DESENVOLVIMENTO DO MÓDULO DE ASSOCIAÇÃO PELO ALGORITMO APRIORI NA *SHELL* DE DATA MINING ORION

Diego Paz Casagrande[1]
Merisandra Côrtes de Mattos[2]
Rafael Charnovscki[3]
Priscyla Waleska Simões[4]
Jane Bettiol[5]

Resumo

Dada a importância da informação aplicada às mais diversas áreas, o conhecimento implícito em bases de dados deve ser explorado e difundido. Assim, esta pesquisa tem por finalidade a exploração e o aprofundamento do conhecimento acerca do *Data Mining*, um campo da Inteligência Artificial bastante recente e em evidência. Levando em conta a perceptível evolução das tecnologias de informação e o uso intenso dos sistemas de bancos de dados, este artigo apresenta o módulo de associação da ferramenta *Data Mining* Orion, na qual está presente a implementação do algoritmo *Apriori*, considerado o mais utilizado para a geração

[1] Acadêmico do Curso de Ciência da Computação. (diegocasagrande@hotmail.com)
[2] Mestre em Ciência da Computação. Professora Orientadora. Curso de Ciência da Computação. Laboratório de Informática Médica e Telemedicina. Grupo de Pesquisa em Inteligência Computacional Aplicada. (mem@unesc.net)
[3] Mestre em Ciência da Computação. Professor Co-orientador. Curso de Ciência da Computação. Laboratório de Informática Médica e Telemedicina. (charnovs@unesc.net)
[4] Mestre em Ciência da Computação. Professora Co-orientadora. Curso de Ciência da Computação. Laboratório de Informática Médica e Telemedicina. Grupo de Pesquisa em Inteligência Computacional Aplicada. Curso de Medicina. (pri@unesc.net)
[5] Doutora em Medicina. Professora Colaboradora. Curso de Medicina.

das regras de associação. No referido módulo, estão presentes os atributos de suporte e confiança, que conferem a propriedade antimonotonia à relação e garantem a validade das regras extraídas. O módulo de associação da *Shell* Orion permite a conexão com diversos Sistemas Gerenciadores de Bancos de Dados (SGBD), como o PostgreSQL, Firebird e MySQL. A implementação da técnica de associação foi realizada no ambiente de desenvolvimento NetBeans 4.1, que utiliza arquitetura Java.

Palavras-chave: *Data Mining*. Associação. Algoritmo Apriori.

Introdução

A importância da informação e o grande volume de dados armazenados acarretam, cada vez mais, a necessidade de meios com os quais se viabilize a interação com as bases de dados, para que delas se extraia conhecimento útil. Diante disso, o *data mining* destaca-se por permitir a exploração e descoberta do conhecimento implícito. Esse processo constitui a etapa fundamental da Descoberta de Conhecimento em Base de Dados (DCBD), a partir da qual se obtêm conhecimentos que antes não teriam sido estudados ou mesmo imaginados pela incapacidade de se trabalhar com grande quantidade de dados.

No processo do *data mining*, necessita-se de uma *shell* que efetue a interação entre o usuário e a base de dados, facilitando, assim, a descoberta do conhecimento em bases de dados. Esse processo automatizado gera várias relações que são desconhecidas e, dificilmente, seriam demonstradas em análises estatísticas. Não obstante isso, tem-se uma carência de ferramentas livres e em português, já que as existentes são comercializadas a um alto custo.

A pesquisa aqui apresentada consiste no desenvolvimento do módulo da tarefa de associação pelo algoritmo *Apriori* na *shell* de *data mining* denominada Orion. Trabalhos de pesquisa como este podem proporcionar uma alternativa de aprendizagem no que se refere ao funcionamento de algoritmos que realizam tarefas específicas constantes do processo do *data mining*.

A tarefa de associação

A tarefa de associação se utiliza dos conceitos da inteligência artificial, de algoritmos estatísticos e da teoria dos conjuntos para representar padrões, tendo

como objetivo encontrar todos os conjuntos de itens que freqüentemente ocorrem de forma conjunta em uma base de dados.

