Avaliação das regras de associação descobertas sob a perspectiva do usuário em relação de medidas objetivas

Viviane Dal Molin de Souza (Mestre)
Curso de Bacharelado em Sistemas de Informação – Faculdade Expoente
Deborah Ribeiro Carvalho (Doutor)
Curso de Bacharelado em Ciência da Computação – Universidade Tuiuti do Paraná

Resumo

A grande maioria dos algoritmos de *Data Mining* apresenta o conhecimento descoberto na forma de uma longa lista de regras, a partir da qual o usuário deve pesquisar e identificar aquelas que realmente possuem qualidade e têm algo a acrescentar ao processo de decisão. Ocorre que muitas vezes esta lista é tão grande que inviabiliza o trabalho de análise a ser desenvolvido. Neste caso, pode ser adotada uma fase de pós-processamento do conhecimento descoberto, a qual pode selecionar um subconjunto das regras descobertas, sob o critério da qualidade, ou seja, do grau de acerto das regras. Desta forma, o usuário receberia um conjunto bem reduzido de regras a ser avaliado, o que facilitaria a sua análise. Este artigo apresenta e discute dezesseis medidas de pós-processamento com o objetivo de avaliar se as medidas selecionadas, sob a perspectiva da qualidade e grau de interesse, efetivamente refletem o real interesse do o usuário como elemento de apoio ao processo decisório.

Palavras-chave: *Data Mining*; descoberta de conhecimento; avaliação do conhecimento; medidas de qualidade; medidas de interesse; regras de associação.

Abstract

The vast majority of Data Mining algorithms has the knowledge discovered in the form of a long list of rules, from which the user must research and identify those that do have quality and have something to add to decision-making. Often happens that this list is so large that turns the work of analysis to be developed. In this case can be taken a stage of post-processing of discovered knowledge, which can select a subset of the discovered rules, under the criteria of quality, or the precision of the rules. Thus the user would receive a greatly reduced set of rules to be evaluated, which would facilitate their analysis. This article presents and discusses sixteen measures of post-processing in order to assess whether the measures selected from the perspective of the quality and level of interest, effectively reflect the real interests of the User as an aid to decision making.

Key words: Data Mining, knowledge discovered, assessment of knowledge, quality measures, measures of interest, association rules.

Introdução

Com a grande quantidade de dados disponíveis, existe uma gama enorme de informações preciosas, mas que muitas vezes o seu volume inviabiliza que o usuário avalie, pois esta atividade ultrapassa a capacidade humana de análise e interpretação. Para facilitar a recuperação e uso destes dados, uma das alternativas que podem ser utilizadas é o processo de KDD – *Knowledge Discovery in Database*. Segundo Fayyad *et al* (1996), o processo KDD é composto de diversas etapas, a saber:

- Seleção de dados: prevê a coleta e seleção dos dados;
- Limpeza: prevê a análise dos dados coletados verificando a existência de ruídos, tratamento de valores ausentes, etc.;
- Transformação ou Enriquecimento dos Dados: trata a questão de que novos dados sejam incorporados / criados a partir dos já existentes;
- *Data Mining*: consiste em aplicar um algoritmo que efetivamente procura por padrões / relações e regularidades em um determinado conjunto de dados. Os algoritmos de *Data Mining* podem ser identificados em três tarefas principais, a saber: classificação, descoberta de regras de associação e *clustering*;
- Interpretação e Avaliação: verifica a qualidade do conhecimento (padrões) descoberto, procurando identificar se o mesmo auxilia a resolver o problema original que motivou a realização do processo KDD.

O volume do conhecimento descoberto muitas vezes é tão grande que dificulta a sua análise e, fundamentalmente, inviabiliza o seu uso no apoio a tomada de decisão. Aliado ao fato de que nestes conjuntos de padrões existem redundâncias, ou relações irrelevantes.

