foi implementada da seguinte forma. Um vetor é utilizado para armazenar contadores de suporte para os conjuntos de condições que envolvem apenas um atributo. Já os contadores de suporte para os conjuntos de condições que envolvem dois ou mais atributos, são armazenados

numa estrutura denominada árvore *hash*, definida em [2].

Inicialmente, cada um dos conjuntos que contém apenas uma condição é associado a um identificador único (um número inteiro). A Tabela 4.1, ilustra o resultado desta operação para o exemplo da Seção 3.4.1. Neste caso, bastaria instanciar um vetor de tamanho 4 para contabilizar o suporte de cada um dos conjuntos de condição que envolvem apenas um atributo.

Conjunto de Condições	**Código Identificador**
{(Sexo = "F")}	0
{(Valor = "≥100,00")}	1
{(Carro = "Sim")}	2
{(Carro = "Não")}	3

Tabela 4.1: Código identificador único para conjuntos de condições envolvendo um atributo.

Para armazenar os contadores de suporte dos conjuntos contendo duas ou mais condições, o EXCEPMINER utilizou uma lista de árvores *hash*. Um exemplo de árvore *hash* é apresentado na Figura 4.2. Esta estrutura é definida da seguinte forma. Um nó interno da árvore *hash* é uma tabela *hash* onde cada entrada aponta para outro nó. O número de entradas nessa tabela é denominado **ordem** da árvore. Um nó folha nesta árvore contém uma lista de conjuntos de condições. A cada conjunto de condições é associado um contador para armazenar a sua freqüência.

A raiz da árvore *hash* é definida com **profundidade** zero. Um nó interno de profundidade d aponta para um nó de profundidade $d + 1$. Uma árvore *hash* que armazena conjuntos de condições envolvendo k atributos deve ter profundidade máxima igual a k.

Os conjuntos de condições são inseridos na árvore *hash* através da aplicação de uma função hash sobre os elementos pertencentes a estes conjuntos. Um conjunto de condições é inserido na folha da árvore *hash* caso o número de conjuntos já

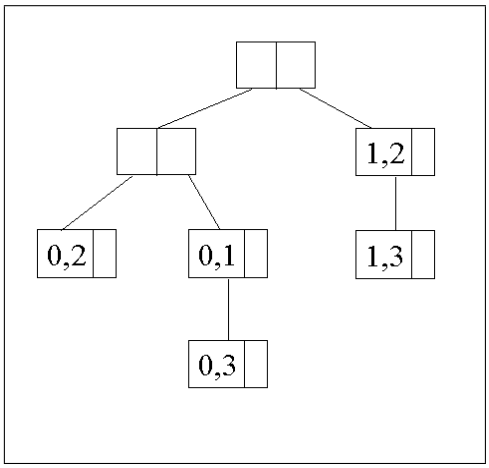

Figura 4.2: Árvore *hash*.

armazenado nesta folha não tenha alcançado um determinado limite, denominado **saturação** da folha. Se o número de conjuntos de condições na folha ultrapassar a saturação e a profundidade da árvore não for a máxima, então esta folha é convertida em um nó interno e nós filhos são criados para este novo nó interno. Se a profundidade máxima foi alcançada, o conjunto de condições é inserido na folha sem que o parâmetro saturação seja levado em consideração. Uma explicação detalhada sobre a implementação de uma árvore hash pode ser encontrada em [34].

Um exemplo será apresentado para melhor ilustrar este processo. Suponha que os cinco conjuntos de condições que envolvem dois atributos, referentes ao exemplo da Seção 3.4.1, devam ser armazenados em uma árvore *hash* com ordem e saturação iguais a 2. Desta forma, a Figura 4.2 corresponde a uma representação exata da árvore *hash* que armazenaria estes cinco conjuntos de condições.

Neste exemplo, o nó que contém o conjunto $\{0, 2\}$ é utilizado para armazenar o suporte do conjunto de condições {*(Sexo = "F"),(Carro = "Sim")*}, já que, de acordo com o mapeamento indicado na Tabela 4.1, o código da condição *(Sexo = "F")* é igual a 0 e o código da condição *(Carro = "Sim")* é igual a 2. Considere que a função *hash* associada à árvore retorne o resto da divisão de um elemento do conjunto de condições pela ordem da árvore. A árvore *hash* é, então, definida da seguinte forma: o nó esquerdo contém os conjuntos de condições cujo primeiro elemento é 0 e o nó direito contém os conjuntos de condições cujo primeiro elemento é 1.

4.2 Bases de Dados

Esta seção descreve brevemente as bases de dados que foram utilizadas no processo de mineração de exceções.

4.2.1 Base de Dados da AIDS

O **Programa Brasileiro para Doenças Sexualmente Transmissíveis e AIDS** [24] é responsável pelo combate ao HIV e à AIDS no Brasil. Dentre as bases de dados referentes a pesquisas mantidas pelo programa, existe uma que armazena informações sobre pessoas portadoras do vírus da AIDS no Brasil, que foram diagnosticadas entre 1980 e 2001. Foram selecionados 172.563 registros e 10 atributos, entre estes os que identificam o grupo de risco, sexo, raça e idade de cada uma das pessoas. Apenas os pacientes considerados adultos (com idade superior a 13 anos) foram selecionados para os testes realizados neste capítulo.

4.2.2 Base de Dados dos Cogumelos

A base de dados dos cogumelos (*mushroom data set*) [5] inclui descrições de exemplos hipotéticos de cogumelos que podem ser encontrados na natureza. A família de cogumelos é bastante extensa, sendo conhecidas cerca de 100.000 espécies. Nenhum teste pode revelar se um cogumelo é venenoso ou comestível. Apenas os caracteres botânicos permitem identificar a espécie com segurança [26]. A Figura 4.3 apresenta alguns destes caracteres botânicos: chapéu, lamelas (lâminas situadas abaixo do chapéu), anel, haste e esporos (material reprodutor). Além disso, os cogumelos podem apresentar diversos cheiros e podem ser encontrados em diferentes ambientes.