Uma regra de associação simples dá-se por meio da utilização de métodos probabilísticos sobre a co-ocorrência de determinados eventos em uma base de dados. Como exemplo para a regra, assumem-se as seguintes variáveis como binárias: IF A=1 AND B=1 THEN C=1, com probabilidade p.

Assim, tem-se p = p(C = 1|A = 1, B = 1), sendo p a probabilidade condicional de C ocorrer, dados os valores de A e B. A partir disso, pode-se chegar a conclusões como: se os itens A e B tiverem o valor 1, o item C também terá. Esse modelo de associação, no entanto, é considerado simplório e usado apenas para dar uma visão geral do processo (HAND; MANNILA; SMYTH, 2001). Pode-se abstrair o processo de encontrar regras de associação relevantes tomando o exemplo clássico do carrinho de supermercado. Considerando-se uma matriz [n][p], onde n são as linhas que representam a compra e p são as colunas que indicam os produtos, aplicam-se as regras de associação a fim de encontrar padrões válidos (HAND; MANNILA; SMYTH, 2001). Considerando a base de dados (Figura 1), apresenta-se o seguinte exemplo, no qual os produtos comprados são indicados pelo valor 1:

Compra	Prod_1	Prod_2	Prod_3	Prod_4	Prod_5	Prod_6
1	1	1	0	0	0	0
2	0	0	1	0	0	0
3	0	0	0	1	1	1
4	0	0	0	1	1	1
5	1	1	0	0	0	0
6	0	0	1	1	0	1
7	1	0	1	0	0	0
8	1	1	0	0	0	0
9	1	1	1	1	1	1
10	1	0	0	1	1	1

Figura 1. Exemplo de base de dados

Com as regras de associação do tipo A a B, pode-se verificar a seguinte relação:

IF (Prod_4 = 1) AND (Prod_5 = 1) THEN (Prod_6 = 1)

Essas regras, na forma A à B, podem possuir atributos que limitam a quantidade de regras extraídas e as validam no conjunto de dados. Tais atributos são denominados de suporte e confiança por possibilitarem o descarte das regras pouco relevantes e menos freqüentes.

O suporte é determinado pela freqüência de certos conjuntos de itens em uma base de dados. Uma forma de determiná-lo se dá por meio da seguinte expressão (ARBEX, 2005):

$$Sup(A \cup B) = \frac{\text{N° de registros com } (A \cup B)}{\text{N° total de transações do BD}}$$

Assim, tomando a base de dados exposta na Figura 4, é possível ter-se como exemplo:
Sup (Prod_1) = 6/10 = 0,6 = 60%
Sup (Prod_2) = 4/10 = 0,4 = 40%
Sup (Prod_1 *U* Prod_2) = 4/10 = 0,4 = 40%

Do exemplo anterior, salienta-se que o conjunto unitário {Prod_1} ocorre seis vezes entre as dez transações descritas, possuindo o suporte igual a 60 %. Já para o {Prod_2}, há quatro ocorrências em dez, com um suporte igual a 40 % e, para o conjunto {Prod_1, Prod_2}, presente em quatro registros no banco de dados, tem-se um suporte igual a 40 %.

Outro quesito igualmente importante para determinar a validade da relação é a medida do seu grau de confiança, a qual pode ser calculada efetuando-se a divisão do suporte do conjunto pelo do antecessor da regra, conforme fórmula a seguir:

$$Conf(A \rightarrow B) = \frac{\text{N°transações que suportam}(A \cup B)}{\text{N°transações que suportam}(A)}$$

Na expressão acima, tem-se como antecessor o item A, sendo ele o elemento que determina a regra. Por exemplo, se A estiver presente, então B estará com confiança de X%, uma vez que a confiança é dada em porcentagem.

Seguindo o exemplo da Figura 1, tem-se a seguinte relação:
Conf(Prod_1 à {Prod_1, Prod_2}) = 40/60 = 0,666 = 66,6%
Conf(Prod_2 à {Prod_1, Prod_2}) = 40/40 = 1 = 100%

Desse modo, tem-se que, em 66,6 % dos registros em que Prod_1 ocorrer, o Prod_2 também ocorrerá. E, na segunda relação, na qual o Prod_2 for o antecessor da regra, tem-se que, em 100% dos casos em que o Prod_2 estiver presente, o Prod_1 também estará.