Dado este fato, muitas vezes surge como necessidade imperativa uma fase de pós-processamento do conhecimento descoberto objetivando viabilizar a análise a ser realizada pelo usuário. Este artigo descreve e experimenta algumas alternativas para este pós-processar regras descobertas a partir da tarefa de associação, sob o ponto de vista da qualidade e do grau de interesse. Vale destacar que uma característica interessante da deste trabalho é que os resultados da avaliação das regras obtidos a partir do pós-processamento são comparados ao resultado da avaliação realizada pelo usuário. Desta forma, é possível avaliar o quanto o uso de tais medidas para pós-processar automaticamente auxiliam na seleção/ranqueamento das regras sob o ponto de vista do que o usuário entende como interessante para a solução do problema proposto.

2 Metodologia

A tarefa de descoberta de regras de associação identifica afinidades entre itens de um subconjunto de dados e estas afinidades são expressas na forma de regras. Para os experimentos foi utilizado algoritmo Apriori (Borgelt, 2004).

Para Zanin (2002), uma variante do problema de regras de associação é analisar a seqüência, ou seja, onde as regras encontradas entre as relações podem ser usadas para identificar seqüências interessantes. Seqüências podem ser úteis para que padrões temporais possam ser identificados, como por exemplo, entre compras em uma loja, ou utilização de cartões de crédito, ou ainda tratamentos médicos.

Um dos padrões mais comuns que podem ser descobertos a partir do processo *Data Mining* são os conjuntos de regras de associação que expressam a probabilidade de um item ocorrer em conjunto a outro. Por exemplo, 80% dos clientes que adquiram o produto "A" também adquiriram o produto "B".

A associação resume-se a encontrar afinidades entre os dados de certa natureza a partir de um grande número de transações. Dessa forma, os algoritmos que descobrem regras de associação objetivam encontrar relacionamentos entre os dados (Berry & Linoff, 1997). Dado um conjunto de registros, onde cada registro é um conjunto de dados, uma regra de associação é uma expressão do tipo $X \rightarrow Y$, (Se X então Y), onde X e Y são itens (ou conjunto de itens) na forma:

$$X \cap Y = \Phi$$

O algoritmo que descobre associações fornece ao usuário padrões expressos na forma de regras, que nem sempre pode ser analisado, dado a sua grande quantidade. Nestes casos, se faz necessário um pós-processamento sobre este conjunto de regras, viabilizando assim a análise por parte do usuário. Com o método implementado e discutido neste artigo o usuário poderá ranquear essas regras descobertas a partir de medidas de qualidade e de grau de interesse.

2.1 *Pós-Processamento do Conjunto de Regras Originais*

Embora as regras de associação sejam padrões valiosos por permitirem uma percepção útil da dependência que existe entre atributos da base de dados, elas também podem apresentar dois inconvenientes: muitas regras geradas (problema da quantidade de regras); e nem todas as regras apresentarem qualidade. Não apenas o conjunto de regras descoberto pelos algoritmos de associação pode ser extensos, este problema também pode ocorrer com o classificador construído. Uma das alternativas para minimizar estes problemas é a adoção de técnicas de pós-processamento. Desta forma, o número de regras descobertas pode ser reduzido, facilitando assim a avaliação.

Existem várias medidas para avaliar as regras descobertas, propostas na literatura, as quais em geral são divididas em dois grupos, ditas subjetivas e objetivas (Freitas, 1998), (Silberschatz & Tuzhilin, 1996). A idéia básica das medidas subjetivas é que o usuário especifica suas crenças ou conhecimento prévio sobre o domínio da aplicação. A partir desta informação, uma regra é considerada surpreendente caso esta represente um conhecimento não esperado em relação à informação coletada (as crenças ou conhecimento prévio).

Em contrapartida, as medidas ditas objetivas, tentam estimar o quanto as regras podem ser surpreendentes ao usuário de uma forma mais automática e indireta, sem exigir que o usuário especifique suas crenças ou conhecimento prévio.

