Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas

EDUARDO CORRÊA GONÇALVES

UFF - Universidade Federal Fluminense
IC - Instituto de Computação
Rua Passo da Pátria, 156 - Bloco E - 3^o andar - Boa Viagem - CEP 24210-240 Niterói (RJ)
egoncalves@ieee.org

Resumo. As regras de associação descrevem padrões de relacionamento entre itens de uma base de dados. Uma de suas típicas aplicações é a análise de transações de compras. Este processo examina padrões de compras de consumidores para determinar produtos que costumam ser adquiridos em conjunto. Um exemplo de regra de associação, que poderia ser minerada a partir de uma base de dados de um supermercado, é dado por: $\{salaminho\} \Rightarrow \{cerveja\}$. Esta regra indica que os clientes que compram $\{salaminho\}$, tendem a também comprar $\{cerveja\}$. Medidas de interesse são comumente utilizadas para avaliar a qualidade de uma regra de associação. Estas medidas podem ser divididas em duas classes: objetivas (identificam estatisticamente a força de uma regra) e subjetivas (a opinião de um analista é levada em consideração para determinar a força da regra). Neste trabalho algumas das principais medidas de interesse encontradas na literatura são comparadas e avaliadas. Com o objetivo de facilitar a discussão, serão apresentados os resultados de uma experiência realizada sobre uma base de dados real, que registra as compras efetuadas por famílias cariocas em supermercados.

Palavras-Chave: Medidas de Interesse, Regras de Associação, Mineração de Dados, Descoberta de Conhecimento.

Objective and Subjective Measures for Association Rules

Abstract. Association rules describe relationship patterns among items in a database. They are commonly used for the purpose of market basket analysis. This process examines purchasing trends, determining what products customers are likely to buy together. An example of association rule, that could be mined from a supermarket database, is given by $\{salami\} \Rightarrow \{beer\}$. This rule indicates that costumers who buy $\{salami\}$ tend to also buy $\{beer\}$. Interest measures are often employed to evaluate the quality of an association rule. These measures can be of two types: objective measures (identify statistically the strength of a rule) and subjective measures (take into account the opinion of analysts). This work compares and discusses some of the most important interest measures found in the literature. In order to facilitate this discussion, we present the results of an experience performed over a real data set that keeps information about purchases made by families residing in Rio de Janeiro.

Keywords: Interest Measures, Association Rules, Data Mining, Knowledge Discovery in Databases.

(Recebido para publicação em 20 de fevereiro de 2005 e aprovado em 29 de março de 2005)

1 Introdução

Nos dias atuais, 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 de 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. Por esta razão, métodos tradicionais como investigação manual, consultas SQL e planilhas de cálculo tornam-se inviáveis [5]. Para lidar com este problema, 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, hierarquias 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 [3, 5, 12].

1.1 Regras de Associação

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. Uma de suas típicas aplicações é a análise de transações de compra (*market basket analysis*). A partir de uma base de dados que armazena 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 exemplo: {*feijão*} $\wedge$ {*couve*} $\Rightarrow$ {*lingüiça*}. Esta regra é utilizada para indicar que os clientes que compram os produtos feijão e couve, tendem a também comprar lingüiça. 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 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.

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 (*itemset*), 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. Tanto o antecedente, quanto o conseqüente de uma regra de associação podem ser formados por conjuntos contendo um ou mais itens. 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*.

Tabela 1: Base de dados de transações

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

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, ou seja, $Conf(A \Rightarrow B) = Sup(A \cup B) \div Sup(A)$.

1.2 Modelo Suporte / Confiança

O modelo típico para mineração de regras de associação em bases de dados 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 1. Suponha suporte e confiança mínimos iguais a 30% e 65%, respectivamente. Uma estratégia de mineração de

regras de associação que fosse baseada no Modelo Suporte/Confiança, geraria como resultado as quatro regras apresentadas na Tabela 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.

Tabela 2: Exemplo dos índices suporte e confiança.

Regra	$\mathbf{Sup_A}$	$\mathbf{Sup_B}$	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

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.

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 [14] 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 [4, 8].

