Geração de Regras de Associação Quantitativas

Bruno Pôssas Wagner Meira Jr. Rodolfo Resende
Departamento de Ciência da Computação – UFMG
Caixa Postal 702 - 30161-970
Belo Horizonte – MG
{bavep,meira,rodolfo}@dcc.ufmg.br

Resumo

As instituições têm investido cada vez mais em explorar a informação e conhecimento presentes nos dados correspondentes às suas atividades. Alguns tipos de informação podem ser obtidos através de uma ou mais consultas aos bancos de dados de uma organização. Entretanto, existem várias informações que não são obtidas simplesmente utilizando as consultas convencionais. A mineração de dados corresponde a um conjunto de técnicas para obtenção de informação que não pode ser obtida através de consultas convencionais. Uma destas técnicas é denominada mineração de regras de associação. As regras de associação são expressões que indicam afinidade ou correlações entre dados. Diversos trabalhos têm descrito várias abordagens sobre como definir e obter regras de associação. A maioria dos trabalhos utiliza regras de associação onde a quantificação dos dados não é considerada. Recentemente alguns trabalhos começaram a propor métodos de obtenção de regras de associação que levam em consideração aspectos quantitativos dos dados envolvidos. Este trabalho apresenta um algoritmo para geração de regras de associação envolvendo aspectos quantitativos dos dados. São apresentados ainda resultados da aplicação do algoritmo em um banco de dados da área bancária.

Abstract

The organizations are increasingly investing in exploring the information and knowledge imbedded in the data related with their activities. Some types of information can be obtained from one or more queries available in the database of the organization. However there are lots of information that can not be obtained simply by using conventional queries. Data mining corresponds to a set of techniques for obtaining information that is not available through conventional queries. One of these techniques is called association rules mining. Association rules are expressions that indicate affinity or correlation among data. Several works have been describing how to define and obtain association rules. The majority of the works uses association rules where the quantification of the data is ignored. Recently some works started to propose methods to obtain association rules that consider quantitative aspects of the data. This work presents an algorithm that generates association rules involving quantitative aspects of the data. We also present results of aplying the algorithm in a banking database.

Palavras chaves: *Mineração de dados, Banco de Dados, Algoritmos*

1 Introdução

Os sistemas de gerência de bancos de dados (*SGBDs*) implementam várias funcionalidades. Na maioria dos *SGBDs* o usuário pode expressar vários tipos de consultas. Infelizmente nem todo tipo de consulta pode ser expresso através dos mecanismos usuais. As regras de associação são expressões que permitem responder a certas consultas que normalmente não podem sequer ser expressas nas linguagens de consulta usuais. O problema da geração de regras de associação a partir de um certo conjunto de dados foi apresentado por Agrawal, Imielinski e Swami [1]. A motivação do trabalho se deu em função de demandas administrativas de organizações da área de supermercados. Essas regras também são aplicadas na solução de vários outros problemas, desde suporte a decisão em telecomunicações até planejamento.

A proposta em Agrawal [1] considera que os produtos de um supermercado correspondem a um conjunto de itens. Um subconjunto dos items adquiridos por um cliente é uma transação. Uma transação pode corresponder aos itens adquiridos em uma única compra ou ao longo de um certo período de tempo. Entretando, a quantidade adquirida de cada item é ignorada. Isto significa que as transações "30 litros de um produto A, 20 kilos de um produto B e 50 unidades de um produto C" e "1 litro de A, 0.5 kilo de B e 2 unidades de C" seriam consideradas equivalentes.

Srikant e Agrawal [3] apresentaram uma nova abordagem de definição e geração de regras de associação levando em conta aspectos quantitativos e categóricos dos produtos. Na Seção 2, a partir do que eles denominam problema de mineração de regras de associação quantitativas, estendemos a sua formalização. O uso de aspectos quantitativos pode tornar regras de associação ainda mais precisas, versáteis e abrangentes.