As medidas subjetivas têm a vantagem de considerarem diretamente as crenças do usuário, porém têm a desvantagem de serem fortemente dependentes do domínio do conhecimento e menos automáticas, exigindo uma participação intensiva do usuário na tarefa de tornar explícitas as suas crenças. De fato, pode-se afirmar que estas medidas não são apenas dependentes do domínio, mas também do usuário, uma vez que mesmo considerando um mesmo domínio de aplicação dois ou mais usuários podem ter crenças ou conhecimento sobre o domínio bastante diverso.

As medidas objetivas têm a desvantagem de ser uma estimativa indireta do quão surpreendente serão as regras para o usuário. Porém, elas têm vantagens como, por exemplo, mais independentes do domínio da aplicação e mais automáticas, liberando o usuário da tarefa de explicitar as suas crenças, o que em geral consume muito tempo do mesmo.

Desta forma, intuitivamente as medidas subjetivas são mais indicadas quando um usuário específico está disponível e tem tempo e experiência suficientes para gerar uma especificação de boa qualidade de suas crenças e conhecimento prévio; enquanto as medidas objetivas são mais indicadas para situações nas quais existe um grande número de usuário ou mesmo quando não houver nem tempo, nem experiência suficiente. Em nenhum dos casos, os dois grupos de medidas são mutuamente exclusivos, ou seja, é possível que sejam usadas medidas oriundas de ambos os grupos em uma determinada aplicação.

O foco deste artigo é testar as medidas de qualidade (objetivas), propostas por (Yao & Zhong, 1999), bem como as medidas de interesse propostas por (Tan, et al, 2002).

2.2 *Medidas Objetivas*

O método adotado usa a análise de medidas de qualidade associadas às regras individualmente. Muitas medidas de qualidade vêm sendo propostas e estudadas e cada uma delas captura características diferentes das regras.

A partir de uma regra do tipo Se <E> então <H>, é possível obter a seguinte tabela de contingência (tabela 1):

Tabela 1. Tabela de Contingência

	H	H'	Total
E	A	B	a + b
E´	C	D	c + d
Total	a + c	b + d	a+b+c+d=n

Os valores das quatro células não são independentes, estão ligados pela restrição $a+b+c+d=n$, sedo n o número total de registros. A partir da respectiva tabela de contingência, podem-se definir algumas medidas de qualidade básica como as descritas a seguir.

A generalidade é definida pela expressão 2:

$$G(E)=\frac{a+b}{n}. \qquad (2)$$

que indica o tamanho relativo do conceito E, sendo $0 \le G(E) \le 1$ o intervalo de valores possíveis.

O suporte absoluto de H provido por E é definida pela expressão 3:

$$AS(H/E) = \frac{a}{a+b}. \quad (3)$$

A quantidade, $0 \leq AS(H \mid E) \leq 1$, mostra o grau em que E implica em H.

A mudança de apoio de H provido por E é definido pela expressão 4:

$$CS(H \mid E) = \frac{an\ (a+b)(a+c)}{(a+b)\ n}. \quad (4)$$

Ao contrário do suporte absoluto, a mudança de suporte varia de -1 até 1. Pode-se considerar G(H) a probabilidade prévia de H e AS(H|E) a posterior de H após conhecer E. A diferença entre a probabilidade anterior e posterior representa a mudança da confiança em que E eventualmente causa H, para o valor negativo, pode-se dizer que E não causa H.

O suporte mútuo de E e H é definido pela expressão 5:

$$MS(E,H) = \frac{a}{a+b+c}. \quad (4)$$

Pode-se interpretar o suporte mútuo $0 \leq MS(H \mid E) \leq 1$, como medida da força de implicação dupla $E \leftrightarrow H$. Mede o grau em que E causa e só causa H.