Assim, tem-se o suporte como a freqüência com que os *itemsets* ocorrem na base de dados e a confiança é tomada como a medida da força da regra,

representada pela porcentagem. Para a aplicação desses componentes às regras de associação, é necessário fixar um valor mínimo e, ao estabelecê-lo, define-se o suporte mínimo e a confiança mínima.

Os algoritmos utilizados nas regras de associação para extrair relações relevantes das bases de dados são: *Apriori,* GART e Algoritmo de Seqüência, entre outros. A seguir, aborda-se o algoritmo *Apriori*, utilizado nesta pesquisa para a extração de conhecimento nas regras de associação.

O algoritmo apriori

O algoritmo *Apriori* é utilizado para encontrar *itemsets* em grandes bases de dados (ARBEX, 2005). De modo geral, o referido algoritmo utiliza conjuntos de itens de tamanho K para gerar os próximos conjuntos de tamanho K+1. Assim, o algoritmo encontra primeiramente os grupos de itens com tamanho k=1, denominado L_1, a partir do qual se encontram os conjuntos de *itemsets* com tamanho k=2, formando L_2, e assim por diante, até que nenhum conjunto possa ser gerado (SOUZA FILHO, 2004).

O algoritmo *Apriori* utiliza duas funções: a *Apriori*-gen (L_{k-1}), responsável por gerar os conjuntos candidatos; e a *subset* (C_k, t), que aloca os novos *itemsets* e efetua a poda (descarte dos *itemsets* que não possuem o suporte mínimo) dos conjuntos pouco freqüentes, conforme demonstrado na Figura 2.

```
L1 = { large 1-itemsets };
for ( k=2; Lk-1≠∅; k++ ) do begin
    Ck = Apriori-gen( Lk-1 ); // Novos Candidatos
    forall transactions t ∈ D do begin
        Ct = subset( Ck, t ); // Candidatos contidos em t
        forall candidates c ∈ Ct do
            c.count++;
        end
    end
    Lk = { c ∈ Ck | c.count >= minsup }
end
Answer = UkLk;
```

Figura 2. Algoritmo Apriori
Fonte: AGRAWAL, R. (2005)

A tarefa de associação e o algoritmo apriori na *Shell de Datamining Orion*

No módulo de associação desenvolvido na *Shell* Orion, as regras são extraídas por meio do algoritmo *Apriori,* tornando-se possível encontrar os relacionamentos entre os itens da base de dados.

A metodologia aplicada para o desenvolvimento desta pesquisa compreendeu as etapas de definição e implementação da forma de conexão com a base de dados; modelagem do módulo de associação da *shell*; modelagem matemática do suporte e confiança e implementação do módulo de associação da *Shell* Orion.

Definição e Implementação da Conexão com a Base de Dados

Com a ferramenta e o módulo de associação da *Shell* Orion, utilizou-se a arquitetura Java, definindo-se, portanto, a conexão com a base de dados por meio do estudo da conexão Java Database Connectivity (JDBC).

A conexão padrão com o banco de dados na arquitetura Java é realizada por meio da API JDBC, desenvolvida em conjunto pela Sun Microsystems e grandes fabricantes de banco de dados, tais como Oracle e Sybase.

Mediante o estudo da realização da forma de conexão com o banco de dados via JDBC, definiu-se na implementação da *shell* Orion as conexões com os seguintes SGBD: Firebird, PostgreSQL e MySQL. A escolha desses sistemas se deu em função de possuírem licença *freeware*, além de apresentarem bom desempenho e confiabilidade. A implementação de diversos gerenciadores de bancos de dados, possibilitando a escolha conforme a necessidade do usuário, proporciona um aumento nas funcionalidades da *shell* Orion.