Em nossa investigação propomos uma definição formal mas ainda intuitiva do problema de geração de regras de associação quantitativas a partir do trabalho de Srikant e Agrawal [3]. No entanto nossa abordagem para a geração das regras é totalmente diferente. Srikant e Agrawal geram regras mapeando uma instância quantitativa do problema em uma instância não-quantitativa. Nossa abordagem utiliza propriedades espaciais proporcionadas pela consideração de aspectos quantitativos. Este trabalho apresenta um algoritmo para a geração de regras de associação e apresenta o resultado da aplicação deste algoritmo em um banco de dados da área bancária. Esse trabalho foi publicado no 14 ^{o} Simpósio Brasileiro de Banco de Dados [7].

O restante do artigo está organizado da seguinte maneira. Na próxima subseção descrevemos os trabalhos correlatos. A Seção 2 apresenta problema de geração de regras de associação quantitativas e um algoritmo para essa geração. A Seção 3 apresenta uma avaliação dos resultados experimentais obtidos. Finalmente, as conclusões e os trabalhos futuros são apresentados na Seção 4.

1.1 Trabalhos Correlatos

Muitos algoritmos para determinação de regras de associação [2, 8] foram propostos na literatura desde a introdução deste problema em [1]. Entretanto, existem poucos trabalhos relacionados à abordagem quantitativa. Uma forma similar do problema é descrita por Piatetsky-Shapiro [9]. As regras quantitativas são obtidas, de acordo com a formulação de Agrawal [3], através da discretização dos valores para um item em intervalos eqüidistantes e posterior utilização de um algoritmo tradicional de regras de associação.

No entanto, essa abordagem gera resultados pouco intuitivos quando aplicados a intervalos cuja separação entre os valores tenha algum significado. Miller [6] explora essa dificuldade e propõe o agrupamento desses valores através de técnicas de clusterização. Ele usa a premissa de que a distância entre os itens deve ser levada em conta para a criação dos intervalos.

2 Regras de Associação Quantitativas

2.1 Definição Formal

Seja $A = \{a_1, a_2, ..., a_n\}$ um conjunto de atributos de uma tabela. Seja V um conjunto de inteiros não-negativos que representam as quantidades para qualquer atributo e V_a o conjunto de quantidades para um atributo a. Definimos um item i como uma dupla $\langle a, q_a \rangle$, onde a representa um atributo (que identifica o item) e $q_a \in V_a$, que representa a quantidade adquirida do referido item. Um *itemrange* ou item de abrangência corresponde a uma tripla $\langle a : l_a - h_a \rangle$, onde a é um atributo, $l_a \in V_a$, $h_a \in V_a$, e $l_a \leq h_a$, que denota uma faixa de quantidades de aquisição do item identificado pelo atributo a.

Seja D um conjunto de transações, onde cada transação T é um conjunto de n itens $t_1 \ldots t_n$. Dizemos que uma transação T *satisfaz* um conjunto de *itemranges* I se para cada $\langle a_I : l_a - h_a \rangle \in I$ existe um $\langle a_T, q_a \rangle \in T$ com $a_I = a_T$ e $l_a \leq q_a \leq h_a$. Uma regra de associação quantitativa é uma implicação da forma $X \rightarrow Y$, onde $X \subset I$, $Y \subset I$, $X \cap Y = \emptyset$ e I corresponde a um conjunto de *itemranges*. Deve ser observado que uma regra contempla uma e somente uma faixa de valores para um dado item. A regra $X \rightarrow Y$ é válida para o conjunto de transações D com confiança c se $c\%$ das transações em D que satisfazem X também satisfazem Y. A regra $X \rightarrow Y$ tem suporte s no conjunto de transações D se $s\%$ das transações em D satisfazem $X \cup Y$. Dado um conjunto de transações D, o problema da geração de regras de associação quantitativas é gerar todas as regras que tenham suporte e confiança maiores que os valores mínimos especificados (*minsup* e *minconf*, respectivamente).

O problema de geração de regras de associação quantitativas pode ser dividido em três etapas: (i) Contar o suporte para os conjuntos de *itemranges*; (ii) Encontrar os conjuntos de *itemranges* que possuam suporte maior do que o suporte mínimo (chamados de cojuntos freqüentes); e (iii) Gerar as regras de associação a partir dos conjuntos freqüentes de *itemranges*. Ressaltamos que, do ponto de vista de elaboração do problema, a geração de regras quantitativas é idêntica à geração de regras de associação tradicionais, entretanto, a consideração das quantidades adquiridas de cada item acrescenta uma dimensão a mais nas regras geradas, o que normalmente aumenta a sua quantidade de informação.