O grau de independência entre E e H é dado pela expressão 6:

$$IND(E,H) = \frac{an}{(a+b)(a+c)}. \quad (6)$$

Isto mostra o grau de desvio de H na subpopulação restrita por E da probabilidade de H no conjunto todo. Com esta expressão, as relações de mudança e suporte se tornam claras. No lugar de usar a taxa, o posterior é definido pela diferença entre AS(H|E) e G(H). Quando E e H são provavelmente independentes, temos CS(E|H) = 0 e IND(E,H) = 1. Ainda mais, CS(H|E) ≥ 0 se e apenas se IND(E,H) ≥ 1, e CS(H|E) ≤ 0 se e apenas se IND(E,H) ≤ 1.

Nas medidas de interesse apresentadas, considera-se que quanto maior for o seu valor (métrica), maior será o grau de interesse atribuído à regra. A medida de interesse Φ-Coeficiente é obtida través da expressão 7:

$$\Phi\text{-Coeficient} = \frac{P(A,B)\ P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}} \quad (7)$$

A medida *Odds Ratio* (O) é dada pela expressão 8:

$$O = \frac{P(A,B)P(\bar{A},\bar{B})}{P(A,\bar{B})P(\bar{A},B)} \quad (8$$

A medida *Kappa* (K) é dada pela expressão 9:

$$K = \frac{P(A,B) + P(\bar{A},\bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})} \quad (9)$$

A medida *Confidence* (C) é dada pela expressão 10:

$$C = \max(P(B \mid A), P(A \mid B)) \quad (10)$$

Cosine (IS) é dada pela expressão 11:

$$IS = \frac{P(A,B)}{\sqrt{P(A)P(B)}} \quad (11)$$

A medida *Piatetsky-Shapiro's* (PS) é dada pela expressão 12:

$$PS = P(A,B) - P(A)\ P(B) \quad (12)$$

A medida *Certainty Factor* (F) é dada pela expressão 13:

$$F = \max\left(\frac{(P(B \mid A) - P(B))}{1 - P(B)}, \frac{(P(A \mid B) - P(A))}{1 - P(A)}\right) \quad (13)$$

Added Value (AV) é dada pela expressão 14:

$$AV = \max(P(B \mid A) - P(B),\ P(A \mid B) - P(A)) \quad (14)$$

Collective Strength (Col) é dada pela expressão 15:

$$Col = \frac{P(A,B) + P(\overline{A},\overline{B})}{(A)P(B) + P(A)P(B)} \times \frac{1 - P(A)P(B) - P(\overline{A})P(\overline{B})}{1 - P(A,B) - P(AB)} \quad (15)$$

Jaccard (J) é dada pela expressão 16:

$$J = \frac{P(A,B)}{P(A) + P(B) - P(A,B)} \quad (16)$$

3 Resultados e Discussão

Foram realizados experimentos com o objetivo de avaliar as medidas propostas por (Tan *et al*, 2002) e (Yao & Zhong, 1999). Os experimentos foram realizados sobre a base de dados de aproveitamento dos alunos da Universidade Tuiuti do Paraná, para os Cursos de Bacharelado em Sistemas de Informação (BSI), Tecnologia em Processamento de Dados (PD) e Bacharelado em Ciência da Computação (CC). As regras de associação foram extraídas a partir do algoritmo Apriori (Borgelt, 2004).

Essa base de dados se refere ao aproveitamento de 205 alunos para BSI, 128 alunos para PD e 364 alunos de CC até o ano de 2003, entendendo por aproveitamento o seu respectivo desempenho nas disciplinas já cursadas. O status que cada aluno pode obter para cada disciplina é: Aprovado, Reprovado ou Desistente. Entende-se por desistente aquele aluno que iniciou a disciplina frequentando e realizando as primeiras avaliações e depois a abandonou. Desta forma, para cada um dos alunos são listadas todas as disciplinas já cursadas. No caso do aluno que na primeira vez reprovou na disciplina X e posteriormente ser aprovado, a disciplina X será listada duas vezes, cada uma das ocorrências com o seu respectivo status.