Para resolver estes problemas, outras medidas de interesse diferentes do suporte e da confiança têm sido propostas com o intuito de identificar as regras que são, de fato, relevantes e úteis dentre as muitas que podem ser mineradas. Existem dois tipos de medidas de interesse: objetivas e subjetivas. As medidas de interesse objetivas empregam índices estatíscos para avaliar a força de uma regra. Já as medidas de interesse subjetivas consideram principalmente a opinião de um analista para determinar a força da regra.

Neste trabalho algumas das principais medidas de interesse encontradas na literatura são comparadas e discutidas. São demonstrados exemplos de utilização destas medidas através de uma experiência realizada sobre uma base de dados de transações real, fornecida pela Fundação Getulio Vargas. O trabalho está dividido da seguinte forma. A Seção 2 demonstra os problemas do Modelo Suporte/Confiança e discorre sobre as medidas de interesse objetivas e subjetivas. Ainda nesta seção, propõe-se um método para a mineração de regras interessantes em bases de dados. Os resultados experimentais obtidos através da aplicação deste método proposto são reportados na Seção 3 Por fim, na Seção 4 são apresentadas as conclusões e idéias para futuros trabalhos.

2 Metodologia

Um método para a mineração de regras interessantes, baseado na utilização e correta interpretação de medidas de interesse objetivas, será proposto nesta seção. O texto está organizado da seguinte forma. A Subseção 2.1 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 subseção, também é apresentado o conceito de dependência entre itens de dados. A Subseção 2.2 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 e propõe-se um método para a mineração de regras interessantes baseado na aplicação das mesmas. Por fim, a Subseção 2.3 aborda a importância das medidas de interesse subjetivas.

2.1 Avaliação do Modelo Suporte/Confiança

Nesta subseção 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 [6]. 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 adqui-

rem 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 associação envolvendo dois itens (ou seja, um item no antecedente e um item no conseqüente). O suporte mínimo foi estabelecido em 3% (13 transações) 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*). Este número confirma um dos principais problemas atribuídos ao Modelo Suporte/Confiança: a saída apresentada para o usuário possui uma quantidade de regras extremamente grande, tornando muito trabalhosa a sua avaliação. A Tabela 3 apresenta algumas das regras de associação obtidas, ordenadas de maneira decrescente de acordo com o valor da medida de confiança.

Tabela 3: Regras de associação mineradas da base de dados da POF.

Regra	$\mathbf{Sup_A}$	$\mathbf{Sup_B}$	Sup	Conf
r_1: cenoura ⇒ batata inglesa	0,7701	0,8175	0,7038	0,9138
r_2: acerola(polpa) ⇒ ovos	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 ⇒ ervilhas	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

A base de dados da POF-RJ-06-98 apresenta uma característica bastante interessante: a distribuição de freqüências dos itens 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 possui suporte baixo, inferior a 10%. Alguns exemplos destes produtos menos populares são ilustrados na Tabela 3: filé de viola, fruta de conde 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 muitas famílias normalmente já iriam adquirir açúcar refinado em suas compras mensais? O Exemplo 1 aborda esta questão.

Exemplo 1 *(Independência entre Itens). Considere a regra r_3 (Tabela 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 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. No entanto, regra {*filé de viola*} ⇒ {*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 3, outro exemplo de regra de associação minerada envolvendo itens independentes é ilustrado pela regra r_2 : {*acerola (polpa)*} ⇒ {*ovos*}. 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 1, dois itens de dados A e B são independentes se:

$$Conf(A \Rightarrow B) = Sup(B)$$

Como a fórmula da confiança é dada por $Conf(A \Rightarrow B) = Sup(A \cup B) \div Sup(A)$, tem-se que A e B são independentes se:

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

Nesta equação, $Sup(A \cup B)$ 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 1 *(Suporte Esperado para 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:*

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

A seguir apresenta-se uma definição formal para o conceito de independência entre itens de dados.

Definição 2 *(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:*

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

O fato de a fórmula da medida de confiança não levar em consideração o valor isolado do suporte do conseqüente das regras de associação traz o problema da mineração de associações envolvendo itens que sejam independentes e, pior ainda, pode levar à mineração de regras envolvendo produtos que possuem dependência negativa. O exemplo a seguir ilustra esta situação.

