Avaliação do conhecimento descoberto em *Data Mining*

Deborah Ribeiro Carvalho (Mestre)
Curso de Ciência da Computação - Universidade Tuiuti do Paraná

Resumo

O produto gerado pelas técnicas de *Data Mining* tem o compromisso de atender às características desejáveis do conhecimento descoberto, quais sejam: que o mesmo seja consistente, compreensível e útil/surpreendente ao usuário. A grande maioria desses algoritmos apresenta modelagens que avaliam o quesito consistência, entretanto, em muitos casos, apenas esta avaliação é insuficiente para o usuário. Este artigo apresenta e descreve algumas medidas que podem ser computadas numa fase de pós-processamento, permitindo assim complementar a avaliação do resultado produzido pelos algoritmos de *Data Mining*, dessa forma facilitando a apropriação do mesmo.

Palavras-chave: *data mining*, pós-processamento, avaliação do conhecimento

Abstract

The *Data Mining* results must comply with the desirable features of the knowledge base: it must be consistent, comprehensible and useful to the user. The vast majority of those algorithms present models that validate whether the results are consistent, but in many cases this is not enough for the user. This article presents and describes measurements that can be calculated in post-processing, allowing the user to fully evaluate the *Data Mining* results and making it easier its use.

Key-words: *data mining*, post-processing, knowledge evaluation

1 Introdução

O processo de descoberta de conhecimento em bases de dados KDD – (*Knowledge Discovery in Databases*) tem como principal objetivo a extração de conhecimento que seja útil a partir de bases de dados. É desejável que este conhecimento atenda a algumas características, tais como: seja tão correto quanto possível, compreensível e surpreendente para o usuário.

Neste processo como um todo, estão envolvidas várias etapas que vão desde a seleção da(s) base(s) de dados sobre a(s) qual(is) será realizado o processamento até a disponibilização do conhecimento descoberto para o usuário. Pode-se dizer que essas etapas fazem parte de três grandes grupos: pré-processamento, aplicação de um algoritmo de *Data Mining* e pós-processamento. 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, na tarefa de classificação considerando a regra

R: se Condição então a classe é *C*

Sob o ponto de vista estatístico, a regra R pode ser descrita pela tabela de contingência Tabela 1.

TABLE 1. VARIABLES TO BE CONSIDERED ON THE EVALUATION OF INTERACTION TECHNIQUES.

	classe *C*	não classe *C*	
Exemplos cobertos *R*	rc	rc'	r
Exemplos não cobertos R	r'c	r'c'	r'
	c	c'	N

Onde rc é o número de exemplos cobertos pela regra *R* e pertencentes à classe *C*;

rc' é o número de exemplos cobertos pela regra *R*, mas não pertencentes à classe *C*

r = rc + rc' é o número de exemplos cobertos por *R*;

c = rc + r'c é o número de exemplos de treinamento da classe *C*;

N = c + c' = r + r' é o número de todos os exemplos de treinamento.

Usando elementos da tabela de contingência é possível definir algumas medidas de qualidade da regra *R* [Monard e Baranauskas 2003]:

Consistência (R) = rc / r
Suporte (R) = rc / N
Sensitividade (R) = rc / r
Especificidade (R) = r'c' / r'
Precisão (R) = (rc + r'c') / n

Porém, muitas vezes essas medidas não são suficientes. Mesmo que o conhecimento descoberto seja altamente correto do ponto de vista estatístico, ele pode não ser de fácil compreensão. Por exemplo, o conjunto de regras descobertas pode ser grande demais para ser analisado, conter muita redundância, etc. 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.

Esta questão sobre pós-processar o conhecimento descoberto tem sido tratada por diversos autores na literatura e este artigo objetiva apresentar e discutir alguns destes métodos propostos.

2 Identificação dos métodos

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).

Neste artigo, são apresentados e discutidos métodos de ambas as naturezas, possibilitando ao leitor identificar as vantagens e desvantagens auferidas pela adoção de cada uma das métricas.

2.1 Métodos subjetivos

Neste critério, o interesse em uma regra depende do usuário, insto é, do conhecimento que o mesmo tem do domínio de aplicação. Porém existem algumas dificuldades inerentes a este conjunto de métodos, como, por exemplo, não é fácil identificar regras relevantes a partir de um grande conjunto de regras descobertas, uma regra pode ser relevante para um usuário e ser considerada inútil para outro, etc. Dessa forma, o interesse em uma regra pode ser considerado essencialmente subjetivo uma vez que depende dos conceitos atuais que o usuário tem a respeito do domínio, e também de seus interesses.

