Descoberta de Conhecimento Interessante Utilizando Medidas Objetivas

Abstract. The process of knowledge acquisition on databases normally shows to the user a large set of discovered rules. It's a hard task for the user, thus, to search and identify those rules which are really interesting and the ones that bring something more to the knowledge already existing about the domain. To help this analysis process, there can be used objective measures of post-processing that allow people to identify interesting rules according to various aspects. In this article, three of these objective measures were tested over the same lists of discovered rules intending to verify the potential on using some measures on data mining processes.

Resumo. O processo de extração de conhecimento em base de dados, normalmente, apresenta ao usuário um extenso conjunto de regras descobertas. Cabe ao usuário, portanto, pesquisar e identificar aquelas regras que são realmente interessantes e que acrescentam algo ao conhecimento já existente sobre o domínio. Para auxiliar esta an*álise, pode-se empregar medidas objetivas de* p*ós-processamento que permitam identificar regras interessantes, segundo aspectos variados. Neste artigo, três medidas objetivas foram testadas sobre as mesmas listas de regras descobertas com o objetivo* de verificar o potencial de uso dessas medidas em processos de data mining.

1. Introdução

A cada ano, companhias acumulam mais e mais informações em seus bancos de dados. Como conseqüência, estes bancos de dados passam a conter grande quantidade de dados sobre vários dos procedimentos dessas companhias. Para setores estratégicos da maioria das organizações o que importa realmente não são os dados em si, mas sim o conhecimento que pode ser extraído destas bases. Através deste conhecimento pode-se tentar detectar tendências e características disfarçadas a fim de melhorar os procedimentos da organização e permitir que a mesma reaja rapidamente a um evento que ainda pode estar para acontecer [Berry e Linoff, 1997].

Assim como outras ciências, a Ciência da Computação pesquisa e estuda formas alternativas de realizar a tarefa de extrair conhecimento a partir de base de dados. Neste caso, parte-se do princípio que os dados contidos em uma base de dados possuem relações escondidas (implícitas), não perceptíveis ou não vislumbradas tão facilmente, que podem vir a ser úteis ao ser humano num processo de tomada de decisão. A área da computação destinada a esta pesquisa denomina-se Descoberta de Conhecimento em

Bases de Dados (KDD - *Knowledge* Discovery in Database). Segundo Fayyad et. al. (1996), o processo KDD é composto de diversas etapas, a saber:

- Seleção dos dados: esta etapa compreende a seleção dos dados necessários para o processamento no KDD.
- Limpeza: nesta etapa deve-se analisar os dados coletados verificando a existência de ruídos, e caso existam, devem ser tratados de modo a eliminá-los. Nesta etapa deve-se também estabelecer as estratégias para resolver o problema de ausência de dados.
- Transformação: tem por intuito armazenar os dados adequadamente e identificar a necessidade de dados adicionais.
- *Data* Mining: consiste de um conjunto de algoritmos com o objetivo de encontrar uma descrição de padrões e regularidades em um dado conjunto de dados.
- Pós-Processamento: consiste em um conjunto de medidas objetivas ou subjetivas com o intuito de podar ou "raquear" os padrões descobertos.
- Interpretação e Avaliação: tem como objetivo interpretar e verificar a validade do conhecimento descoberto.

Existem diversas formas de representação do conhecimento descoberto, entre elas as regras de associação. Em geral, o conjunto de regras descobertas pelos algoritmos de *data* mining é muito extenso, com centenas ou milhares de regras, dificultando a avaliação do conhecimento descoberto, e em alguns casos, inviabilizando todo o processo de KDD. Além disso, em alguns casos pode ocorrer a descoberta de conhecimento muito óbvio ou irrelevante e que pouco acrescenta ao conhecimento já existente sobre o domínio [Padmanabhan and Tuzhilin, 1999].

Desta forma, surge a necessidade de um refinamento, onde as regras redundantes sejam eliminadas e as mais interessantes sejam destacadas. Este refinamento no conjunto de regras descoberto pode ser realizado utilizando-se uma medida de pós-processamento (subjetiva ou objetiva) que permitirá que o trabalho de interpretação realizado pelo analista seja facilitado e mais produtivo.