A base de dados é composta por 8.124 tuplas. Os caracteres botânicos dos cogumelos são descritos através de 23 atributos categóricos. Entre eles, está um atributo alvo que é utilizado para classificar cada cogumelo como comestível ou venenoso.

Ao longo dos últimos anos, a base de dados dos cogumelos vem sendo freqüentemente utilizada como *benchmark* para algoritmos de classificação. Devido à sua popularidade, esta base foi escolhida em [15], com o objetivo de testar a técnica para mineração de exceções negativas que foi apresentado nesta dissertação.

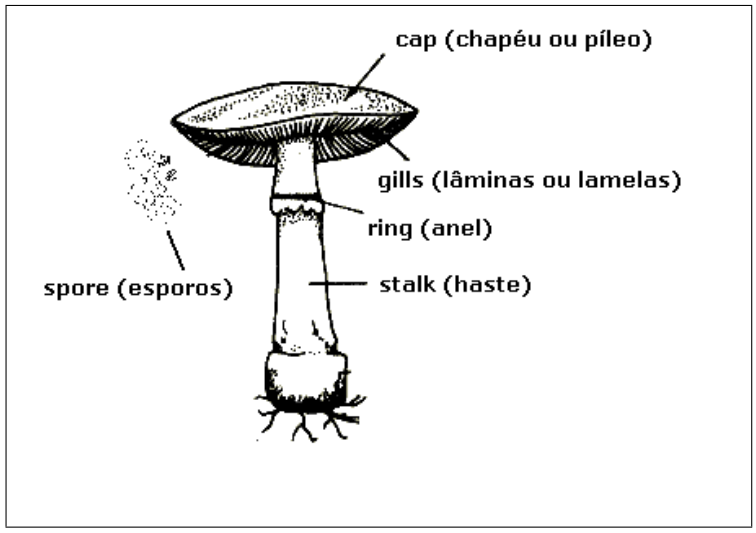

Figura 4.3: Cogumelo com lamelas.

4.2.3 Base de Dados da Aterosclerose

A **aterosclerose** é uma doença que atinge as artérias de grande e médio calibre, como as artérias coronárias, as artérias carótidas e as artérias dos membros inferiores. Esta doença é caracterizada pelo depósito de gordura, cálcio e outros elementos nas paredes das artérias, reduzindo seu calibre e trazendo um déficit sanguíneo aos tecidos irrigados por elas [29, 30]. A aterosclerose é uma das principais causas de morte no mundo ocidental.

Em meados dos anos 70, um projeto visando a prevenção da aterosclerose foi iniciado em *Praga*, capital da República Tcheca. Este estudo foi batizado como STULONG - *Longitudinal Study of Atherosclerosis Risk Factors* [10]. O projeto STULONG durou mais de 20 anos. O primeiro passo deste projeto foi a realização de exames iniciais (*entry examinations*) em 1417 pacientes com idade entre 38 e 53 anos, entre os anos de 1975 e 1979. Nestes exames iniciais, os pacientes foram requisitados a preencher um formulário com seus dados pessoais (como estado civil,

grau de instrução e data de nascimento) e hábitos gerais (consumo de cigarro e álcool, prática de atividades físicas, etc.). Estes pacientes também foram submetidos a um exame físico completo (peso, altura, índice de massa corpórea, etc.), a um exame bioquímico (sangue, urina, etc.) e a um eletrocardiograma.

Os seguintes aspectos foram definidos pelos especialistas do projeto STULONG como fatores de risco da aterosclerose: hipertensão arterial, hipercolesteromia, glicemia, alto nível de ácido úrico, hipertrigliceridemia, obesidade, caso positivo na família e o hábito de fumar muitos cigarros por dia até um ano antes dos exames iniciais. De acordo com estes fatores de risco e com os resultados dos exames iniciais, o pacientes foram classificados em três grupos:

A. *Grupo Normal.* Homens que não apresentaram nenhum fator de risco.

B. *Grupo de Risco.* Homens que apresentaram um ou mais fatores de risco, mas que não manifestaram nenhuma doença cardiovascular.

C. *Grupo Patológico.* Homens com alguma doença cardiovascular ou outra doença séria identificada.

Os pacientes classificados nos grupos A e B foram submetidos a exames periódicos ao longo das duas décadas de duração do projeto STULONG. Tanto os dados relativos aos exames iniciais como aqueles relativos aos exames periódicos foram disponibilizados ao público [8]. Com isto, diversas técnicas de mineração de dados foram aplicadas sobre a base de dados e os especialistas do projeto STULONG puderam obter novos conhecimentos. Em [16], a técnica de mineração de exceções proposta nesta dissertação foi utilizada para se obter uma série de exceções a partir da tabela que contém os dados relativos aos exames iniciais. Neste mesmo trabalho, é possível encontrar regras de associação fortes, que foram mineradas de acordo com a aplicação das medidas de interesse $lift$ e RI, comentadas no Capítulo 2.

4.2.4 Base de Dados do Censo de Washington

Esta base de dados, também conhecida como *adult database* ou *census income data set*, está disponibilizada em [5]. Ela contém dados do censo que foi realizado no ano de 1990, na cidade de Washington nos Estados Unidos. Cada tupla apresenta informações relativas a um indivíduo entrevistado pelo censo.

A base de dados é composta por 48.842 tuplas e 14 atributos, que indicam, entre outras características, a idade, educação, estado civil e ocupação de cada pessoa. Um atributo alvo é utilizado para especificar se o entrevistado possui renda anual superior a US$ 50.000,00.