Podemos distinguir três questões básicas a serem tratadas quando geramos regras de associação quantitativas: (1) **Agrupamento**: Como os múltiplos valores para um dado item podem ser agrupados, e qual a melhor representação para esse agrupamento. A utilização de *itemranges*, como mencionado, define de forma bem clara como são feitos os agrupamentos. (2) **Combinação**: Uma vez que cada conjunto pode conter itens com múltiplos valores, a combinação de dois conjuntos para criação de um novo conjunto não é trivial. Uma vez que somente é considerada uma faixa por atributo em cada regra (ou seja, cada atributo somente aparece uma vez por regra), a tarefa de combinação torna-se trivial e consiste de uma união entre os conjuntos originais. (3) **Relevância das regras**: O número potencial de regras de associação geradas cresce exponencialmente com as quantidades possíveis para um dado item. Desta forma torna-se necessária a criação de um critério de relevância da regra, que mede a sua especificidade. Assim, como discutido na Seção 2.2, a relevância da regra é função não apenas dos tradicionais suporte e confiança, mas também da extensão da faixa de quantidades de cada atributo e da variação dessa extensão entre os vários itens.

2.2 Relevância da Regra

Como apresentado na Seção 2.1, o número de regras de associação cresce como função do número de valores possíveis para um dado item. Nesta seção propomos um novo critério, *relavância*, para essas regras, que difere dos apresentados por Agrawal [3] e Miller [6].

Para motivar e explicar a nossa métrica de relevância, considere as seguintes regras:

$$\begin{array}{lcll} \langle A:5-10\rangle & \rightarrow & \langle B:3-4\rangle & (10\%\ sup,\ 80\%\ conf) \\ \langle A:8-10\rangle & \rightarrow & \langle B:3-4\rangle & (8\%\ sup,\ 80\%\ conf) \end{array}$$

Claramente a primeira regra contém a segunda, mas intuitivamente a segunda regra apresenta uma informação mais precisa sobre a correlação entre ocorrências simultâneas dos items A e B. Desta forma, acreditamos que quanto "mais específica" uma regra, maior a sua utilidade e relevância, para um dado valor fixo de confiança. Completando, essas regras nos mostram o compromisso entre regras que tenham maior relevância e sejam menos freqüentes e regras que tenham menor relevância e sejam mais freqüentes.

Mais formalmente, a especificidade dos intervalos e o suporte da regra são os fatores que determinam a sua relevância: $Rel = E * sup$, onde E é a especificidade de uma regra e sup o seu suporte. Quanto maior esse valor, maior a relevância da regra.

A especificidade de uma regra é calculada como sendo a razão entre o somatório das especificidades dos *itemranges* e o tamanho da regra (ou seja, o número de *itemranges* que compõe a regra), que é uma medida normalizada, permitindo comparar regras de tamanhos diferentes. A especificidade de *itemranges*, por sua vez, é definida como a razão entre o número de valores possíveis do atributo constantes do *itemrange* e o número de valores possíveis de valores do atributo. Essa medida de especificidade leva em consideração distribuições de freqüência não uniformes, uma vez que faixas de valores esparsas tem baixa especificidade. Uma métrica correlata de relevância da regra é o desvio D entre as especificidades dos *itemranges*. Esse desvio quantifica a discrepância entre os *itemranges* que compoem a regra e quanto menor for o desvio, melhor a relevância da regra. Formalmente, a espeficidade e o desvio podem ser, respectivamente, definidos como:

$$E = \frac{\sum_{i=1}^{tam} \frac{\|v_i\|}{\|V_i\|}}{tam} \quad , \quad D = \frac{\sum_{i=1}^{tam} |E - \frac{\|v_i\|}{\|V_i\|}|}{tam * E}$$

onde tam é o tamanho da regra, v_i é o conjunto de quantidades do *itemrange* associado ao atributo i e V_i é o conjunto de quantidades possíveis de ocorrência do atributo i.