A seguir são descritos alguns métodos que adotam o critério subjetivo para avaliação do conhecimento descoberto.

Gebhardt (1991) discute a questão na qual em casos de aprendizado, os procedimentos de generalização devem encontrar um conjunto de tamanho considerável de generalizações parciais onde cada uma delas cubra uma parte dos exemplos de treinamento e em geral e poucos exemplos negativos. No entanto, identificar este tamanho de generalizações parciais não é trivial. Este artigo propõe um procedimento para selecionar a partir de um conjunto de generalizações um subconjunto no qual qualquer generalização que seja suficientemente similar e que ao mesmo tempo seja inferior a outra seja eliminada. A base deste procedimento é uma medida de evidência e de afinidade sobre qualquer par de generalizações. A evidência $V(G)$ de uma generalização é uma medida de sua importância, significância, surpresa; ela está baseada na proeminência de sua extensão, mas deve levar em conta alguma propriedade como, por exemplo, a complexidade, preferências do usuário, etc. A afinidade $S(G_j, G_k)$ entre duas generalizações é primariamente uma medida de sobreposição de duas extensões, que

também deve considerar algumas propriedades de suas expressões ou mesmo preferências do usuário. $S = 1$ indica uma similaridade forte.

A generalização G_j será suprimida por G_k se $V(G_j) < S(G_j, G_k) * V(G_k)$. A condição $S \pounds 1$ garante que duas generalizações não sejam mutualmente suprimidas.

Matheus *et al.* (1994) descrevem um experimento que analisa dados de seguro de saúde usando o sistema KEFIR. O objetivo é a partir do uso deste identificar situações que constituam importantes desvios em relação à norma estabelecida. Os indicadores avaliam diferentes características de provisionamento de saúde, tais como custo, uso e qualidade. O sistema KEFIR define o interesse de algumas variáveis, em termos de seus benefícios, como por exemplo, potencial de redução de custos de uma determinada ação corretiva que recupere o desvio ao seu normal. Essas ações corretivas são especificadas previamente pelo especialista do domínio para várias classes de desvios. Essa abordagem em definir uma medida subjetiva de interesse é executada com sucesso pelo KEFIR, entretanto ela é bastante específica em função de: a) trabalhar apenas com padrões expressos em desvios; b) préclassificar todos os padrões a serem descobertos em um conjunto finito de classes, de forma a determinar uma ação corretiva para cada classe (tendo em vista ao KEFIR trabalhar com um determinado domínio da saúde); e c) ele parte de diversas hipóteses do domínio sobre os benefícios estimados a serem processados.

Silberschatz e Tuzhilin (1996) apresentam um estudo das medidas de interesse subjetivas, classificando-as em acionáveis e surpreendentes. A forma de avaliação do quão surpreendente é o conhecimento descoberto é realizada em função das crenças do usuário. A questão de interesse do padrão é expressa em termos do quanto este afeta o sistema de crenças. Os autores defendem que as características se acionável e/ou surpreendente são independentes entre si, entretanto que a acionabilidade parece ser um conceito mais amplo e de maior dificuldade de formalização. Uma vez que os padrões mais inesperados sejam acionáveis e que os mais acionáveis sejam inesperados, os autores propõem capturar o quesito acionabilidade a partir do quanto o mesmo seja inesperado. Desta forma os autores se dedicam a tratar o inesperado como uma medida de interesse e em definir interesse do padrão em termos do quanto ele *mexe* com o sistema de crenças existente. Estes mesmos autores em 1995 (Silberschatz e Tuzhilin 1995) propuseram usar um sistema de crenças probabilísticas como um *framework* para descrever interesse subjetivo. Especificamente,

um sistema de crença é usado para definir o que é esperado.

Liu *et al*. (1999) propõem uma técnica, chamada expectativa do usuário (*user-expectation method*), na qual o usuário inicialmente define os padrões que representam a sua expectativa de acordo com sua experiência ou mesmo *feeling*. A partir destas expectativas, o sistema utiliza a lógica *fuzzy* para processar os padrões descobertos em relação àquelas expectativas. Como resultado deste processamento pode-se obter distintos produtos: confirmação das expectativas inicialmente identificadas e /ou identificar padrões surpreendentes.