4.3 Resultados

Esta seção apresenta resultados selecionados dentre os que foram obtidos através da mineração de exceções sobre as bases de dados previamente descritas. Estes resultados são comentados de maneira a avaliar tanto o interesse objetivo, quanto o interesse subjetivo dos padrões obtidos.

As exceções relacionadas a cada regra de associação ou fato, são apresentadas no formato apresentado pela Tabela 4.2. As exceções negativas (tabela superior) e as exceções positivas (tabela inferior) são apresentadas separadamente.

As exceções negativas são apresentadas ordenadas decrescentemente, de acordo com valor do índice GI^-. A primeira coluna (**Id**) é utilizada para identificar a exceção e também para apresentar a sua ordem de acordo com o valor para a medida GI^-. Por exemplo, se a exceção é a que tem o décimo maior valor de GI^-, então o campo (**Id**) mostra o valor EN_{10}. A segunda coluna apresenta o conjunto de prova Z, relacionado à exceção. A terceira e a quarta colunas exibem, respectivamente, os valores dos índices MF^- e GI^-. A quinta coluna apresenta o valor do índice de dependência negativa entre o antecedente da regra avaliada (A) e o conjunto de prova (Z). A sexta coluna apresenta o valor do índice de dependência negativa entre o conseqüente da regra avaliada (B) e o conjunto de prova (Z). Por

Exceções Negativas						
Id	**Conjunto de Prova**	$\mathbf{MF^-}$	$\mathbf{GI^-}$	$\mathbf{DN_{A,Z}}$	$\mathbf{DN_{B,Z}}$	$\mathbf{DC^-}$
EN_1	Z_1	x.xxxx	x.xxxx	x.xxxx	x.xxxx	x.xxxx
EN_2	Z_2	x.xxxx	x.xxxx	x.xxxx	x.xxxx	x.xxxx

Exceções Positivas						
Id	**Conjunto de Prova**	$\mathbf{MF^+}$	$\mathbf{GI^+}$	$\mathbf{DP_{A,Z}}$	$\mathbf{DP_{B,Z}}$	$\mathbf{DC^+}$
EP_1	Z_1	x.xxxx	x.xxxx	x.xxxx	x.xxxx	x.xxxx
EP_2	Z_2	x.xxxx	x.xxxx	x.xxxx	x.xxxx	x.xxxx

Tabela 4.2: Formato de apresentação dos resultados.

fim, a sétima coluna apresenta o valor do índice DC^- para a exceção.

As exceções positivas são apresentadas de maneira semelhante. A diferença está apenas no fato de que os índices MF^-, GI^-, $DN(A,Z)$, $DN(B,Z)$ e DC^- são substituídos, respectivamente, pelos índices MF^+, GI^+, $DP(A,Z)$, $DP(B,Z)$ e DC^+. As exceções positivas são apresentadas em ordem decrescente, de acordo com o valor do índice GI^+.

É importante notar também que as exceções negativas são sempre identificadas (campo **Id**) pelo prefixo EN, enquanto as exceções positivas são sempre identificadas pelo prefixo EP.

Não há nenhum parâmetro de entrada para configurar um valor mínimo para as medidas DC^- e DC^+ (desvios de confiança). Os valores destas medidas foram calculados pelo programa EXCEPMINER e são apresentados juntos com os resultados, com o objetivo de enriquecer o conhecimento a respeito das exceções mineradas.

4.3.1 Base de Dados da AIDS

Nesta subseção, exceções mineradas a partir da base de dados da AIDS são apresentadas. A regra avaliada é *(Transmissão Sexual = "Não")* $\Rightarrow$ *(Drogas = "Sim")*,

que possui os seguintes valores para suas medidas de interesse: $Sup = 0,1305$, $Conf = 0,8809$ e $lift = 2,944$. Esta regra, já apresentada no Capítulo 1, indica que a grande maioria das pessoas que não contraíram AIDS através de relações sexuais, são usuárias de drogas. O conjunto $\mathcal{A} = \{$Sexo, Idade, Região$\}$ contém os atributos escolhidos para compor os conjuntos de prova. Além disso, os seguintes parâmetros foram utilizados: $SupMin = 0,01\%$, $MF_{min} = 0,40$, $GI_{min} = 0,01$ e $TMax = 3$. O valor baixo para GI_{min} foi utilizado com o intuito de possibilitar a comparação entre exceções que possuem valores mais altos para o grau de inesperabilidade com exceções que possuem valores baixos. Como resultado, foram mineradas 41 exceções negativas e 5 exceções positivas. A Tabela 4.3, apresenta alguns dos resultados obtidos.

Exceções Negativas						
Id	**Conjunto de Prova**	$\mathbf{MF^-}$	$\mathbf{GI^-}$	$\mathbf{DN_{A,Z}}$	$\mathbf{DN_{B,Z}}$	$\mathbf{DC^-}$
EN_1	(Sexo = "F"), (Região = "Norte")	0,9428	0,1260	0,8168	0,8054	0,6876
EN_3	(Idade = "F"), (Região = "Nordeste"), (Idade = "25-29")	0,7934	0,1169	0,6616	0,6765	0,3895
EN_{40}	(Idade = "$\geq$ 50")	0,8153	0,0130	0,5548	0,8024	0,5854

Exceções Positivas						
Id	**Conjunto de Prova**	$\mathbf{MF^+}$	$\mathbf{GI^+}$	$\mathbf{DP_{A,Z}}$	$\mathbf{DP_{B,Z}}$	$\mathbf{DC^+}$
EP_1	(Sexo = "M"), (Região = "Sudeste") (Idade = "20-24")	0,4238	0,0374	0,3864	0,3847	0,0610

Tabela 4.3: Exceções relativas à regra *(Transmissão Sexual = "Não")* $\Rightarrow$ *(Drogas = "Sim")*, da base de dados da AIDS.