Modelagem do Módulo de Associação da Shell Orion

A modelagem do módulo de associação foi realizada por meio da *Unified Modeling Language* (UML), desenvolvendo-se o diagrama de caso de uso, a fim de demonstrar as interações do usuário com o módulo.

Modelagem Matemática do Suporte e Confiança

Nessa etapa, demonstrou-se como o módulo de associação da *shell* Orion trata os atributos de suporte e confiança. Para isso, utilizou-se a base de dados já categorizada, no que se refere ao tema: prevalência da asma e rinite em adolescentes escolares no município de Criciúma.

Primeiramente, realiza-se o cálculo do suporte a fim de encontrar a freqüência com que o item ou grupo de itens ocorre na base de dados. O cálculo da confiança é realizado após a geração das regras, sendo podadas as que possuem confiança com valor inferior ao mínimo estabelecido. Com isso, tem-se a Figura 3, demonstrando uma regra extraída conforme relatório gerado pelo módulo de associação da *shell* Orion.

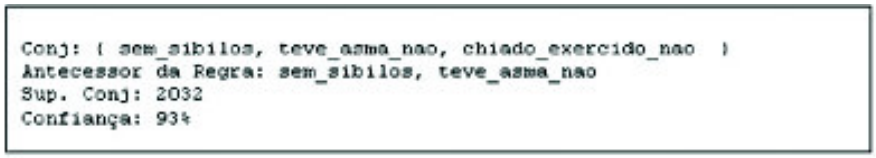

Figura 3. Exemplo de regra gerada pelo módulo de associação

De acordo com a Figura 3, apresenta-se a seguir o cálculo da confiança e do suporte:

Total de registros na base de dados: 3010
Suporte do conjunto: 2032

O valor do suporte do conjunto indica que ele está presente em 2032 registros da base de dados. Assim, verifica-se que o suporte do conjunto é de aproximadamente 67%.

A confiança pode ser calculada da seguinte forma:

Suporte do Conjunto: 2032 ou aproximadamente 67%
Suporte do Antecessor da Regra: 2166 ou aproximadamente 71%
Confiança: 2032 / 2166 = 0,93 ou aproximadamente 93%

Assim, tem-se que, para 93% dos registros em que consta o conjunto {sem_sibilos, teve_asma_nao, chiado_exercido_nao}, o antecessor da regra também está presente. Esse conjunto obedece à condição de suporte mínimo, estando presente em 2032 transações na base de dados analisada.

Com isso, verifica-se a importância do suporte e da confiança para a geração das relações e a conseqüente extração do conhecimento das bases de dados, uma vez que esses atributos determinam a quantidade de regras geradas e proporcionam uma maior validade das informações encontradas.

Implementação do Módulo de Associação da Shell Orion

A implementação do módulo de associação e do algoritmo *Apriori* foi realizada no ambiente de desenvolvimento NetBeans 4.1, implementando-se na linguagem Java e permitindo-se a conexão com os sistemas gerenciadores de bancos de dados PostgreSQL, MySQL e Firebird. Os recursos utilizados obedecem à licença *freeware* e estão disponibilizados na Internet.
A *shell* de *data mining* Orion dispõe da opção de ajuda, que traz informações do projeto e do funcionamento dos módulos. No módulo de associação, também é possível salvar o relatório gerado pelo sistema em formato RTF. Na interface principal (Figura 4), ao inicializar o sistema, o usuário escolhe qual *driver* JDBC usará. Após, deve-se escolher a tabela que tem seus atributos visualizados em uma lista. Assim, com todos os parâmetros indicados, escolhe-se a tarefa de *data mining*.
No módulo de associação (Figura 5), é possível selecionar o algoritmo de *data mining*, visualizar os conjuntos de dados iniciais, definir o suporte e a confiança necessários para a execução do algoritmo