Existem também alguns trabalhos que propõem que a modelagem dos algoritmos de *Data Mining* já aproprie a habilidade de descobrir regras mais interessantes, minimizando assim a etapa de pós-processamento. Um exemplo é o trabalho de Romão *et al*. (2002) que propõe um Algoritmo Genético (AG) especificamente modelado para descobrir regras *fuzzy* interessantes para predição. Para uma regra de predição ser considerada interessante, ela deve representar um conhecimento que o usuário desconhecia previamente, bem como, deve contradizer as suas respectivas crenças originais. A adoção da lógica *fuzzy* objetiva melhorar a compreensibilidade das regras descobertas pelo AG.

2.2 Métodos objetivos

Hsu *et al*. (2000) descrevem um experimento realizado sobre uma base contendo dados de pacientes com diabete, o qual descreve um método que permite ao usuário entender melhor o conjunto de regras descobertas. Um dos métodos propostos foi um algoritmo chamado LCD, o qual usa testes de dependência, independência e independência condicional de variáveis, restringindo assim uma possível relação causal entre as mesmas. Essa técnica fundamenta-se na condição de Markov. Dado A como sendo um nó na rede causal Bayesiana, e dado B como qualquer nó que não seja um nó descendente da rede causal de A, então a condição de Markov diz que A e B são independentes, condicionadas sobre os ancestrais de A.

No experimento descrito no artigo o algoritmo LCD foi modificado para determinar a relação causal envolvendo múltiplos fatores. O algoritmo assume os testes de dependência e independência condicional usando o teste chi-quadrado. O algoritmo LCD implementado também permite identificar as regras de senso-comum e suas respectivas regras de exceção.

Fabris e Freitas (1999) apresenta o paradoxo de Simpson como forma de avaliar o quão surpreendentes são os padrões descobertos. Por exemplo, o

paradoxo encontrado na comparação de mortes por tuberculose em Nova Iorque e Richmond durante o ano de 1910, Tabela 2. Em geral a taxa de mortalidade por tuberculose de Nova Iorque era menor que a taxa de Richmond. Entretanto, o contrário foi observado quando os dados foram particionados de acordo com categorias raciais: brancos e não-brancos. Em ambos os casos (brancos e não-brancos), Richmond teve uma taxa de mortalidade menor. Sendo assim, ao particionar os dados unicamente por cidade o atributo de interesse era a ocorrência de morte, e ao particionar por cidade e categoria racial, o atributo de interesse era categoria racial.

Em particular, a ocorrência deste paradoxo em um conjunto de treinamento pode facilmente enganar o algoritmo de *Data Mining*, levando-o a não interpretar corretamente a relações entre alguns atributos. Por exemplo os algoritmos que induzem árvores de decisão, em geral constróem uma árvore selecionando um atributo por vez. Desta forma eles podem selecionar um atributo que aparentemente tem uma certa relação com o atributo meta, quando na realidade a relação verdadeira, levando-se em conta a interação entre atributos, pode ser distinta da então identificada.

TABELA 2: PARADOXO DO SIMPSOM NOS DADOS SOBRE MORTES POR TUBERCULOSE.

	Nova Iorque		Richmond	
População total	4.766.883		127.682	
Número mortes	8.878		286	
Percentual	0,19%		0,22%	
	Branco	Não-branco	Branco	Não-branco
População total	4.675.174	91.709	80.895	46.733
Número mortes	8.365	513	131	155
Percentual	0,18%	0,56%	0,16%	0,33%

Hussain *et al.* (2000) apresentam um método para identificar no conjunto de padrões descoberto as regras de exceção. Esta medida está baseada na estimativa de interesse relativo entre a regra em questão e a correspondente regra que representa o senso comum. Os autores partem do princípio que a contradição ao senso comum pode ser surpreendente. A tabela 3 apresenta a estrutura adotada para definir as exceções, onde A e B representam uma condição ou um conjunto de condições.

Fica claro a partir da estrutura (Tabela 3) que o item de referência B é o que explica a causa da exceção, em relação ao senso comum $A \rightarrow X$.

Quando nenhuma outra informação é fornecida, um evento com baixa probabilidade de ocorrer oferece mais informação se comparado a um evento com mais alta probabilidade. A partir da teoria da informação, o número de bits requeridos para descrever a ocorrência é definido como:

$I = - LOG_2 P$

Onde P é a probabilidade de um evento ocorrer.