A exceção negativa EN_1 foi a que apresentou os maiores valores para os índices MF^-, GI^- e DC^-. Esta exceção pôde revelar uma situação inesperada: a grande maioria das mulheres da região norte que não contraíram AIDS através

de relações sexuais, não são usuárias de drogas. De fato, após a descoberta desta exceção, foi realizada uma consulta SQL à base de dados da AIDS e desta maneira foi possível descobrir que existem 18 mulheres residentes na região norte, que não contraíram AIDS através de relações sexuais. Deste total de 18 mulheres, 13 contraíram o vírus HIV através de transfusão sanguínea. A Figura 4.4, ilustra o resultado da consulta à base de dados da AIDS.

Microsoft Access - [F_NORTE_DROGAS : Consulta seleção]

Arquivo Editar Exibir Inserir Formatar Registros Ferramentas Janela Ajuda

SEXO	REGIAO	UF	IDADE	TRANSMISSAO_SEXUAL	TRANSFUSAO	DROGAS
F	NORTE	PA	35-39 ANOS	NAO	SIM	NAO
F	NORTE	RR	25-29 ANOS	NAO	SIM	NAO
F	NORTE	PA	ACIMA DE 50 ANOS	NAO	SIM	NAO
F	NORTE	PA	40-49 ANOS	NAO	NAO	SIM
F	NORTE	PA	40-49 ANOS	NAO	SIM	NAO
F	NORTE	PA	35-39 ANOS	NAO	SIM	NAO
F	NORTE	RO	20-24 ANOS	NAO	SIM	NAO
F	NORTE	RO	30-34 ANOS	NAO	SIM	NAO
F	NORTE	AM	ACIMA DE 50 ANOS	NAO	SIM	NAO
F	NORTE	AP	25-29 ANOS	NAO	NAO	SIM
F	NORTE	RO	25-29 ANOS	NAO	SIM	NAO
F	NORTE	RO	20-24 ANOS	NAO	SIM	NAO
F	NORTE	PA	25-29 ANOS	NAO	SIM	NAO
F	NORTE	PA	40-49 ANOS	NAO	SIM	NAO
F	NORTE	TO	35-39 ANOS	NAO	NAO	SIM
F	NORTE	PA	35-39 ANOS	NAO	SIM	NAO
F	NORTE	PA	30-34 ANOS	NAO	NAO	SIM
F	NORTE	PA	20-24 ANOS	NAO	NAO	SIM

Figura 4.4: Resultado da consulta à base de dados da AIDS.

Este é o exemplo típico de informação que muito dificilmente poderia ser descoberta por um algoritmo de mineração de regras de associação convencional, pois o suporte do conjunto de condições {(Transmissão Sexual = "Não"),(Sexo = "F"),(Região = "Norte")} é extremamente baixo (0,01%). No entanto a técnica para mineração de exceções foi capaz de destacar a exceção EN_1 como a mais interessante no teste realizado. Neste exemplo, também é possível notar que a dependência negativa entre A e Z ($DN_{A,Z} = 0,8168$) e entre B e Z ($DN_{B,Z} = 0,8054$) é muito alta; no entanto, a dependência negativa entre Z e a regra $A \Rightarrow B$ é ainda maior ($MF^{-} = 0,9428$).

Ainda que neste teste nenhuma das exceções negativas tenha apresentado um valor muito alto para medida GI^-, foi possível notar, através de uma análise subjetiva, que os maiores valores para esta medida foram capazes de evidenciar as exceções mais interessantes, como por exemplo EN_1 e EN_3. Por outro lado, baixos valores para GI^- apontaram para padrões negativos mais óbvios. A exceção negativa EN_{40} é um exemplo. Esta exceção foi gerada apenas pelo fato de que pessoas com idade acima de 50 anos raramente são usuárias de drogas. A forte dependência negativa entre B e Z foi responsável por reduzir o suporte da regra ($MF^- = 0,8153$ e $DN_{A,Z} = 0,8024$).

O suporte e a confiança da regra *(Transmissão Sexual = "Não")* $\Rightarrow$ *(Drogas = "Sim")* são, originalmente, muito altos. Por este motivo nenhuma exceção positiva forte em relação a esta regra pôde ser minerada. Ainda assim, a exceção positiva EP_1 é útil para indicar que a regra torna-se um pouco mais forte entre os homens, residentes na região sudeste, com idade entre 20 e 24 anos. A confiança aumenta cerca de 6,10% em relação ao esperado (observe a coluna DC^+) e o suporte aumenta em mais de 40% em relação à expectativa (coluna MF^+).

4.3.2 Base de Dados dos Cogumelos

Nesta subseção, algumas exceções mineradas a partir da base de dados dos cogumelos são apresentadas. A regra avaliada é *(Habitat = "Grasses")* $\Rightarrow$ *(Class = "Edible")*, que possui os seguintes valores para suas medidas de interesse: $Sup = 0,1733$, $Conf = 0,6555$ e $lift = 1,265$. Esta regra indica que a maior parte dos cogumelos encontrados na grama são comestíveis. A Tabela 4.4, apresenta algumas exceções relacionadas a esta regra. Quase todos os outros atributos diferentes de $Habitat$ e $Class$ foram selecionados para formar os conjuntos de prova. Apenas os atributos $VeilType$ e $StalkRoot$ não foram utilizados (o primeiro por conter muitos valores nulos e o segundo por ser um atributo inútil, já que possui apenas um valor de domínio). Os seguintes parâmetros foram utilizados: $SupMin = 0,20\%$, $MF_{min} = 0,40$, $GI_{min} = 0,10$ e $TMax = 3$. Como resultado, foram mineradas 1.574 exceções negativas e 6.618 exceções positivas. A Tabela 4.4 apresenta qua-

tro resultados selecionados para serem comentados (duas exceções negativas e duas exceções positivas).