O objetivo deste artigo é testar medidas objetivas de pós-processamento avaliando a qualidade do conhecimento adquirido e identificando o conhecimento considerado interessante.

2. Descoberta de Regras de Associação

A descoberta de regras de associação foi introduzida por Agrawal (1993) para gerar relacionamentos ou padrões freqüentes entre conjuntos de dados. As regras de associação revelam quais são as relações relevantes dos objetos no conjunto de dados. Por exemplo, um conjunto de sintomas freqüentemente ocorre juntamente com um outro conjunto de sintomas. [Carvalho, 1999]

Dessa forma, as regras de associação objetivam encontrar relacionamentos entre conjuntos de dados. Dado um conjunto de transações, onde cada transação é um conjunto de dados, uma regra de associação é uma expressão do tipo $X \rightarrow Y$, (Se X

então Y, sendo X o antecedente e Y o conseqüente), onde X e Y são conjuntos de itens, $X \cap Y = \varnothing$.

Em geral, os algoritmos que descobrem regras de associação trabalham com duas medidas para avaliar cada uma das regras: o suporte e a confiança. Suporte e confiança servem para limitar a quantidade de regras extraídas e são definidos antes da extração.

Suporte é o número de transações contendo o conjunto de itens, dividido pelo número de total de transações. Dessa forma, esta medida quantifica a incidência da regra no conjunto de dados, ou seja, indica a freqüência com que X e Y ocorrem juntos no conjunto de dados. O suporte é equivalente a probabilidade P(X,Y), isto é, a probabilidade de que X e Y ocorram simultaneamente [Melanda e Rezende, 2002]. O suporte de uma regra é calculado pela equação 1 abaixo:

$$Suporte(Sup) = \frac{Nr.de\_registros\_com\_XeY}{Nr\_total\_de\_registros} \quad (1)$$

A medida de confiança indica a freqüência com que X e Y ocorrem juntos em relação ao número total de registros em que X ocorre, é equivalente à probabilidade condicional P(Y/X), ou seja, representa a probabilidade de ocorrência de Y, dado que X ocorra [Melanda e Rezende, 2002]. A confiança de uma regra é calculada pela equação 2 abaixo:

$$Confiança(Conf) = \frac{Nr.de\_registros\_com\_XeY}{Nr\_de\_registros\_com\_X} \quad (2)$$

A tarefa de descobrir regras de associação consiste em extrair do banco de dados todas as regras que possuam os fatores de suporte e confiança maiores ou iguais a um suporte e confiança especificado pelo usuário. A definição desses fatores serve para introduzir uma medida capaz de distinguir associações interessantes, dado que as regras geradas em forma binária elevam em muito o espaço de busca de qualquer algoritmo de *Data* Mining.

3. Pós-processamento

A transformação de dados em informações úteis, ou seja, em conhecimento, tem ocorrido tradicionalmente através de análises e interpretações realizadas a partir de procedimentos manuais.

Em geral, os dados são analisados por um ou mais especialistas familiarizados com a natureza dos dados, que comparam gráficos e relatórios, cruzam informações, utilizando-se do conhecimento prévio que possuem sobre o domínio da aplicação. Dessa forma, os especialistas atuam como uma espécie de interface entre os dados e os possíveis usuários das informações. Este procedimento, usualmente, é lento, subjetivo e custoso, tornando-se impraticável quando se trabalha com grande volume de dados [Melanda e Rezende, 2002].

Outros procedimentos que podem ser utilizados para transformar dados em conhecimento, são os procedimentos automáticos, como por exemplo os algoritmos de regras de associação. Embora as regras de associação sejam padrões valiosos por

oferecerem uma percepção útil da dependência que existe entre atributos da base de dado, elas também possuem dois inconvenientes [Tan e Kumar, 2000]:

- muitas regras são geradas (problema da quantidade de regras)
- nem todas as regras são interessantes (problema da qualidade da regra)

Como forma de superar essa problemática, podem ser utilizadas técnicas de pós-processamento. Desta forma o número de regras descobertas pode ser reduzido, facilitando assim a avaliação.