Similarmente, para uma dada regra AB $\rightarrow$ X com confiança Pr(X|AB), irá requerer - $\log_2$ Pr(X|AB) e - $\log_2$ Pr($\neg$X |AB) número de bits para descrever os eventos X e $\neg$X dado AB. Entretanto, a diferença de bits na descrição da regra AB em termos de A $\rightarrow$ X e B $\rightarrow$ X pode não trazer surpresa. Quanto maior a diferença na descrição em relação a regra AB $\rightarrow$ X, mas ela é interessante.

Cago e Bento (1998) descrevem uma métrica de distância entre duas regras, a partir da qual é selecionado o subconjunto de regras mais heterogêneo. Os autores propõem esta métrica com o objetivo que as regras redundantes sejam eliminadas e assim se destaquem as regras mais interessantes.

A medida de distância entre duas regras com o mesmo conseqüente é baseada em três fatores: o número de atributos em uma regra e ausente em outra, o número de atributos em ambas as regras com sobreposição de valores e o número de atributos em ambas as regras com pequena ou nula sobreposição. Baseado nestes fatores a métrica proposta pelos autores é:

$dist(r_i, r_j) = (\alpha \# DA_{i,j} + \beta \# DV_{i,j} - \varpi \# EV_{i,j}) / (\# F_i + \# F_j)$ se $\# NO_{i,j} = 0$

$= 2$ caso contrário onde

- # $DA_{i,j}$ Número de atributos na regra i e não na regra j mais o número de atributos na regra j e não na regra i
- # $NO_{i,j}$ Número de atributos em ambas as regras i e j, mas sem nenhuma sobreposição de valores
- # $DV_{i,j}$ Número de atributos em ambas as regras i e j, mas com pequena sobreposição de valores (< 66%)
- # $EV_{i,j}$ Número de atributos em ambas as regras i e j, mas com sobreposição de valores (> 66¨%)
- # F_i + # F_j Número de atributos na regra i mais o número de atributos na regra j

Para todos os atributos que aparecem em ambas as regras é verificada a interseção de seus valores. Se não existe interseção é sabido que as duas regras não podem ser aplicadas para os mesmos casos e desta forma, é determinado o valor 2 (dois) para a distância entre as duas regras.

Na métrica de distância as características que mais evidenciam a diferença entre duas regras ($DA_{i,j}$ e $DV_{i,j}$) aumentam o valor retornado pela função. Aquelas que aparecem em ambas as regras ($EV_{i,j}$) decrescem o valor da função. Os termos $DA_{i,j}$, $DV_{i,j}$ e $EV_{i,j}$ são ponderados pelas constantes a, b e v. Os autores

TABELA 3: ESTRUTURA DAS REGRAS DE EXCEÇÃO.

A $\rightarrow$ X regra de senso comum (alta cobertura e alta confiança (taxa de acerto?)
A, B $\rightarrow \neg$ X regra de exceção (baixa cobertura, baixa confiança)
B $\rightarrow \neg$ X regra de referência (baixa cobertura e/ou baixa confiança)

propuseram em seu artigo os seguintes valores: a=1, b=2 e v=2, limitando que o intervalo de valores possível seja entre –1 (regras com poucas situações em comum) e 1 (forte sobreposição).

Liu *et al.* (1999) apresentam o sistema DM-II (*Data Mining – Integration and Interestingness*). Quanto a questão específica de descoberta de conhecimento interessante para o usuário o sistema proposto, para a tarefa de *Data Mining* especificamente, adota o teste chi-quadrado como forma de podar o conjunto de regras gerado, objetivando assim a redução das regras ditas insignificantes. Após o primeiro conjunto de regras ter sido descoberto, é definido um subconjunto denominado regras DS (*direction setting rules*) com o objetivo de constituir um sumário das regras que não foram podadas no primeiro processo.

Essencialmente as regras DS são regras significativas que estabelecem as direções para as demais regras. A direção é determinada por correlação positiva, correlação negativa ou independente, a qual é obtida, novamente, a partir do teste chi-quadrado. Sumariza o conjunto descoberto, mas não necessariamente descobre regras surpreendentes.