Exceções Negativas						
Id	**Conjunto de Prova**	$\mathbf{MF^-}$	$\mathbf{GI^-}$	$\mathbf{DN_{A,Z}}$	$\mathbf{DN_{B,Z}}$	$\mathbf{DC^-}$
EN_4	(CapShape = "Flat"), (StalkShape = "Enlarging"), (StalkSurfaceBelowRing = "Smooth")	1,0000	0,9286	0,0714	0,0289	1,0000
EN_{16}	(GillColor = "White"), (StalkSurfaceBelowRing = "Ibrous")	1,0000	0,8275	-0,0806	0,1725	1,0000

Exceções Positivas						
Id	**Conjunto de Prova**	$\mathbf{MF^+}$	$\mathbf{GI^+}$	$\mathbf{DP_{A,Z}}$	$\mathbf{DP_{B,Z}}$	$\mathbf{DC^+}$
EP_1	(CapColor = "Brown"), (StalkShape = "Tapering"), (StalkColorAboveRing = "White")	0,4042	0,3131	0,0911	-0,0174	0,3445
EP_9	(Bruises = "False"), (RingType = "Evanescent")	0,4169	0,3064	0,1104	-0,6402	0,3445

Tabela 4.4: Exceções relativas à regra *(Habitat = "Grasses")* $\Rightarrow$ *(Class = "Edible")* da base de dados dos cogumelos.

A exceção negativa EN_4 indica que, em geral, cogumelos encontrados na grama são comestíveis, a menos que possuam os seguintes caracteres botânicos:

1. O formato do chapéu não é curvo e este possui pequena espessura. Este caracter botânico é indicado pela condição (CapShape = "Flat").
2. A haste é mais larga na parte próxima à base (StalkShape = "Enlarging").
3. A superfície da haste é lisa, na parte situada abaixo do anel (StalkSurfaceBelowRing = "Smooth").

Este é um exemplo muito interessante de exceção. Os valores baixos para as medidas $DN_{A,Z}$ e $DN_{B,Z}$ indicam que o conjunto de prova é quase que totalmente independente de A e de B. No entanto o valor da medida MF^- é máximo. Isto significa que Z não reduz o suporte de A e também não reduz o suporte de B; no entanto Z reduz o suporte da regra $A \Rightarrow B$ a zero.

A exceção negativa EN_{16} indica que, em geral, cogumelos encontrados na grama são comestíveis, a menos que possuam os seguintes caracteres botânicos:

1. As lamelas são da cor branca (GillColor = "White").
2. A superfície da haste não é lisa, na parte situada abaixo do anel (StalkSurfaceBelowRing = "Ibrous").

Esta exceção apresenta outra curiosidade: Z e A possuem uma leve dependência positiva ($DN_{A,Z} = -0,0806$); porém outra vez o valor de MF^- é máximo.

Exceções positivas interessantes também foram obtidas neste teste. A exceção EP_1 indica que há uma chance maior de os cogumelos encontrados na grama serem comestíveis quando apresentam os seguintes caracteres botânicos:

1. Cor do chapéu marrom (CapColor = "Brown").
2. Haste mais estreita na parte próxima à base (StalkShape = "Tapering").
3. Haste com a cor branca, na parte acima do anel (StalkColorAboveRing = "White").

Esta exceção apresenta um caso semelhante ao da exceção EN_4. Existe uma leve dependência positiva entre A e Z; B e Z são quase totalmente independentes; porém Z aumenta em muito a probabilidade de ocorrência de $A \Rightarrow B$. Além disso, o valor $DC^+ = 0,3445$ significa que a confiança da exceção candidata $A \Rightarrow B\ [Z]$ é igual a 100%. Segundo a equação (Definição 3.5.3), tem-se que: $DC^+ = 1 - (ConfEsp(A \Rightarrow B\ [Z]) \div Conf(A \Rightarrow B\ [Z]))$. Dado que a confiança da regra original *(Habitat = "Grasses")* $\Rightarrow$ *(Class = "Edible")* é igual a 0,6555, então o seguinte

cálculo pode ser efetuado: $0,3445 = 1 - (0,6555 \div Conf(A \Rightarrow B\ [Z])) \equiv Conf(A \Rightarrow B\ [Z]) = 1,0000$.

A exceção EP_9 indica que há uma chance maior de os cogumelos encontrados na grama serem comestíveis quando apresentam os seguintes caracteres botânicos

1. Nenhuma marca ou machucado (Bruises = "False").
2. Anel com formato descontínuo (RingType = "Evanescent").

Este também é um caso interessante, pois existe uma forte dependência negativa entre Z e B ($DN_{A,Z} = -0,6402$), porém mais uma vez a confiança da exceção candidata $A \Rightarrow B\ [Z]$ é igual a 100% ($DC^{+} = 0,3445$). O suporte da exceção candidata também ficou muito acima do esperado ($MF^{+} = 0,4169$).

4.3.3 Base de Dados da Aterosclerose

Nos testes realizados com a base de dados da aterosclerose, o tamanho máximo dos conjuntos de prova (parâmetro $TMax$) foi limitado a um elemento. Os outros parâmetros utilizados foram $SupMin = 0,80\%$, $MF_{min} = 0,30$, $GI_{min} = 0,10$. O conjunto de atributos selecionados para geração dos conjuntos de prova é dado por: $\mathcal{A}$ = *{Education, Age, BasicGroup, PhysicalActivityAfterJob, DailyBeerConsumption, DailyWineConsumption, DailyLiquorConsumption, BMI, Cholesterol, Tryglicerides, BloodPressure}*. Alguns destes atributos serão descritos adiante, juntamente com os comentários sobre a exceções mineradas.