Exemplo 2 *(Dependência Negativa entre Itens). Considere a regra r_6 (Tabela 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 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 3). Desta vez seria verdadeiro considerar que a compra de milho verde levou os consumidores a também comprar ervilhas? O Exemplo 3 aborda esta questão.

Exemplo 3 *(Dependência Positiva entre Itens). Considere a regra r_4 (Tabela 3). A probabilidade de qualquer cliente comprar* {ervilhas} *é de 37.91%, enquanto a probabilidade de um cliente comprar este produto dado que compra* {milho verde} *sobe para 82.01%. Portanto os clientes que compram* {milho verde} *têm maior probabilidade de comprar* {ervilhas}. *Estes produtos possuem uma **dependência positiva**.*

Ainda na Tabela 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. As regras r_1, r_4 e r_5 representam casos de associações interessantes entre itens da base da POF.

Definição 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 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.

2.2 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 5 *(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:

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

2.2.1 Lift

A medida de interesse *lift* [4], 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 6 *(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:*

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

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 ∞. e possui interpretação bastante simples: quanto maior o valor do *lift*, mais interessante a regra, pois A aumentou ("*lifted*") B numa maior taxa.

Exemplo 4 *(Lift). O valor do índice lift para a regra r_3 :* {açúcar refinado} $\Rightarrow$ {filé de viola} *(Tabela 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} $\Rightarrow$ {banana prata} *(Tabela 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} $\Rightarrow$ {ervilhas} *(Tabela 3) é calculado por:* $0,8201 \div 0,3701 = 2,21$*, indicando que os itens* {milho verde} *e* {ervilhas} *possuem dependência positiva (o suporte real da regra é 2,21 vezes maior do que seu suporte esperado).*

2.2.2 Rule Interest

O índice *Rule Interest* (RI), introduzido em [9] e também conhecido na literatura como PS (letras incidas do nome de seu autor), novidade e *leverage*, é outra medida de interesse objetiva que pode ser utilizada para avaliar dependências. Esta medida indica o valor da diferença entre suporte real e o suporte esperado de uma regra de associação.

Definição 7 *(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:*

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

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. Quanto maior o valor da medida, mais interessante é a regra.

Exemplo 5 *(RI). O valor do índice RI para a regra r_3 :* {açúcar refinado} $\Rightarrow$ {filé de viola} *(Tabela 3) é calculado por:* $0,0758 - (0,0877 \times 0,8649) = 0$*, indicando que os itens {açúcar refinado} e {filé de viola} são independentes.*

O valor do índice RI para a regra r_6 : {banana nanica} $\Rightarrow$ {banana prata} *(Tabela 3) é calculado por:* $0,0735 - (0,1209 \times 0,7607) = -0,06$*, 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} $\Rightarrow$ {ervilhas} *(Tabela 3) é calculado por:* $0,2701 - (0,3294 \times 0,3791) = 0,14$*, indicando que os itens* {milho verde} *e* {ervilhas} *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 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 6 *(RI $\times$ Lift). O valor do índice RI para a regra r_1 :* {cenoura} $\Rightarrow$ {batata inglesa} *(Tabela 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} $\Rightarrow$ {melancia} *(Tabela 3), que apresenta suporte igual a 3,08% 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.

2.2.3 Convicção

Tanto o *lift* quanto o *RI* possuem como característica o fato de serem medidas **simétricas**, ou seja, $Lift(A \Rightarrow B) = Lift(B \Rightarrow A)$ e $RI(A \Rightarrow B) = RI(B \Rightarrow A)$. Isto ocorre porque estes índices possuem o objetivo de mensurar dependência entre os itens, ao invés de medir implicação (o sentido da seta "⇒"). Em [4], a medida de interesse **convicção** é proposta com o objetivo de avaliar uma regra de associação como uma verdadeira implicação.

