Avaliando a relação entre quão corretas e interessantes podem ser as regras de associação descobertas

Viviane Dal Molin de Souza (Analista de Sistemas)
Curso de Ciências da Computação - Universidade Tuiuti do Paraná

Deborah Ribeiro Carvalho (Doutora)
Curso de Ciências 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 tem 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. Nestes casos pode ser adotada uma fase de pós-processamento do conhecimento descoberto, a qual pode selecionar um sub-conjunto 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 cinco medidas de pós-processamento com o objetivo de avaliar se as medidas selecionadas sob a perspectiva da qualidade, são efetivamente interessantes para o usuário como elemento de apoio ao processo decisório.

Palavras-chave: data mining, descoberta de conhecimento, avaliação do conhecimento descoberto, medidas de qualidade, regras de associação.

Abstract

The great majority of the Data Mining algorithms presents the knowledge discovered in the form of a huge list of rules, from which the user must search and identify those that really possess quality and has something to add to the decision process. Very often this list is so great that makes impracticable the analysis work. In these cases a phase of post-processing of the discovered knowledge can be adopted, which can select a subgroup of the discovered rules, under the criterion of the quality, or either, of the degree of rightness of the rules. From this form the user would receive a set well reduced of rules to be evaluated, what he would facilitate its analysis. This article presents and argues five measures of post-processing with the objective to evaluate if the measures selected under the perspective of the quality, are effectively interesting for the user as element of support to the power to decide process.

Key words: data mining, discovery of knowledge, evaluation of the discovered knowledge, measured of quality, rules of association rules.

1 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 (1996), o processo KDD é composto de diversas etapas, a saber:

- Seleção de dados: prevê a coleta e seleção dos dados;
- Limpeza: prevê a analise 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*: esta etapa consiste em aplicar um algoritmo que efetivamente procura por padrões /

relações e regularidades em um determinado conjunto de dados;
- 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 decoberto 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.

Os algoritmos de *Data Mining* podem ser identificados em três tarefas principais, a saber: classificação, descoberta de regras de associação e *clustering*.

Neste artigo serão avaliadas as regras descobertas a partir da tarefa de associação sob o ponto de vista da qualidade em relação ao quanto o usuário entende que a regra seja interessante.

A preocupação das regras descobertas por algoritmos de *Data Mining* serem interessantes remonta do fato de que não adianta uma regra ser 100% correta, mas refletir situações já conhecidas pelo usuário, ou seja, não agregarem nada ao processo decisório. Por isso, quando se realiza um experimento *Data Mining*, como caminho alternativo para auxiliar no processo decisório é fundamental que o conhecimento descoberto (no caso deste artigo, as regras de associação), sejam corretas, compreensíveis e úteis / interessantes.

2 Descoberta de regras de associação

A tarefa de descoberta de regras de associação identifica afinidades entre itens de um subconjunto de dados e estas afinidades são expressa na forma de regras.

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 uma 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 e 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 (conjunto de itens) na forma $X \cap Y = \varnothing$.

Os algoritmos de associação fornecem ao usuário uma gama enorme de regras, e que nem sempre podem ser analisadas por causa da quantidade, assim se faz necessário um refinamento desta gama de regras para que o usuário possa conseguir efetuar a análise das mesmas. Com o método implementado neste trabalho o usuário tem pode selecionar as regras que realmente apresentam qualidade do ponto de vista da precisão preditiva, além de poder ordenar as regras por alguma medida de qualidade. Auxiliando assim a tomada de decisão.

3 Pós-processamento do conjunto de regras descobertas

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 descobertas pelos algoritmos de associação podem 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 (Silberschatz e Tuzhilin, 1996) (Freitas, 1998). 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 do domínio bastante diversos.

As medidas objetivas têm a desvantagem de serem uma estimativa indireta do quão surpreendentes serão as regras para o usuário. Porém 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 exclusivas, 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 e Zhong (1999).

3.1. Medidas Objetivas

O método adotado usa a analise 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.