Existe um grande número de propostas na literatura para "minerar" o conhecimento descoberto. Em geral as propostas se enquadram em duas categorias básicas: métodos subjetivos e objetivos. No método subjetivo, é preciso que o usuário estabeleça previamente o conhecimento ou crenças, a partir do qual o sistema irá minerar o conjunto original de padrões descoberto pelo algoritmo de *Data* Mining, buscando por padrões que sejam interessantes ao usuário. Por outro lado, o método objetivo não necessita que um conhecimento prévio seja estabelecido. Pode-se dizer que o método objetivo é *data-driven* e o subjetivo é *user-driven* (Freitas, 1999).

O foco deste artigo são as medidas objetivas, especificamente a medida de taxa de acerto, a medida de interesse (IS) proposta por Tan e Kumar (2000) e outra proposta por Liu et al (1999).

3.1. Taxa de Acerto (TA)

O conhecimento descoberto pelo processo KDD, durante a etapa de *Data* Mining, tem por objetivo identificar em dados padrões válidos, novos, potencialmente úteis e compreensíveis [Fayyad et al., 1996].

Na sua grande maioria, os algoritmos de *Data* Mining produzem, como parte dos resultados, informações de natureza estatística que permitem ao usuário identificar o quão correto e confiável é o conhecimento descoberto. Por exemplo, dada a seguinte regra:

R: SE X ENTÃO Y

Sob o ponto de vista estatístico, a regra acima pode ser descrita pela tabela de contingência [Monard e Baranauskas, 2003] (Tabela 1):

Tabela 1. Tabela de Contingência para o cálculo da Taxa de Acerto

	Y	Não Y	
Exemplos cobertos pela regra	RY	rY'	R
Exemplos não cobertos pela regra	r'Y	r'Y'	r'
	C	c'	N

onde:

- rY é o número de exemplos cobertos pela regra *R* e pertencentes a classe Y;
- rY' é o número de exemplos cobertos pela regra *R*, mas não pertencentes a classe Y;
- r = rY + rY' é o número de exemplos cobertos por *R*;

- c = rc + r'c é o número de exemplos de treinamento que possuem a classe Y;
- N = c + c' = r + r' é o número total de exemplos (registros) de treinamento.

A Taxa de acerto é uma medida calculada utilizando a tabela de contingência (tabela 1) e tem por objetivo atribuir um valor de quão correta é uma regra descoberta. A equação 3 é utilizada para o cálculo do valor da taxa de acerto de uma regra:

$$TaxadeAcerto(TA) = (rY + r'Y)/N \quad (3)$$

Muitas vezes, apenas a medida de taxa de acerto, calculada pela tabela de contingência, não é suficiente para descobrir regras interessantes ou ainda regras de fácil compreensão. Por este motivo, métricas que avaliem o grau de interesse e de compreensibilidade podem ser computadas em uma fase de pós-processamento, como uma forma de avaliação adicional da qualidade do conhecimento descoberto, complementando (e *não* substituindo) medidas estatísticas sobre o grau de correção daquele conhecimento.

3.2. Medida proposta por Tan e Kumar (IS)

Segundo Tan e Kumar (2000), a solução para o problema da qualidade da regra está em especificações de uma medida de interesse para representar a novidade, a utilidade ou o significado de um padrão. Requisitando as regras descobertas de acordo com seu grau de interesse, as regras selecionadas podem ser apresentadas a um analista. Algumas destas medidas são aplicáveis aos conjuntos de itens como também às regras.

O fator do interesse é uma medida amplamente usada para os padrões de associação. Esta métrica é definida para ser a razão entre a probabilidade comum de duas variáveis com relação às suas probabilidades previstas sob a suposição de independência (Tan e Kumar, 2000). O fator de interesse pode ser calculado pela fórmula 4 abaixo:

$$I(A,B) = \frac{P(A,B)}{P(A)P(B)} \quad (4)$$

Este artigo [Tan e Kumar, 2000], sugere que uma boa medida de interesse, derivada da correlação estatística para regras que tenham baixo suporte (menor que 30%) e alto fator de interesse, pode ser dado pela fórmula 5 abaixo:

$$IS = \sqrt{I * P(A,B)} = \sqrt{\frac{P(A,B)P(A,B)}{P(A)P(B)}} \quad (5)$$

A medida acima é um produto de dois quantificadores importantes: fator de interesse e o suporte e leva em consideração tanto os aspectos do interesse como o do suporte de um padrão.