Os dois conjuntos de regras descobertos totalizaram 456.063 regras (conjunto 1 - BSI), 62.700 regras (conjunto 1 - PD) e 1.271.715 regras (conjunto 2 - CC). A partir destes conjuntos foram selecionadas regras que estivessem relacionadas com o problema proposto para o experimento, a questão referente à alta desistência e reprovação nas disciplinas dos referidos cursos. Sendo assim, foram selecionadas regras que associassem disciplinas com desistência e/ou reprovação, totalizando 154.443 regras (conjunto 2 - BSI), 39.339 regras (conjunto 2 - PD) e 418.028 regras (conjunto 2 - CC).

Regra 1
LOGICA_DE_PROGRAMACAO_Aprovado INTRODUCAO_A_COMPUTACAO_Aprovado HABILIDADES_ACADEMICAS_Aprovado PROGRAMACAO_COMPUTADORES_I_Aprovado
→ ESTRUTURA_DE_DADOS_E_GRAFOS_Desistente
Regra 2
INTRODUCAO_A_COMPUTACAO_Aprovado LOGICA_MATEMATICA_Aprovado
→ ESTRUTURA_DE_DADOS_E_GRAFOS_Desistente

Figura 1 – Regras Descobertas

A figura 1 mostra algumas das regras descobertas que associam a desistência na disciplina de Estrutura de Dados a outras situações (conjunto 2).

A partir da figura 1 (regra 1) é possível perceber que uma das associações identificadas a partir dos dados é o fato de que o aluno mesmo tendo sido aprovado na disciplina Programação de Computadores I (PROGRAMACAODECOMPUTADORESI Aprovado), da primeira série, não garante que o mesmo seja aprovado na disciplina Estrutura de Dados e Grafos (ESTRUTURADEDADOSEGRAFOS Desistente), disponível na segunda série. Mesmo considerando-se a forte dependência do aprendizado da primeira.

Não apenas estas duas regras foram avaliadas pelos colegiados dos cursos, mas, sim, um conjunto de 45 regras (conjunto 3 – BSI, conjunto 3 – PD e conjunto 3 - CC). O critério para selecionar estas 45 regras, a partir do conjunto 2 de regras, é descrito a seguir.

Os conjuntos de regras (conjunto 2 – BSI, conjunto 2 – PD e conjunto 2 - CC) foram ranqueados cinco vezes, cada um dos ranqueamento considerou uma medida distinta de qualidade como critério classificador. A partir de cada um dos ranqueamentos foram selecionadas 9 regras: as três com os melhores valores, as três piores e finalmente as três medianas em relação à respectiva medida.

Os três conjuntos de 45 regras (conjunto 3 – BSI, conjunto 3 – PD e conjunto 3 - CC) foram oferecidos para avaliação a membros do colegiado dos respectivos cursos, sob o foco do problema da alta desistência e/ou reprovação. Aos membros do colegiado coube

selecionar entre as 45 regras apresentadas aquela mais interessante (agrega algo ainda não percebido pelo gestor) e a regra menos interessante (algo já conhecido pelo gestor).

A partir da análise do usuário foi identificada a correlação existente entre cada uma das medidas de qualidade e de grau de interesse em relação à medida de avaliação dos membros do colegiado.

Examinando as correlações existentes para o Curso de Bacharelado em Sistemas de Informação, foi possível perceber que a medida com a maior correlação em relação ao real interesse do colegiado foi a *Cosine*, chegando a 0.684 e a segunda maior com a medida Φ- Coeficiente, 0.634. Segundo (Callegari-Jacques & Sidia, 2003), um valor de correlação compreendido no intervalo entre 0.6 e 0.9 é considerado uma correlação forte. A situação ideal seria acima de 0.9 para a qual é considerada uma correlação muito forte, mas que não foi o caso para os experimentos realizados. Pode-se concluir que para este problema, o método de avaliação que mais se aproxima das expectativas do usuário são as medidas Φ- Coeficiente e *Cosine*. Então, uma destas duas medidas poderia ser escolhida como critério de ordenação do conjunto de regras a ser oferecido para o usuário. Inclusive vale destacar que existe uma forte correlação entre as medidas de interesse, como por exemplo, 0.984 entre a Φ- Coeficiente e Kappa, bem como de 0.991 entre a *Cosine* e *Kappa*.