Um universo finito consistente de objetos e cada um pode ser considerado um modelo de disposição de dados e é denotado por U. A partir de regras do tipo Se <E> então <H>, pode-se dizer que E e H são conceitos e que podem ser definidos usando certa linguagem. Busca-se a representação exata dos conceitos E e H. Assim, para o conceito E, m(E) denota o conjunto dos elementos de U que satisfaçam as condições expressadas por E. Da mesma maneira, o conjunto m(H) consiste nos elementos que satisfazem H. Desta forma m pode ser interpretado como uma função significativa que associa esse conceito com um subconjunto de U.

Usando o conjunto de cardinalidades obtem-se o seguinte quadro ou tabela de contingência representando as informações quantitativas sobre a regras E $\rightarrow$ H:

	H	H'	Total
E	$\|m(E) \cap m(H)\|$	$\|m(E) \cap m(H')\|$	$\|m(E)\|$
E´	$\|m(E') \cap m(H)\|$	$\|m(E') \cap m(H')\|$	$\|m(E')\|$
Total	$\|m(H)\|$	$\|m(H')\|$	$\|U\|$

A tabela de contingência pode ser reescrita da seguinte forma:

	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, *n* denota o número total de registros. A partir de uma tabela de contingência pode-se definir algumas medidas de qualidade básica.

A generalidade é definida pela expressão 1:

$$G(E) = a + b / n, \quad (1)$$

que indica o tamanho relativo do conceito E. Um conceito é mais geral se cobrir mais partes do universo. Se G(E) = a, então(100a)% dos objetos do universo satisfazem E. A quantidade pode ser vista como uma probabilidade de selecionar elementos ao acaso e satisfazer E. Também temos $0 \leq G(E) \leq 1$.

O suporte absoluto de H provido por E é definida pela expressão 2:

$$AS(H \mid E) = a / a + b. \quad (2)$$

A quantidade, $0 \leq AS(H \mid E) \leq 1$, mostra o grau em que E implica em H. Se $AS(H \mid E) = \alpha$, então (100a)% de objetos que satisfaçam E também satisfazem H. Isso pode ser visto como uma probabilidade condicional de um elemento escolhido ao acaso satisfazer H, dado que o elemento satisfaz E. Em termos teóricos, é o grau com que m(E) esta incluido em m(H). Claramente, $AS(H \mid E) = 1$, se e apenas se $m(E) \subseteq m(H)$.

A mudança de apoio de H provido por E é definido pela expressão 3:

$$CS(H \mid E) = an – (a + b)(a + c) / (a + b)n. \quad (3)$$

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

O suporte mutuo de E e H é definido pela expressão 4:

$$MS(E,H) = a / a + b + c \quad (4)$$

Pode-se interpretar o suporte mutuo $0 \leq MS(H \mid E) \leq 1$, 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 5:

$$IND(E,H) = an / (a + b)(a + c) \quad (5)$$

Isto mostra o grau de desvio de H na sub-população 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 $AS(H \mid E)$ e G(H). Quando E e H são provavelmente independentes, temos $CS(E \mid H) \geq 0$ e $IND(E,H) \geq 1$. Ainda mais, $CS(H \mid E) \geq 0$ se e apenas se $IND(E,H) \geq 1$, e $CS(H \mid E) \leq 0$ se e apenas se $IND(E,H) \leq 1$.

4 Experimentos realizados

Foram realizados experimentos com o objetivo de avaliar as medidas propostas por Yao e Zhong (1999). Os experimentos foram realizados sobre a base de dados de aproveitamento dos alunos da Universidade de Tuiuti do Paraná, para o Curso de Bacharelado em Ciência da Computação. As regras de associação foram extraídas a partir do algoritmo Apriori (Borgelt, 2004).

Essa base de dados se refere ao aproveitamento de 364 alunos 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 364 alunos são listadas todas as disciplinas já cursadas. No caso de alunos que na primeira vez reprovaram na disciplina X e posteriormente foram aporvados, para estes alunos a disciplina X será listada duas vezes, cada uma das ocorreências com o respectivo status.

O conjunto total (conjunto 1) de regras descobertas totalizou 1.271.715 regras. A partir deste conjunto foras selecionadas regras que estivessem relacionadas com o problema proposto para o experimento, a questão referente a alta desistência e reprovação nas disciplinas do referido curso. Após este processo de seleção, regras que associassem alguma disciplina com

desistência e/ou reprovação, foram obtidas 418.028 regras (conjunto 2).