3.3. Medida proposta por Liu et al. (DS)

O método adotado por Liu et al. (1999) usa correlações estatísticas como base para encontrar as regras que representam as relações fundamentais do domínio, podendo assim distinguir entre regras insignificantes e significantes. Dentro do conjunto de

regras significantes, o método permite distinguir um subconjunto capaz de resumir o conteúdo apontado pelas regras significantes.

A figura 1 mostra o fluxo da técnica proposta, que consiste em dois passos: poda e resumo. O método é capaz de podar as regras insignificantes e encontrar um subconjunto especial de regras que representam um resumo das regras não podadas, assim consideradas interessantes. Esse subconjunto de regras é chamado de conjunto de regras *DS – direction* setting rules.

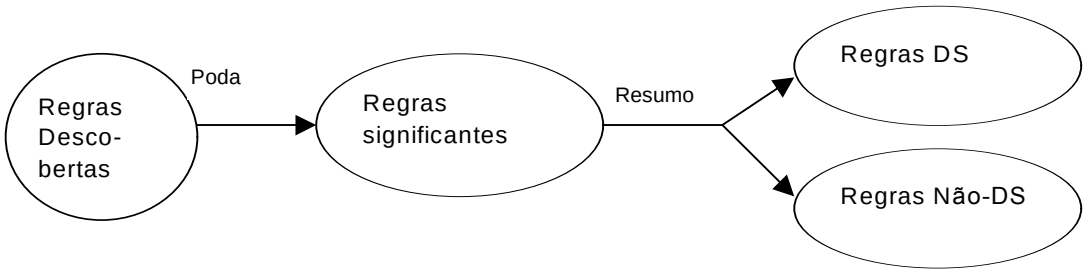

Figura 1. O método proposto por Liu et. al. (1999).

As regras DS representam as relações essenciais, ou estrutura do domínio dos dados. As regras não-DS, apesar de serem consideradas interessantes, apresentam apenas informações adicionais.

O método é focado na mineração de regras de associação a partir de uma tabela relacional, que consiste em um conjunto de registros descritos por um número de atributos.

O método não usa confiança mínima, pois ela não reflete a base das relações do domínio representado pelos dados [Bayardo et al., 1999]. Ao invés dele, o método usa correlações estatísticas como base para encontrar as regras que representam as relações fundamentais do domínio.

Regras que são muito específicas (com muitas condições) possuem tendência de sobrecarregar os dados e tem pouco poder preditivo. Embora regras de associação não sejam normalmente usadas para predição, regras que apenas capturam as irregularidades e características especiais dos dados não possuem valor, assim devem ser podadas. Um exemplo desse tipo de regras é mostrado no quadro 1.

R1: Trabalho = sim > empréstimo = aprovado [sup = 60%, conf = 90%]
R2: Trabalho = sim, histórico de crédito = bom > empréstimo = aprovado [sup = 40%, conf = 91%]

Quadro 1. Exemplo de Regra Redundante a ser podada

Pode-se perceber que R2 é insignificante uma vez que ela apenas fornece informações extras em relação a R1. A maior superioridade de R2 com relação à confiança é provavelmente devido à chance de correlação verdadeira. R2 dessa forma deve ser podada, pois R1 é mais geral e simples. Regras gerais e simples são preferidas.

O método avalia a significância de uma regra usando o teste Chi-quadrado (x^2) para correlação a partir de estatísticas.