Analisando as correlações existentes para o Curso de Tecnologia em Processamento de Dados, não foi possível identificar uma medida que mais se correlacionasse com a avaliação do colegiado. A maior correlação foi de -0.401 relacionada à medida de qualidade Generalidade. O mesmo ocorreu com o Curso de Ciência da Computação para o qual a maior correlação encontrada foi de -0.267 para a medida Suporte Mútuo. Neste caso, pode-se concluir que para este problema, em princípio não é possível indicar uma destas medidas como critério de ordenação do conjunto de regras.

Conclusões

A elevada quantidade de regras de associação, comumente gerada pelos algoritmos, motiva a pesquisa por alternativas de pós-processamento, capazes de analisar as regras geradas, como por exemplo, ranqueamento, contribuindo assim com a tarefa de análise efetuada pelo especialista no problema, tornando esse trabalho mais produtivo, ou mesmo viabilizando a análise.

Porém a literatura propõe várias medidas de qualidade e de interesse que nos permitam ordenar;

mas a questão que surge é: "qual delas utilizar em determinada situação?"

Este artigo discutiu e experimentou cinco medidas de qualidade e onze medidas de interesse, comparando os resultados obtidos a partir destas com os resultados obtidos a partir da avaliação do usuário. Analisando os resultados desta comparação foi possível identificar que:

- para o curso BSI, o Φ-Coeficiente e *Cosine* foram as medidas mais adequadas para ordenação do o conjunto de regras descoberto, tendo em vista a possuírem os maiores graus de associação entre as variáveis medida de qualidade/interesse versus avaliação do usuário;
- para o curso PD e CC nenhuma destas medidas seria potencialmente indicada para tal função.

Uma sugestão de trabalho futuro seria pesquisar e identificar se existe uma relação entre estes tipos de problemas e/ou domínios para os quais outras medidas serão mais indicadas.

Referências Bibliográficas

BERRY, M. J. A.; LINOFF, (1997).G. *Data Mining Techniques*: for marketing, sales and customer support. John Willey & Sons, Inc. USA.

BORGELT, C.. *Working Group Neural Networks and Fuzzy Systems, Departament of knowledge Processing and Language Engineering*. Otto-von-Guericke-University of Magdeburg, Alemanha. Disponível em: http://www-ics.cs.uni-magdeburg.de/iws.html Acesso em: 05 jun. 2004

CALLEGARI-JACQUES, S. M., (2003). *Bioestatística* – Princípios e Aplicações. Artemd Editora.

FAYYAD, U. M; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (1996). *Advances in Knowledge Discovery and Data Mining*. USA: American Association for Artificial Intelligence

FREITAS, A. A. (1998) *On objective measures of rule surprisingness*. Principles of Data Mining & Knowledge Discovery (Proc. 2nd European Symp., PKDD'98. Nantes, France, Sep. 1998). LNAI 1510. 1-9. Springer-Verlag.

SILBERSCHATZ, A.; TUZHILIN, A. (1996). *What makes patterns interesting in knowledge discovery systems*. IEEE Trans. Knowledge & Data Eng. 8(6).

TAN, P. N.; KUMAR V.; SRIVASTAVA J. (2002) *Selecting the Right Interestingness Measure for Association Patterns*. Canadá.

YAO, Y. Y.; ZHONG, N. (1999) *An Analysis of Quantitative Measures Associated with Rules*, Pacific-Asia Conference on Knowledge Discovery and Database.

ZANIN, E. (2002). *Ferramenta para pré-processamanto na descoberta de conhecimento em bases de dados*. Monografia para obtenção do grau de Bacharel de Ciência da Computação – UTP, Curitiba.