Definição 8 *(Convicção). 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 da* convicção *para $A \Rightarrow B$ é computado por:*

$$Conv(A \Rightarrow B) = \frac{Sup(A) \times Sup(\neg B)}{Sup(A \cup \neg B)}$$

A medida da convicção foi desenvolvida baseada no seguinte argumento: na lógica proposicional, uma implicação $A \rightarrow B$ pode ser reescrita por $\neg A \vee B \equiv \neg(A \wedge \neg B)$. Seguindo este argumento, $Sup(A \cup \neg B)$, que representa a probabilidade de ocorrência (suporte real) do antecedente sem o conseqüente na base de dados, foi colocado no denominador da fórmula da medida de convicção. Já no numerador da fórmula encontra-se o suporte esperado do antecedente sem o conseqüente. A medida é então capaz de avaliar o quanto A e $\neg B$ se afastam da independência.

A medida da convicção apresenta algumas características bastante interessantes:

- A medida leva em consideração tanto o suporte do antecedente, como o suporte do conseqüente.
- Caso exista a independência completa entre o antecedente e o conseqüente da regra, o valor da convicção será igual a 1.
- Regras onde o antecedente nunca aparece sem o conseqüente (confiança de 100%) terão valor de convicção igual a ∞.

Exemplo 7 *(Convicção). Considere a regra r_4 :* {milho verde} ⇒ {ervilhas} *(Tabela 3). Para calcular o valor da medida da convicção desta regra é necessário obter inicialmente:*

1. $Sup(\neg B) = 1 - Sup(B) = 1 - 0,3791 = 0,6209$.

2. $Sup(A \cup \neg B) = Sup(A) - Sup(A \cup B) = 0,3294 - 0,2701 = 0,0593$.

O valor da convicção é calculado por:

$$(0,3294 \times 0,6209) \div 0,0593 = 3,449.$$

Este valor indica que a probabilidade da compra do produto {milho verde} *ocorrer sem que o produto* {ervilhas} *seja comprado é 3,449 vezes menor do que o esperado.*

A medida da convicção varia entre 0 e ∞. Os autores do índice realizaram uma avaliação através da mineração de uma base de dados censitários. Neste teste, mais de 20.000 regras de associação foram mineradas. Os desenvolvedores da medida identificaram que as regras mais interessantes apresentaram um valor de convicção entre 1,01 e 5. Foi percebido que muitas das regras com valor de convicção acima de 5 representavam informações óbvias ou ilusórias.

Além dos índices $lift$, RI e convicção, comentados nesta subseção, diversos trabalhos encontrados na literatura apresentam outras medidas de interesse objetivas importantes, tais como: **coeficiente de correlação**, **teste chi-squared para independência**, **J-Measure**, entre outras. Uma análise a respeito de cerca de vinte diferentes medidas de interesse objetivas pode ser encontrada em [11].

2.2.4 Método Proposto

Ao contrário dos modelos definidos em [1] (que utiliza apenas o suporte e a confiança para avaliar as regras de associação), em [4] (que propõe a utilização da convicção em substituição à medida de confiança) e em [9] (que utiliza apenas o RI para avaliar dependências), neste trabalho propõe-se um método que utiliza, conjuntamente, os índices suporte, confiança, $lift$, RI e convicção no processo de mineração de regras de associação. A utilização de todas estas medidas permite que usuários possam realizar análises alternativas sobre uma mesma regra, pois cada uma das medidas é capaz de destacar uma característica a respeito da associação.

A abordagem proposta também possui o objetivo de reduzir significativamente a chance de que sejam minerados um número excessivo de regras óbvias ou ilusórias, através da adição do $lift$, do RI e da convicção como parâmetros de entrada para o sistema de mineração de dados. Esta abordagem foi avaliada através de

novo teste realizado sobre a base de dados da POF. Os resultados obtidos são apresentados na Seção 3.

2.3 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 [10], 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 [3]. 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 [10], 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 [8, 13] 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.

Para melhor ilustrar este conceito, será apresentado um exemplo hipotético. Suponha que um analista, acostumado a trabalhar com a base de dados da POF, possua a seguinte expectativa: "as famílias que compram couve costumam comprar brócolis". Um algoritmo desenvolvido para mineração de regras inesperadas a partir da base da POF poderia, por exemplo, encontrar a seguinte regra inesperada (em relação à crença citada): "as famílias que compram couve, mas também compram lingüiça, **não** costumam comprar brócolis".