A poda pode reduzir o número de regras substancialmente. Entretanto, o número de regras deixadas pode ser ainda grande. Este passo encontra um subconjunto chamado de *direction* setting rules (ou regras DS), que resumem as regras não-podadas. Essencialmente, as regras DS são regras de associações significantes que mostram a direção seguida pelas regras não-DS. A direção de uma regra é o tipo de correlação que ela possui. Por exemplo, correlação positiva, correlação negativa ou independência, que também é verificada usando o teste x^2. O quadro 2 abaixo demonstra isso.

R3: Trabalho = sim > empréstimo = aprovado [sup = 40%, conf = 70%]
R4: Casa própria = sim > empréstimo = aprovado [sup = 30%, conf = 75%]

Quadro 2. Exemplo de Regras DS

O teste x^2 mostra que ter um trabalho é positivamente correlacionado com a concessão do empréstimo, e possuir casa própria é também positivamente correlacionado com a concessão do empréstimo. Então a associação abaixo não é tão surpreendente:

R5: Trabalho = sim, casa própria = sim > empréstimo = aprovado [sup = 20%, conf = 90%]

Quadro 3. Associação obtida das regras R3 e R4

A regra intuitivamente segue R3 e R4. Pode-se usar R3 e R4 para prover um resumo das três regras. R3 e R4 são regras DS, pois mostram a direção (correlação positiva) que é seguida pela regra três. Nos conjunto de dados da vida real, existe um grande número de regras iguais à R5.

Do exemplo, nota-se que as regras DS mostram as relações essenciais do domínio dos dados. As regras não-DS não são surpreendentes quando já se tem conhecimento das regras DS. Entretanto, isto de forma alguma, diz que as regras não-DS não sejam interessantes. As regras não-DS podem prover detalhes extras sobre o domínio. Por exemplo, a regra não-DS R5 no exemplo acima, possui uma maior confiança, que pode ser do interesse do usuário. Usar as regras DS para formar um resumo é semelhante ao resumo de um texto qualquer. Do resumo, pode-se descobrir a essência do texto. Se os detalhes de algum ponto específico do texto são interessantes, o resumo direciona a ler o texto em si. Da mesma forma, as regras DS mostram a essência do domínio e direcionam o usuário as regras não-DS. As regras não-DS são basicamente combinações de regras DS.

4. Experimentação e Resultados

Para a realização dos experimentos foram utilizadas duas bases de dados (Senco e Duarte, 2003) e (Ferreira e Petrin, 2003):

- Base de dados House-votes, obtida a partir do site da University of Califórnia Irvine – UCI. Esta base possui 232 registros, compostos por 17 atributos categóricos e nenhum valor de atributo faltando.
- Base de compras de livros mostrada na Tabela 2.

As regras de associação, para ambas as bases, foram geradas a partir do algoritmo Apriori, com o valor determinado de 10 para o Suporte e 10 para Confiança.

Tabela 2. Cadastro de clientes *versus* compra de livros

Sexo	Pais	Idade	Meta Compra
Masculino	França	Jovem	S
Masculino	Inglaterra	Jovem	S
Feminino	França	Jovem	S
Feminino	Inglaterra	Adulto	S
Feminino	França	Adulto	N
Masculino	Alemanha	Jovem	N
Masculino	Alemanha	Jovem	N
Feminino	Alemanha	Jovem	N
Feminino	França	Adulto	N
Masculino	França	Adulto	N

Dado que o algoritmo proposto por LIU et. al. (1999), tem como objetivo identificar as regras podadas e não podadas, e dentre as não podadas quais são as regras DS, foram obtidos os seguintes resultados a partir das duas bases testadas (Ferreira e Petrin, 2003):

- base compra de livros foram descobertas 170 regras de associação, das quais 140 (82.35%) foram inicialmente podadas. Das apenas 30 regras restantes, 27 foram consideradas DS e 3 não_DS;
- base house-votes foram geradas 2424 regras de associação das quais 1344 (55.45%) foram podadas. Das 1080 regras não podadas, 150 foram consideradas DS e 930 não-DS.

A partir da tabela 3 é possível comparar o grau de interesse atribuído por ambos os métodos para 25 regras consideradas DS. Também foi possível fazer um a análise sobre o valor da Taxa de Acerto obtido para essas mesmas regras.