A figura 1 mostra algumas das regras descobertas a partir do algoritmo apriori (conjunto 2), bem como os seus respectivos valores de qualidade, descritos na seção 3.1.

ESTRUTURADEDADOSEGRAFOS Desistente <- LOGICADEPROGRAMACAO Aprovado INTRODUCAOACOMPUTACAO Aprovado HABILIDADESACADEMICAS Aprovado PROGRAMACAODECOMPUTADORESI Aprovado
(G=0.233516,SA=0.082353,MA=-0.000065, SM=0.064815,I=0.999216)

ESTRUTURADEDADOSEGRAFOS Desistente <- INTRODUCAOA COMPUTACAO Aprovado LOGICAMATEMATICA Aprovado
(G=0.233516,SA=0.082353,MA=-0.000065, SM=0.064815, I=0.999216)

A partir da figura 1 é possível perceber que uma das associações identificadas a partir dos dados é o fato do aluno ser aprovado na disciplina Programação de Computadores I (PROGRAMACAODECOMPUTADORESI Aprovado), disciplina do primeira série, não garante que este aluno seja aprovado na disciplina Estrutura de Dados e Grafos (ESTRUTURADEDADOSEGRAFOS Desistente), disciplina da segunda série e que em princípio dependeria do bom desempenho daquela disciplina da primeira série.

Não apenas estas duas regras foram avaliadas pelo colegiado do curso, mas sim um conjunto de 45 regras (conjunto 3). O critério para selecionar estas 45 regras a partir do conjunto2 de regras é descrito a seguir.

O conjunto de regras (conjunto 2) foi ranqueado cinco vezes, cada um dos ranqueamento considerou uma medida distinta de qualidade. A partir de cada um dos ranqueamentos foram selecionadas 9 regras: as três que a apresentassem os melhores, as três que apresentassem os piores e as três que ocupassem a posição mediana no ranqueamento em relação a respectiva medida.

O conjunto de 45 regras (conjunto 3) foi oferecido a dois membros do colegiado do Curso de Bacharelado em Ciência da Computação para avaliação, sob o foco do problema da alta desistência e reprovação. Aos usuários (membros do colegiado) coube identificar cada uma das 45 regras como sendo: interessante (agrega algo ainda não percebido pelo gestor), ± interessante (agrega pouco ao conhecimento prévio do gestor) e nada interessante (algo já conhecido pelo gestor).

A partir da análise do usuário foi identificada a correlação existente entre cada uma das medidas de qualidade a medida de avaliação do usuário. Foi possível perceber que a medida de qualidade com maior correlação em relação a medida gerada pelo usuário foi o suporte mútuo, chegando a 0.621. Segundo (Callegari-Jacques, 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.

Os demais valores de correlação entre a avaliação do usuário e as medidas de qualidade são: Generalidade = 0.282, Suporte Absoluto = - 0.362, Mudança de Apoio = -0.375, Grau de Independência = -0.464.

5 Conclusão

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

Porém muitas vezes a literatura propõem várias medidas de qualidade que nos permitam ordenar; e a questão que surge é: "qual delas utilizar em determinada situação?"

Este artigo discutiu e experimentou cinco medidas de qualidade, comparando os resultados obtidos a partir destas com os resultados obtidos a partir da avaliação do usuário. Analisando os resulatdos desta comparação foi possível identificar que o Suporte Mútuo foi a medida mais adequada para ordenar o conjunto de regras descoberto, tendo em vista a ser aquela com maior grau de associação entre as variáveis medida de qualidade versus avaliação do usuário.

Uma sugestão de trabalho futuro seria a comparação do método implementado neste trabalho com outros métodos que possuem como objetivo principal o pós-processamento do conhecimento descoberto na fase de *Data Mining*, objetivando identificar não apenas as regras com maior qualidade, mas sim também regras com maior grau de interesse.

Referências bibliográficas

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

BORGELT, C. (2004) *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.

CALLEGARI-JACQUES, Sidia 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, Alex Alves. (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, Springer-Verlag, 1-9.

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

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

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