Uma exceção deste tipo pode ser útil para identificar clientes com diferentes perfis de compra. Um especialista pode concluir, por exemplo, que a regra de associação $\{couve\} \Rightarrow \{brócolis\}$ 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. É importante deixar claro que a mineração de exceções é uma tarefa desafiadora, visto que uma abordagem pouco cuidadosa para a mineração de padrões negativos pode levar à geração de um número grande de regras envolvendo itens muito raros (como por exemplo "clientes que não compram caviar, também não compram cereja"). Para contornar este problema, o método empregado em [7] considera que uma exceção deve ser minerada apenas quando possui valor de suporte significativamente inferior a uma determinada estimativa. Como trabalho futuro, intenciona-se utilizar este princípio no desenvolvimento de um algoritmo para mineração de regras inesperadas a partir da base de dados da POF.

3 Resultados

Com o objetivo de realizar uma avaliação prática das medidas de interesse objetivas suporte, confiança, $lift$, RI e convicção, um novo teste foi realizado sobre a base de dados da POF. O programa original para mineração de regras de associação, apresentado na Seção 2, foi alterado para receber como parâmetros de entrada valores mínimos para os índices $lift$, RI e convicção, além da confiança e do suporte.

A base da POF foi outra vez minerada e os valores mínimos das medidas de interesse foram configurados da seguinte forma: suporte mínimo de 3%, confiança mínima de 60% (como no primeiro teste), $lift$ mínimo de 1,10, RI mínimo de 1% e convicção mínima de 1,10.

Como resultado, o programa apresentou como saída um total de 3.892 regras de associação geradas a partir dos conjuntos freqüentes de tamanho dois. Este número representa menos da metade das regras mineradas no primeiro teste, descrito na Seção 2. A redução do número de regras geradas é explicada pelo fato de a utilização do $lift$ e do RI como parâmetros de entrada evitar a geração de regras envolvendo itens independentes e itens que possuem dependência negativa.

Os resultados demonstraram que a utilização con-

junta das medidas de suporte, confiança, *lift*, RI e convicção permite que usuários possam realizar análises alternativas sobre uma mesma regra, enriquecendo o poder de entendimento a respeito das associações. Para justificar este ponto de vista, considere uma das regras de associação mineradas na avaliação, que é ilustrada na Figura 1.

R: Strogonoff de Frango (caixa) ⇒ Lasanha (caixa)
Suporte(R) = 3,32%.
Confiança(R) = 77,78%
Suporte(Strogonoff de Frango) = 4,27%
Suporte(Lasanha) = 14,45%
Lift(R) = 5,381.
RI(R) = 2,701%.
Convicção(R) = 3,849.

Figura 1: Medidas de interesse de uma regra de associação

O valor da medida de suporte indica que 3,32% das famílias cariocas entrevistadas pela Fundação Getulio Vargas adquiriram ambos os produtos em suas compras mensais: {*Strogonoff de Frango (caixa)*} e {*Lasanha (caixa)*}.

A confiança da regra indica que a probabilidade de uma família ter comprado o produto {*Lasanha (caixa)*}, dado que ela também comprou o produto {*Strogonoff de Frango (caixa)*}, é de 77,78%.

O valor do $lift$ indica que a compra do produto {*Lasanha (caixa)*} é 5,381 vezes maior entre as famílias que também compram {*Strogonoff de Frango (caixa)*}.

O valor do índice RI indica que a diferença entre o valor de suporte real e o valor de suporte esperado da regra é de 2,701%.

Finalmente, o valor da medida de convicção serve para indicar que a probabilidade da compra do produto {*Strogonoff de Frango (caixa)*} ocorrer sem que o produto {*Lasanha*} seja comprado é 3,849 vezes menor do que o esperado.

A análise que acaba de ser apresentada demonstra que cada uma das medidas objetivas é empregada para fornecer uma informação específica a respeito de uma regra de associação. Este fato evidencia que todas estas medidas são importantes para o usuário de uma ferramenta de mineração de dados. A abordagem adotada para a mineração de regras de associação permite com que este usuário possa, numa etapa de pós-processamento, trabalhar os resultados minerados de diversas formas. O usuário poderia, por exemplo, ordenar as regras mineradas de acordo com cada um dos índices, analisar estes resultados e descobrir quais as regras mais fortes de acordo com cada medida. Considere as Figuras 2 e 3 que apresentam, respectivamente, regras com valores altos de $lift$ e de RI. Observe como estas regras possuem características diferentes e como a utilização conjunta de todas as medidas de interesse é capaz de auxiliar na interpretação das associações na fase de pós-processamento.