Pode-se observar que as medidas apontam para um mesmo objetivo: "descoberta de conhecimento novo, correto e interessante", e essas medidas possuem uma correlação entre si, não é alta e sim de 0,48.

Também foi possível observar uma forte relação entre as medidas de interesse e a taxa de acerto de cada uma das regras, ou seja, a maioria das regras que obtiveram valores de IS altos apresentou valores de TA alto.

No entanto, nem sempre uma regra com alta taxa de acerto é a mais interessante e nem sempre uma regra com alto valor de IS é a mais correta. Observa-se também que

regras com baixo suporte tiverem baixo valor de IS, isso ocorre devido ao fato de que as regras não dependem apenas do suporte, mas também do valor obtido pelo fator de interesse.

Tabela 3. Comparação entre as duas medidas de interesse (Senco e Duarte, 2003) e (Ferreira e Petrin, 2003).

Regras	DS	IS	TA
sim <- Inglaterra (20.0%, 100.0%)	$X^2 = 8.0000$	IS = 7.07	80
Inglaterra <- sim (20.0%, 50.0%)	$X^2 = 7.0000$	IS = 7.07	80
adulto <- Inglaterra (10.0%, 50.0%)	$X^2 = 6.8333$	IS = 3.54	60
Inglaterra <- adulto (10.0%, 25.0%)	$X^2 = 5.2500$	IS = 3.54	60
Masculino <- Alemanha (20.0%, 66%)	$X^2 = 6.0000$	IS = 5.16	60
Alemanha <- Masculino (20.0%, 40.0%)	$X^2 = 5.2381$	IS = 5.16	60
jovem <- Alemanha (30.0%, 100.0%)	$X^2 = 7.0000$	IS = 7.07	70
Alemanha <- jovem (30.0%, 50.0%)	$X^2 = 6.5714$	IS = 7.07	70
nao <- Alemanha (30.0%, 100.0%)	$X^2 = 7.0000$	IS = 7.07	70
Alemanha <- nao (30.0%, 50.0%)	$X^2 = 6.5714$	IS = 7.07	70
jovem <- sim (30.0%, 75.0%)	$X^2 = 5.5000$	IS = 6.12	60
sim <- jovem (30.0%, 50.0%)	$X^2 = 5.5000$	IS = 6.12	60
Franca <- adulto (30.0%, 75.0%)	$X^2 = 6.0000$	IS = 6.71	70
adulto <- Franca (30.0%, 60.0%)	$X^2 = 5.8333$	IS = 6.71	70
Feminino <- adulto (30.0%, 75.0%)	$X^2 = 6.0000$	IS = 6.71	70
adulto <- Feminino (30.0%, 60.0%)	$X^2 = 5.8333$	IS = 6.71	70
nao <- adulto (30.0%, 75.0%)	$X^2 = 5.5000$	IS = 6.12	60
adulto <- nao (30.0%, 50.0%)	$X^2 = 5.5000$	IS = 6.12	60
Feminino <- Franca (30.0%, 60.0%)	$X^2 = 5.2000$	IS = 6.00	60
Franca <- Feminino (30.0%, 60.0%)	$X^2 = 5.2000$	IS = 6.00	60
Masculino <- Inglaterra jovem (10.0%, 100.0%)	$X^2 = 4.5833$	IS = 4.47	60
jovem <- sim Franca (20.0%, 100.0%)	$X^2 = 4.2500$	IS = 5.77	60
sim <- Franca jovem (20.0%, 100.0%)	$X^2 = 6.3333$	IS = 7.07	80
Franca <- adulto Masculino (10.0%, 100.0%)	$X^2 = 5.3333$	IS = 4.47	60
Nao <- adulto Masculino (10.0%, 100.0%)	$X^2 = 5.1667$	IS = 4.08	50

Por este motivo, para avaliar o grau de interesse e o quão correta é uma regra, deve ser aplicada a etapa de pós-processamento, como uma forma de avaliação adicional

da qualidade e interesse do conhecimento descoberto, complementando e não substituindo outras medidas estatísticas.