O consumo excessivo de cigarros é apontado como um dos maiores causadores de doenças cardiovasculares. Por isto, duas regras de associação envolvendo pacientes fumantes foram selecionadas para avaliação. A primeira regra é *(DailyBeerConsumption = ">1l")* $\Rightarrow$ *(Smoking = ">20 cig/day")* ($Sup = 4,48\%$, $Conf = 37,58\%$ e $lift = 1,444$). Esta regra indica que os pacientes que bebem muita cerveja (o equivalente a mais de um litro de cerveja por dia) têm maior chance de também fumar muito (mais de 20 cigarros por dia). Duas exceções negativas a esta regra

são ilustradas na Tabela 4.5. Nenhuma exceção positiva a esta regra foi encontrada, considerando os parâmetros de entrada adotados.

Exceções Negativas						
Id	**Conjunto de Prova**	**MF$^-$**	**GI$^-$**	**DN$_{A,Z}$**	**DN$_{B,Z}$**	**DC$^-$**
EN_1	(Age = "≥50")	0,5304	0,2658	0,2649	-0,0113	0,3614
EN_2	(BasicGroup = "A")	0,9192	0,1837	0,4229	0,7354	0,8600

Tabela 4.5: Exceções relativas à regra *(DailyBeerConsumption = ">1l")* ⇒ *(Smoking = ">20 cig/day")* da base de dados da aterosclerose.

As pessoas que bebem muito tendem a também fumar muito, mas isto torna-se muito mais fraco entre os pacientes que têm mais de 50 anos (EN_1). O suporte da regra é 53,04% inferior ao esperado (coluna MF^-) e a confiança apresenta um valor 36,14% inferior ao esperado (coluna DC^-).

A regra original também torna-se muito mais fraca entre os pacientes que pertencem ao grupo A (os pacientes saudáveis, que não apresentam nenhuma doença cardiovascular e nenhum fator de risco de aterosclerose). Neste caso, representado pela exceção EN_2, note os valores extremamente altos para as medidas MF^- e DC^-. Já o valor do grau de inesperabilidade não foi tão alto se comparado com o valor do grau de inesperabilidade de EN_1 (veja a coluna GI^-). Consultando a base de dados da aterosclerose, foi possível descobrir que um total de 19 pacientes pertencem ao grupo A e costumavam fumar mais de 20 cigarros por dia (suporte de 1,52%), mas todos estes pacientes já pararam de fumar há mais de um ano (um campo denominado $ExSmoker$ indica esta situação). A consulta indicou que apenas um dentre estes 19 pacientes costumava beber muita cerveja diariamente.

A segunda regra avaliada é *(Education = "apprentice school")* ⇒ *(Smoking = "15-20 cig/day")* ($Sup = 9,84\%$, $Conf = 34,35\%$ e $lift = 1,189$). Esta regra indica que os pacientes cujo grau máximo de instrução é equivalente a *"apprentice school"* têm maior chance de fumar muito (entre 15 e 20 cigarros por dia). Na República Tcheca, *"apprentice school"* corresponde a um grau de instrução não muito elevado, sendo inferior ao segundo grau. Duas exceções negativas a esta regra são ilustradas

na Tabela 4.6. Nenhuma exceção positiva a esta regra foi encontrada, considerando os parâmetros de entrada adotados.

Exceções Negativas						
Id	**Conjunto de Prova**	$\mathbf{MF^-}$	$\mathbf{GI^-}$	$\mathbf{DN_{A,Z}}$	$\mathbf{DN_{B,Z}}$	$\mathbf{DC^-}$
EN_1	(PhysicalActivityAfterJob = "great activity")	0,4306	0,1479	0,2827	0,1270	0,2062
EN_2	(DailyBeerConsumption = "does not drink beer")	0,4729	0,1248	0,3481	0,0901	0,1915

Tabela 4.6: Exceções relativas à regra *(Education = "apprentice school")* $\Rightarrow$ *(Smoking = "15-20 cig/day")* da base de dados da aterosclerose.

Os resultados demonstram que a associação entre possuir um baixo nível de escolaridade e fumar muito torna-se mais fraca entre as pessoas que praticam atividade física de forma intensa em seu tempo livre (EN_1) e também entre os pacientes que não consomem cerveja (EN_2).

Os resultados desta subseção encontram-se também em [16], assim como uma série de regras de associação fortes mineradas da base de dados da aterosclerose.

4.3.4 Base de Dados do Censo de Washington

Na base de dados do censo, o atributo alvo $IncomeClass$ é utilizado para especificar se uma pessoa possui renda anual superior a US$ 50.000,00. Esta condição pode ser representada pelo fato $F = (IncomeClass = \text{“} > 50K\text{”})$, que possui suporte igual a 24,08%. Este fato não é forte, já que existe uma fração bem maior de pessoas que recebem salário anual menor ou igual a US$ 50.000,00 (75,92%).

Um teste foi realizado para descobrir exceções positivas relacionadas ao fato $F = (IncomeClass = \text{“} > 50K\text{”})$. Os parâmetros utilizados foram $SupMin = 0,10\%$, $MF_{min} = 0,30$ e $TMax = 1$. Lembre-se que para a mineração de exceções a fatos, não existe a necessidade de configurar um valor mínimo para o grau de inesperabilidade. O conjunto de atributos $\mathcal{A}$ = *{Education, Gender, HoursPerWeek,*

Age} foi utilizado na geração dos conjuntos de prova. O objetivo foi descobrir qual dentre estes atributos aumenta mais a chance de uma pessoa ganhar bem. A Tabela 4.7 apresenta algumas das 20 exceções positivas mineradas.