2.3 Propriedades

Nesta seção realizamos uma comparação entre regras de associação quantitativas e não quantitativas.

As regras de associação quantitativas têm vários aspectos em comum com as regras de associação tradicionais. Conforme discussão da Seção 2.4 várias propriedades dessas últimas são também aplicáveis, o algoritmo para regras de associação quantitativas é muito semelhante ao tradicional *Apriori*, se diferenciando pelo conceito de *itemranges*.

A propriedade mais interessante de regras de associação quantitativas é a flexibilidade em termos de níveis de informação provida. Para tornar clara essa propriedade, vamos explorá-la através de um exemplo envolvendo três atributos, figuradamente produtos vendidos em uma padaria: Leite (L), Pão (P) e Manteiga (M). Um conjunto de dez transações envolvendo esses produtos poderia ser:

1.	L	2					**6.**	L	1	P	4		
2.			P	3	M	1	**7.**	L	1	P	3		
3.	L	1	P	2			**8.**	L	1	P	2		
4.	L	2	P	4	M	1	**9.**	L	2	P	6	M	1
5.	L	3	P	8	M	1	**10.**	L	1	P	2		

Encontramos a seguinte regra contendo L e P, em termos do Apriori tradicional, com suporte e confiança indicados:

$$L \quad \rightarrow \quad P \qquad (80\%\ sup,\ 89\%\ conf)$$

A utilização de regras de associação quantitativas pode prover informações com um nível de detalhe maior. Para as dez transações apresentadas anteriormente teríamos, por exemplo, as seguintes regras entre L e P:

$$\begin{array}{lcll} \langle L:1-1\rangle & \rightarrow & \langle P:2-2\rangle & (30\%\ sup,\ 60\%\ conf) \\ \langle L:1-1\rangle & \rightarrow & \langle P:3-4\rangle & (20\%\ sup,\ 40\%\ conf) \\ \langle L:1-1\rangle & \rightarrow & \langle P:2-4\rangle & (50\%\ sup,\ 100\%\ conf) \\ \langle L:1-3\rangle & \rightarrow & \langle P:2-8\rangle & (80\%\ sup,\ 100\%\ conf) \\ \langle L:2-3\rangle & \rightarrow & \langle P:4-8\rangle & (30\%\ sup,\ 75\%\ conf) \end{array}$$

Essas regras nos mostram o compromisso entre regras que tenham maior relevância e sejam menos freqüentes e regras que tenham menor relevância e sejam mais freqüentes. Por exemplo, a quarta regra diz que "80% das pessoas que compram entre 1 e 3 litros de leite, também compram de 2 a 8 pães", já a quinta regra é muito mais precisa, dizendo que "30% das pessoas que compram entre 2 e 3 litros de leite, também compram de 4 a 8 pães". Apesar desse exemplo ser bastante intuitivo, ou seja, "quem compra mais leite compra mais pães", a quantificação dessas correlações não é trivial a partir das regras de associação tradicionais. Finalmente, como será discutido na próxima seção, o algoritmo proposto determina essas faixas de forma automática e transparente.

2.4 Algoritmo *Apriori* Quantitativo

Durante cada iteração do algoritmo somente os conjuntos determinados como freqüentes na iteração anterior são usados para gerar conjuntos candidatos, cujo suporte é determinado na iteração corrente. Um passo de corte elimina qualquer conjunto candidato que tenha um subconjunto que não seja freqüente. O algoritmo termina no passo k, se não há nenhum conjunto candidato de tamanho k ($k - itemsets$). Como mencionado, a grande diferença entre o algoritmo proposto e o apriori original é com relação à geração dos intervalos que é realizada simultaneamente à contagem de suporte.