5. Conclusão e Trabalhos Futuros

A elevada quantidade de regras de associação descoberta por algoritmos de *Data* Mining motivam o desenvolvimento de algoritmos de pós-processamento, com abordagem objetiva, capazes de analisar as regras geradas, eliminando as regras redundantes, e assim contribuindo com a análise do conhecimento gerado, tornando esse trabalho mais produtivo, ou mesmo viabilizando a análise em muitos casos.

Assim, foram implementados três métodos que pós-processa o conjunto de regras descoberto. Pode-se observar, que a maioria das regras que obtiveram valores de interesse alto – seja através do IS ou através do DS – apresentaram valores de TA alto. No entanto, nem sempre uma regra com alta taxa de acerto é a mais interessante e nem sempre uma regra com alto valor de interesse é a mais correta.

Uma outra importante observação, é que o método DS possui uma grande dependência da quantidade de condições existentes nas regras analisadas, onde regras com poucas condições, principalmente regras com uma única condição, possuem uma grande tendência de serem classificadas como regras DS; regras com duas condições demonstram uma tendência de serem classificadas como regras não-DS; e regras com três ou mais condições denotam uma grande tendência de serem podadas. Isto reduz, consideravelmente, o número de regras a serem analisadas e nos permite aplicar outros métodos apenas sobre o conjunto de regras DS e não-DS.

Dessa forma, constata-se que a implementação de mais de um método de pós-processamento torna o processo de descoberta de conhecimento mais robusto. Isto ocorre, uma vez que o uma medida complementa a outra e nos permite analisar melhor os conhecimento descoberto.

Como trabalho futuro, sugere-se a comparação destes métodos testados com outros objetivando identificar as vantagens, desvantagens e possíveis otimizações de código, neste caso, visando um ganho de performance.

6. References

Agrawal, R.; Imielinski, T. (1993) "Mining associations between sets of items in massive databases". In: International Conference on Management of Data, 207–216.

Bayardo, R.; Agrawal, R.; Gunopulos, D. (1999) "Constraint-based rule mining in large, dense databases", In: ICDE-99.

Berry, M. J. A., Linoff, G. (1997) "Data Mining Techniques: for marketing, sales, and customer support", John Wiley & Sons, Inc., USA.

Carvalho, D. R. (1999) "Data Mining Através de Introdução de Regras e Algoritmos Genéticos". Dissertação para obtenção do grau de Mestre – PUCPR, Curitiba.

Fayyad, U. M; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R.(1996) "Advances in Knowledge Discovery and Data Mining". USA: American Association for Artificial Intelligence. 1996.

Ferreira, C. A.; Petrin, C. R. (2003) "Data Mining – Descoberta de Conhecimento Interessante". Monografia para obtenção do grau de Cientista da Computação, Universidade Tuiuti do Paraná.

Freitas, A.. (1999) "On Rule Interestingness Measures". Knowledge – Based Systems Journal 12 (5-6), 309-315.

Liu, B.; Hsu, W.; Ma, Y. (1999) "Pruning and Summarizing the Discovered Associations", In: KDD-99.

Melanda, E. A.; Rezende, S. O. (2002) "Medidas Objetivas para Análise e Interpretação de Regras de Associação". In: I Workshop de Teses e Dissertações do SBIA'02, Recife/Porto de Galinhas, PE, 2002.

Monard, M. C.; Baranauskas, J.A. (2003) "Indução de regras e Árvores de Decisão". In Sistemas Inteligentes. Rezende, S. O. Editora Manole Ltda. 115-140.

Padmanabhan B.; Tuzhilin A. (1999) "Unexpectedness as a measure of interestingness in knowkdge discovery". Decision Support Systems, (27):303–318.

Senco, S. C.; Duarte, J. V. (2003) "Pós-processamento de Regras de Associação". Monografia para obtenção do grau de Bacharel em Sistemas de Informação, Universidade Tuiuti do Paraná.

Tan, P., Kumar, V. (2000) "Interestingess Measures for Association Patterns: A Perspective". In: Workshop on Postprocessing in Machine Learning and Data Mining - KDD. Boston.