Exceções Positivas		
Id	**Conjunto de Prova**	**MF$^+$**
EP_1	(Education = "Doctorate")	0,6702
EP_2	(Education = "Prof-school")	0,6692
EP_3	(Education = "Masters")	0,5682
EP_4	(HoursPerWeek = "56-60")	0,4474
EP_5	(HoursPerWeek = "51-55")	0,4428
EP_9	(Age = "46-50")	0,3961
EP_{10}	(Age = "51-55")	0,3934
EP_{17}	(Gender = "Male")	0,3308

Tabela 4.7: Exceções relativas ao fato *(IncomeClass = ">50K")* da base de dados do censo de Washington.

Os resultados obtidos indicam que um elevado grau de educação é o fator que mais aumenta a chance de uma pessoa ganhar mais de US$ 50.000,00 (EP_1, EP_2 e EP_3). O segundo fator que mais contribui é o número de horas trabalhadas por semana (atributo $HoursPeerWeek$). Como era de se esperar, a probabilidade de ocorrência do fato $F = (IncomeClass = \text{“} > 50K\text{”})$, aumenta bastante entre as pessoas que trabalham mais 50 horas por semana (EP_5 e EP_6). Em seguida, tem-se o atributo idade. As pessoas com mais de 45 anos têm maior chance de ganhar bem (EP_9 e EP_{10}), muito provavelmente porque, devido à longa experiência de trabalho, estas pessoas podem já estar ocupando cargos de gerência ou chefia. Por fim, a exceção EP_{17} evidencia uma dependência positiva entre o entrevistado ser do sexo masculino e ganhar bem. Desta forma, é possível inferir que a condição (Gender = "Female") e o fato $F = (IncomeClass = \text{“} > 50K\text{”})$ possuem dependência negativa.

Capítulo 5

Conclusões e Trabalhos Futuros

A contribuição central desta dissertação foi a definição de um método cooperativo para a mineração de exceções negativas e positivas em bases de dados multidimensionais. Nesta abordagem, usuários podem explorar um conjunto de regras de associação, para descobrir o quanto a força de cada uma destas regras desvia-se de seus valores médios em diferentes subconjuntos selecionados da base de dados. As exceções são mineradas quando estes desvios apresentam valores significativos. O objetivo é descobrir informações previamente ocultas, úteis e inesperadas.

Neste trabalho, foram revisadas as definições de regras de associação transacionais e de regras de associação multidimensionais, assim como as suas principais medidas de interesse objetivas e subjetivas. Também foi revisado o conceito de regra de associação negativa, que é bastante relacionado ao conceito de exceção. O problema da mineração de regras de associação negativas em bases transacionais foi introduzido em [27]. Segundo esta proposta, uma regra de associação negativa é extraída da base de dados quando não alcança um suporte esperado. Este suporte esperado é computado baseado na existência de uma taxonomia que classifica hierarquicamente os itens que pertencem ao domínio de conhecimento da aplicação. Em [21], apresentou-se um novo tipo de regra negativa, denominado regra de associação relacional negativa. Estas regras podem ser mineradas a partir de bancos de dados relacionais e são extraídas se possuírem o suporte ou a confiança inferior a uma determinada expectativa.

A técnica para mineração de exceções negativas e positivas, apresentada nesta dissertação, é uma extensão da proposta apresentada em [21]. Uma exceção representa uma regra cuja força afastou-se significativamente de seus valores médios num determinado subconjunto da base de dados. As exceções negativas representam regras que tornam-se fracas, enquanto as exceções positivas representam regras que tornam-se mais fortes nestas fatias da base de dados. Foram propostas medidas de interesse capazes de avaliar as exceções negativas e positivas e propôs-se um procedimento para a mineração destas exceções. Este procedimento foi implementado e avaliado através da mineração de bases de dados reais de diferentes áreas.

Como trabalhos futuros, primeiro pretende-se adaptar a técnica de mineração de exceções para também ser utilizada em bases de dados transacionais. Neste caso, a força de uma regra de associação transacional será avaliada dentro de diferentes conjuntos de prova, compostos por itens especificados pelo usuário. Intenciona-se ainda integrar o procedimento para mineração de exceções junto à ferramenta de mineração de dados MIDAS-UFF, que está sendo desenvolvida no Instituto de Computação da Universidade Federal Fluminense.

Referências Bibliográficas

[1] R. Agrawal, T. Imielinski e R. Srikant, "Mining Association Rules between Sets of Items in Large Databases", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington, Estados Unidos, 1993, 207–216.

[2] R. Agrawal e R. Srikant, "Fast Algorithms for Mining Association Rules", *Proceedings of the* 20^{th} *International Conference on Very Large DataBases Conference*, Santiago, Chile, 1994, 487–499.

[3] A. Agresti, *Categorical Data Analysis*, J. Wiley Publishing, 1990.

[4] M. L. A. Berry e G. Linoff, *Data Mining Techniques: for Marketing, Sales and Customer Support*, J. Wiley Computer Publishing, 1997.

[5] C. L. Blake e C. J. Merz, "UCI Repository of Machine Learning Databases", Informações Disponíveis em [*http://www.ics.uci.edu/˜mlearn/MLRepository.html*], Department of Information and Computer Science, University of California, Irvine, 1998.

[6] S. Brin, R. Motwani e C. Silverstein, "Beyond Market Baskets: Generalizing Association Rules to Correlations", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Arizona, Estados Unidos, 1997, 265–276.

[7] S. Brin, R. Motwani, J. D. Ulman e S. Tsur, "Dynamic Itemset Counting and Implication Rules for Market Basket Data", *Proceedings of the ACM SIGMOD*

International Conference on Management of Data, Arizona, Estados Unidos, 1997, 255–264.

[8] Homepage da ECML/PKDD 2004 Discovery Challenge. Informações Disponíveis em [*http://lisp.vse.cz/challenge/ecmlpkdd2004/*], 2004.

[9] R. Elmasri e S. B. Navathe, *Fundamentals of Database Systems*, Addison-Wesley, 4^a Edição, 2003.