Uma vez determinados todos os conjuntos freqüentes e as suas faixas de valores, as regras de associação são derivadas. A estrutura geral do algoritmo é apresentada na Figura 1. A sintaxe e a semântica das construções utilizadas em nossos algoritmos são similares às encontradas nos trabalhos de Agrawal e outros.

```
1.  L1 = {frequent 1-itemsets}
2.  for (k=2; Lk-1<>0; k++) {
3.    Ck = generate_candidates(Lk-1);
4.    for all transactions T in DB
5.      for all subsets t in T
6.        if (c is in Ck: c=t) c.count++;
7.    Lk = {c in Ck | c.count >= minsup};
8.  }
9.  for all Lk, k>2
10.    generate_rules(Lk, Lk);
```

Figura 1: Algoritmo *Apriori* Quantitativo

São empregadas duas estruturas de dados para a geração de regras de associação quantitativas: árvores de conjuntos e de intervalos. A árvore de conjuntos mantém os *itemsets*, à semelhança do *Apriori* tradicional. Esta árvore é organizada em níveis e cada nível possui uma ou mais listas de nodos. Cada nodo representa um *itemset* e contém um identificador de item e um contador de ocorrências do *itemset*. Os componentes de cada *itemset* são o item armazenado no nodo propriamente dito e os itens armazenados em todos os ancestrais do nodo. Desta forma, o i-ésimo nível da árvore contem *itemsets* de tamanho i.

Cada um dos nodos da árvore de conjuntos possui uma árvore de intervalos. As árvores de intervalos são inspiradas em árvores KD [4] e têm por objetivo armazenar as informações sobre os *itemranges* e sua frequência de ocorrência. Estas árvores são binárias e cada nodo possui, além de um conjunto de *itemranges* a que denominamos *rangeset*, um contador de ocorrências do referido *rangeset*, o discriminante do nodo. Essa árvore satisfaz duas propriedades: (i) **acumulação ancestral:** o valor de ocorrência de um nodo é igual à soma dos valores de ocorrência dos nodos filhos e (ii) **inclusão ancestral:** os *itemranges* dos nodos filhos são sub-intervalos dos *itemranges* do nodo pai [1].

[1] Assumimos que um intervalo é sub-intervalo dele próprio.

3 Resultados

Acreditamos que a forma mais eficiente de validar a implementação do algoritmo quantitativo é através da sua utilização em uma aplicação real. Desta forma, apresentamos nesta seção os resultados obtidos pela aplicação do algoritmo a dados de transações bancárias e de cartões de crédito.

3.1 Descrição dos Dados

Os dados utilizados, foram retirados de um concurso de Mineração de Dados promovido pelo congresso *PKDD 99* [5]. Utilizamos duas tabelas desse banco de dados para gerar o arquivo de entrada para o nosso algoritmo: (i) tabelas de transações bancárias e (ii) tabela de distritos. Fizemos a combinação dessas duas tabelas, preparando uma amostra para a execução do *Apriori* quantitativo. A tabela de transações contém 442954 registros e a amostra utilizada continha todos esses registros e possuia um tamanho de aproximadamente 110 MB.

3.2 Análise dos Resultados

Através da análise do resultado da amostra de dados utilizada, comprovamos a eficácia do algoritmo proposto. As principais relações encontradas foram: (i) taxas de desemprego e criminalidade, (ii) faixas de salário e gastos com cartões de crédito e retirada de dinheiro, (iii) cidades e distritos com maiores salários e menor criminalidade, (iv) retirada de dinheiro e número de grandes cidades por região, etc. Alguns exemplos interessantes dessas relações são apresentados nas regras abaixo:

ENTERPRENEURS 81-167	→	VYBERKARTOU 100-8000	(01% sup, 97% conf)
UNEMPLOYMENT95 1.79-2.77	→	CRIMES95 1003-6041	(10% sup, 67% conf)
UNEMPLOYMENT96 2.21-4.28	→	CRIMES96 1099-6261	(30% sup, 72% conf)
INHABITANTS10000 1-5	→	CRIMES95 818-85677	(28% sup, 93% conf)
INHABITANTS10000 1-5	→	CRIMES96 888-99107	(28% sup, 91% conf)
VYBER 1-1200	→	INHABITANTS500_1999 1-70	(05% sup, 84% conf)
SALARY 8110-12541	→	VYBER 1-1500	(29% sup, 99% conf)

Analisando as regras obtidas, podemos concluir que: (1) Em cidades em que o número de empresários está no intervalo de 81 a 167, a retirada de dinheiro através de cartões de crédito varia de US$ 100 a US$ 8000. (2) O aumento no número de desempregados de 1995 para 1996 proporcionou um aumento no número de crimes cometidos. (3) Houve um aumento na criminalidade de 1995 para 1996 nas regiões que possuem cidades com mais de 10000 habitantes. (4) Pessoas que retiram até US$ 1200 dos bancos, moram em regiões cujo número de cidades com o número de habitantes entre 500 e 1999 apresenta um variação de 1 a 70. (5) Pessoas com salário entre US$ 8110 e US$ 12541 retiram dos bancos valores entre US$ 1 e US$ 1500.