R: Champignon ⇒ Palmito em Conserva
Suporte(R) = 5,21%.
Confiança(R) = 73,33%
Suporte(Champignon) = 7,11%
Suporte(Palmito em Conserva) = 15,88%
Lift(R) = 4,619.
RI(R) = 4,85%.
Convicção(R) = 3,154.

R: Alface Americana ⇒ Brócolis
Suporte(R) = 5,45%.
Confiança(R) = 69,70%
Suporte(Alface Americana) = 7,82%
Suporte(Brócolis) = 25,59%
Lift(R) = 2,723.
RI(R) = 3,90%.
Convicção(R) = 2,455.

Figura 2: Exemplos de regras com valor alto para o índice $lift$

4 Conclusões

Neste trabalho foram revistas diferentes medidas de interesse obetivas e subjetivas que podem ser utilizadas em processos de mineração de regras de associação. As medidas de interesse objetivas suporte e confiança foram avaliadas através de um teste realizado sobre uma base de dados real (base de dados da POF). Os resultados obtidos demonstraram que a medida da confiança possui a deficiência de não considerar a questão da dependência entre os itens de dados. Para solucionar este problema, podem ser utilizadas as medidas objetivas $lift$, RI e convicção. As duas primeiras que são capazes de analisar a dependência entre o antecedente e o conseqüente de uma regra de associação, enquanto a última fornece informações a respeito da implicação. Novas avaliações foram realizadas e os resultados obtidos indicaram que a utilização conjunta das medidas de interesse suporte, confiança, *lift*, RI e convicção diminui

R: Biscoitos Recheados ⇒ Achocolatado em Pó
Suporte(R) = 35,07%.
Confiança(R) = 80,43%
Suporte(Biscoitos Recheados) = 43,60%
Suporte(Achocolatado em Pó) = 56,87%
Lift(R) = 1,414.
RI(R) = 10,27%.
Convicção(R) = 2,204.

R: Carne Seca: Dianteiro ⇒ Lingüiça
Suporte(R) = 18,25%.
Confiança(R) = 60,16%
Suporte(Carne Seca: Dianteiro) = 30,33%
Suporte(Lingüiça) = 36,49%
Lift(R) = 1,648.
RI(R) = 7,18%.
Convicção(R) = 1,594.

Figura 3: Exemplos de regras com valor alto para o índice RI

a chance da mineração de regras óbvias e irrelevantes e, além disso, possibilita aos usuários a realização de análises alternativas sobre uma mesma regra, enriquecendo o poder de entendimento a respeito das associações.

Este trabalho também abordou a importância da utilização de medidas de interesse subjetivas na mineração de regras de associação. Uma regra costuma ser interessante subjetivamente quando é útil e surpreendente para um analista que a examina. Por esta razão, como trabalho futuro, intenciona-se desenvolver um algoritmo para mineração de regras inesperadas em bases de dados de transações.

Referências

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

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

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

[4] S. Brin, R. Motwani, J. D. Ulman e S. Tsur, Dynamic Itemset Counting and Implication Rules for Market Basket Data, *Proc. of the ACM SIGMOD Intl. Conf. on Management of Data*, Arizona, Estados Unidos, 1997, 255–264.

[5] 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.

[6] Fundação Getulio Vargas, Instituto Brasileiro de Economia, Divisão de Gestão de Dados (FGV/IBRE/DGD). Informações Disponíveis em [*http://www.fgv.br/ibre/dgd*], 2005.

[7] E. C. Gonçalves, I. M. B. Mendes e A. Plastino, Mining Exceptions in Databases, *Proc. of the* 17^{th} *Australian Joint Conf. on Artificial Intelligence (LNAI 3339)*, Cairns, Australia, 2004, 1081-1086.

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

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

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

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

[12] R. Viana, Mineração de Dados: Introdução e Aplicações. *SQL Magazine, Ed. 10, Ano 1*, 2004.

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

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