[10] European Centre of Medical Informatics, Statistics and Epidemiology of Charles Unversity and Charles University Hospital. Informações Disponíveis em [*http://euromise.vse.cz/STULONG*], 2004.

[11] U. M. Fayyad, G. Piatetsky-Shapiro e P. Smith, "From Data Mining to Knowledge Discovery: An Overview", *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996, 1–34.

[12] Fundação Getulio Vargas, Instituto Brasileiro de Economia, Divisão de Gestão de Dados (FGV/IBRE/DGD). Informações Disponíveis em [*http://www.fgv.br/ibre/dgd*], 2004.

[13] V. Ganti, J. Gehrke e R. Ramakrishnan, "Mining Very Large Databases", *IEEE Computer Vol. 32, No. 8*, 1999, 38–45.

[14] M. Goebel e L. Gruenwald, "A Survey of Data Mining and Knowledge Discovery Tools", *SIGKDD Explorations Vol. 1*, 1999, 20–33.

[15] E. C. Gonçalves, I. M. B. Mendes e A. Plastino, "Mining Exceptions in Databases", *Proceedings of the* 17^{th} *Australian Joint Conference on Artificial Intelligence*, Cairns, Australia, 2004 (a ser publicado).

[16] E. C. Gonçalves e A. Plastino, "Mining Strong Associations and Exceptions in the STULONG Data Set", *Proceedings of the ECML/PKDD 2004 Discovery Challenge*, Pisa, Itália, 2004, 44–55.

[17] J. Han e M. Kamber, *Data Mining: Concepts and Techiniques*, Morgan Kaufmann Publishers, 2001.

[18] M. Kamber, J. Han, J. e J. Y Chiang, "Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes", *Proceedings of the* 3^{rd} *SIGKDD International Conference on Knowledge Discovery and Data Mining*, Califórnia, Estados Unidos, 1997, 207–210.

[19] S. Lipschutz, *Probabilidade*, Makron Books, 4^a Edição, 1994.

[20] B. Liu e W. Hsu, "Post-Analysis of Learned Rules", *Proceedings of the* 13^{th} *National Conference on Artificial Intelligence*, Portland, Estados Unidos, 1996, 828–834.

[21] I. M. B. Mendes, "Regras de Associação Negativas", Dissertação de Mestrado apresentada ao Curso de Pós-Graduação em Computação Aplicada e Automação da Universidade Federal Fluminense, 2002.

[22] B. Padmanabhan e A. Tuzhilin, "Unexpectedness as a Measure of Interestingness in Knwoledge Discovery" *Decision Support Systems Vol. 27*, 1999, 303–318.

[23] G. Piatetsky-Shapiro, "Discovery, Analysis and Presentation of Strong Rules" *Knowledge Discovery in Databases*, AAAI/MIT Press, 1991, 229–248.

[24] Programa Brasileiro para Doenças Sexualmente Transmissíveis e AIDS. Informações Disponíveis em [*http://www.aids.gov.br*], 2004.

[25] M. Radelet, "Racial Characteristics and the Imposition of the Death Penalty" *American Sociological Review Vol. 46*, 1981, 918–927.

[26] P. Raven, R. Evert e S. Eichhorn, *Biologia Vegetal*, Ed. Guanabara Koogan, 1992.

[27] A. Savasere, E. Omiecinski e S. Navathe, "Mining for Strong Negative Associations in a Large Database of Costumer Transactions", *Proceedings of the* 14^{th} *International Conference on Data Engineering*, Flórida, Estados Unidos, 1998, 494–502.

[28] A. Silberschatz e A. Tuzhilin, "On Subjective Measures of Interestingness in Knowledge Discovery", *Proceedings of the* 1^{st} *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Montreal, Canadá, 1995, 275–281.

[29] Sociedade Brasileira de Cardiologia. "Resumo das III Diretrizes Brasileiras sobre Dislipidemias e I Diretriz de Prevenção de Aterosclerose do Departamento de Aterosclerose da Sociedade Brasileira de Cardiologia", 2004.

[30] Sociedade Brasileira de Cirurgia Vascular. Informações Disponíveis em [*http://www.sbcv.org.br*], 2004.

[31] R. Srikant e R. Agrawal, "Mining Quantitative Association Rules in Large Relational Tables", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Montreal, Canadá, 1996, 1–12.

[32] E. Suzuki e J. M. Zŷtkow, "Unified Algorithm for Undirected Discovery of Exception Rules", *Proceedings of the 4^{th} European Conference on Principles of Data Mining and Knowledge Discovery*, Lyon, França, 2000, 169–180.

[33] P. Tan, V. Kumar e J. Srivastava, "Selecting the Right Interestingness Measure for Association Patterns ", *Proceedings of the 8^{th} ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Edmonton, Canadá, 2002, 32–41.

[34] C. N. Targa, "Mineração Eficiente de Regras de Associação através da Indexação de Conjuntos Candidatos", Dissertação de Mestrado apresentada ao Curso de Pós-Graduação em Computação da Universidade Federal Fluminense, 2002.

[35] K. Wang, Y. Jiang e L. V. S. Lakshmanan, "Mining Unexpected Rules by Pushing User Dynamics.", *Proceedings of the 9^{th} ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, Estados Unidos, 2003, 246–255.

[36] Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Informações Disponíveis em [*http://www.cs.waikato.ac.nz/ml/weka/*], 2004.

[37] I. H. Witten e E. Frank, *Data Mining: Pratical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, 2000.

[38] X. Wu, C. Zhang e S. Zhang, "Mining both Positive and Negative Association Rules", *Proceedings of the 19^{th} International Conference on Machine Learning*, Sydney, Austrália, 2002, 658–665.

[39] Z. Zheng, R. Kohavi e L. Mason, "Real World Performance of Association Rule Algorithms", *Proceedings of the 7^{th} ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, São Francisco, Estados Unidos, 2001, 401–406.