Figura 4. Interface principal da Shell Orion

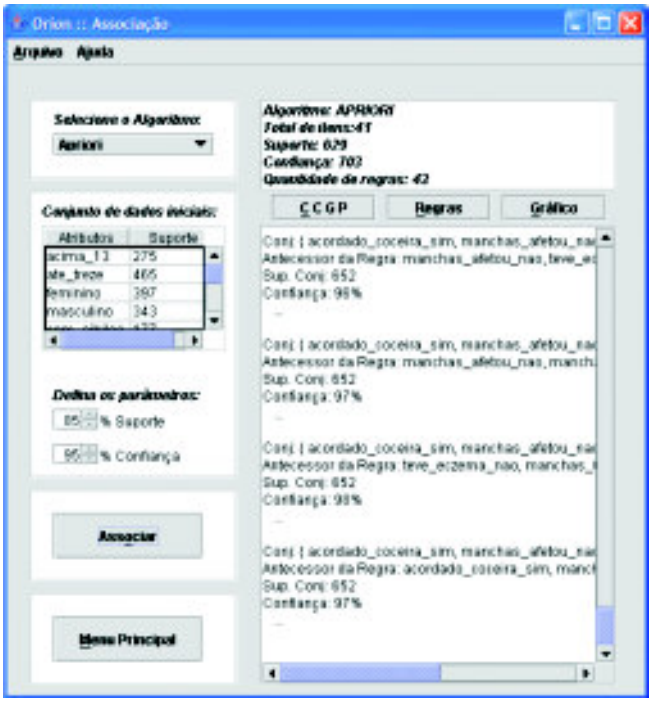

Figura 5. Interface do módulo de associação

Conclusão

Na realização desta pesquisa, os objetivos propostos foram atingidos, visto que o módulo de associação da *Shell* Orion apresenta as regras por meio das quais se viabiliza a busca por relações entre itens ou conjuntos de itens que auxiliam no processo de entendimento dos padrões presentes nas bases de dados. Nesse módulo, aplicou-se aos dados o algoritmo *Apriori*, que faz buscas recursivas a fim de encontrar conjuntos de itens, apresentando regras que satisfazem às condições de suporte e confiança. Durante a etapa de realização dos testes, pode-se perceber a dificuldade em processar grandes quantidades de dados, dificuldade que ocasionou alguns problemas logo solucionados. Com a realização dos testes, verificou-se a aplicação do algoritmo sobre a base de dados, gerando dezenas de relações que, embora imprevistas, podem contribuir com o propósito da base de dados.

Association rules and apriori algorithm in orion data mining shell

Abstract

Because of the importance of the information applied in several areas, the implicit knowledge must to be explored and diffused. This way, this research has as finality to explore and to deepen the knowledge about the Data Mining, an artificial intelligence camp very recently and in evidence. Considering the perceptible information technologies evolution and the intense use of data bank system, this article presents an association module tool of Data Mining Orion. In this module is present the implementation Apriori algorithms considered the most used for the rules association generation. In the association module are the attributes of support and trust, that grants the characteristics of the anti-monotony relation and guarantee the validity of the rules extracted. The Shell of Data Mining Orion allows the connection of Data Base Management System (DBMS) Postgres and Firebird. The implementation of the association rules was realized in the Netbeans 4.1 development ambient, that uses Java technology.

Keywords: Data Mining. Association Rules. Apriori Algorithm.

Referências

AGRAWAL, Rakesh; SRIKANT, Ramakrishnan. ***Fast Algorithms for Mining Association Rules***. San Jose, CA – USA, IBM Almaden Research Center. Disponível em: <www.almaden.ibm.com/*software*/ quest/Publications/papers/ vldb94.pdf > Acesso em: 15 jun 2005.

ARBEX, Eduardo Compasso et al. Iniciação Cientifica: *Data Mining*. Associação Educacional Dom Bosco em Resende, RJ. Disponível em: <http:// www.inf.aedb.br/datamining/index.html>. Acesso em: 14 Jun 2005.

HAND, David; MANNILA, Heikki; SMYTH, Padhraic. ***Principles of Data Mining***. Massachusetts: MIT Press, 2001.

SOUZA FILHO, Hélcio Gomes de. **Extração de Regras de Associação de um Banco de Dados Relacional**. Rio de Janeiro, Dissertação (Mestrado em Engenharia Civil). Universidade Federal do Rio de Janeiro, 2004.