O número de regras geradas utilizando o critério de relevância, proposto na Seção 2.2, diminuiu consideravelmente, em comparação com a geração sem o critério apresentado. A Figura 2 apresenta o decaimento do número de regras geradas de acordo com o aumento da relevância para a amostra de dados analisada. O tempo de execução do algoritmo *Apriori* quantitativo foi aproximadamente 20% menor do que o *Apriori* tradicional aplicado à discretização da amostra avaliada.

4 Conclusões e Trabalhos Futuros

Neste trabalho apresentamos uma definição formal para o problema de geração de regras de associação quantitativa e propomos um algoritmo, baseado no algoritmo *Apriori* [2], para resolução desse problema. A abordagem quantitativa adiciona uma nova dimensão na análise de conjunto de itens possibilitando um resultado mais preciso.

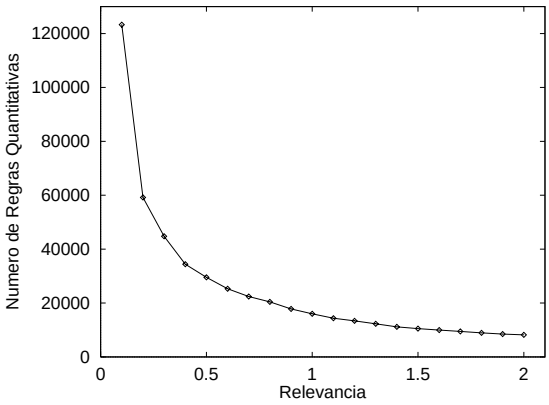

Figura 2: Número de Regras Quantitativas x Relevância

Propomos a utilização de uma variação de uma árvore KD [4] para armazenamento das faixas de valores para um dado conjuntos de itens (*itemrange*). O uso de quantidades associadas aos itens faz com que haja aumento do número de regras de associação. Apresentamos uma métrica, que define a relevância de uma regra. Esta métrica é utilizada para controle do número de regras, sendo geradas somente as regras mais específicas. Comprovamos que essa métrica diminui consideravelmente o número de regras geradas.

Pretendemos continuar este trabalho estendendo os algoritmos apresentados na Seção 2.4 para múltiplas faixas de valores e permitindo o armazenamento de qualquer tipo de faixas de valores, como por exemplo, faixa de valores reais. Também pretendemos analisar mais detalhadamente a aplicação do algoritmo proposto a outros de problemas reais.

Referências

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Intl. Conf. Management of Data*, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *The 20th Int. Conf. on Very Large Databases*, September 1994.

[3] R. Agrawal and R. Srikant. Mining quantitative association rules in large relational tables. Technical report, IBM Almaden Research Center, San Jose, CA, 1996.

[4] J. Bentley. Multidimensional binary search trees used for associative searching. In *Communications of ACM*, September 1975.

[5] P. Berka. 3rd european conference on principles and practice of knowledge discovery in databases. Disponível em `http://lisp.vse.cz/pkdd99/`.

[6] R. Miller and Y. Yang. Association rules over interval data. Technical report, Ohio State University, May 1997.

[7] B. Pôssas, F. Ruas, W. Meira, and R. Resende. Geração de regras de associação quantitativas. In *14th Simpósio Brasileiro de Banco de Dados.*, Outubro 1999.

[8] A. Savasere, E. Omiecinski, and S. Navathe. An efficient algorithm for mining association rules in large databases. In *2lst Intl. Conf. on Very Large Databases*, 1995.

[9] G. Shapiro-Piatetsky and W. Frawley. *Knowledge Discovery in Databases.* AAAI Press/MIT Press, New York and Toronto, 1991.