Liu *et al.* (2000) apresentam uma nova técnica de organizar o conjunto de regras descobertas em diferentes níveis de detalhes. O algoritmo consiste em duas fases, a primeira se preocupa em encontrar as regras gerais, descendo pela árvore de decisão a partir do nó raiz para encontrar o nó mais próximo que represente a classe da maioria, ou seja uma regra significante. Estas regras são denominadas regras gerais *top-level*. A segunda é encontrar as exceções, as exceções das exceções e assim por diante. Os autores determinam que se um nó de uma árvore forma uma regra de exceção ou não usando dois critérios: significância e simplicidade. Algumas das regras de exceção encontradas por este método poderiam ser consideradas pequenos disjuntos. Entretanto, este método não se propõe a discutir esta questão, ele objetiva apenas a sumarizar um grande conjunto de regras.

Liu *et al.* (2001) apresentam que uma regra pode ser significante, porém pode não ser potencialmente útil para ação. Por exemplo, a partir de uma base de dados de 1000 tuplas, o domínio de valores do atributo meta é sim ou não. A distribuição de freqüência das tuplas é 50% sim e 50% não. Considerando as três regras descobertas:

R1: pressãosanguínea = alta $\rightarrow$ Meta = sim [Suporte = 6% confiança = 60%]

R2: pressãosanguínea = alta, sexo = masculino $\rightarrow$ meta = sim [Suporte = 3.6% confiança = 90%]

R3: pressãosanguínea = alta, nívelglicose = anormal $\rightarrow$ meta = sim [Suporte = 3% confiança = 100%]

A partir de R1 é possível perceber que o número de tuplas com pressão sangüínea = alta é 60 (6% * 1000), a cobertura da regra R1 é 100 (60/60%). Fazendo o mesmo tipo de análise em relação a R2, são 36(3.6% * 1000) tuplas que tem pressão sangüínea alta, são do sexo masculino e o valor do atributo meta é sim. A respectiva cobertura é 30 (3.0% * 90%). E assim por diante considerando R3.

Assumindo que as três regras não sejam podadas, ou seja, são significativas pode ser realizada a seguinte análise. O número de tuplas cobertas por R2 e R3 com valor sim para o atributo meta é 58. O número de tuplas cobertas por R2 ou R3 é 62 (incluindo ambos os valores para o atributo meta). Desde que R2 e R3 têm um nível muito mais alto de confiança e elas cobrem em sua maioria as instancia com meta = sim, elas deveriam ser usadas primeiramente em uma aplicação (elas têm uma qualidade maior). Isto implica que as tuplas remanescentes da cobertura por R1 – (R2 $\cup$ R3) são 2 tuplas com meta = sim e 36 tuplas com meta = não. Claramente, R1 – (R2 $\cup$ R3) denominada R1', não constitui uma regra significativa para o valor sim do atributo meta, dada a sua respectiva confiança ser muito baixa (2/(2+36) = 5;3%) comparado com a confiança *default* do atributo meta = sim que é (500 – 58) / (1000-62) = 47%. Em outras palavras, o valor sim para o valor sim do atributo meta R1' é pior que a solução aleatória dada pela confiança *default*. Desde que R1 é efetivamente R1' quando fazendo parte do grupo de 3 regras, R1 é dita não *acionável*.

A técnica proposta neste artigo trabalha em duas fases: as regras são geradas e podadas de acordo com um critério de significância, no caso específico foi adotado o teste chi-quadrado; em seguida, são analisadas as regras restantes a partir das regras com mais condições para as regras com menor número de condições, ou seja, usando regras especializadas com maior nível de qualidade para determinar se uma regra mais geral é potencialmente *acionável*. É importante observar que identificar as regras não-acionáveis opera de forma inversa em relação aos processos tradicionais de poda, os quais em geral podam as regras especializadas.

Freitas (1998) introduz uma outra métrica de avaliar a surpresa da regra, chamada *AttSurp* (Attribute Surprisingness). O artigo propõe que *AttSurp* seja definida a partir do ganho de informação (medida da teoria da informação) (Mitchell, 1997). As regras que forem compostas por atributo(s) com baixo ganho de informação tendem a ser mais surpreendentes. Esses atributos podem ser considerados irrelevantes se tomados individualmente, entretanto combinados a outros atributos podem vir a se tornar relevantes.

Matematicamente o cálculo do AttSurp é expresso por:

$$AttSurp = 1 / \sum_{i=1}^{K} GanhoInformação(A_i / K)$$

onde GanhoInformação(A_i) é o ganho de informação do i-ésimo atributo que ocorre no antecedente da regra e *k* é o número de atributos neste antecedente.

Suzuki e Kodratoff (1998) apresentam um algoritmo para descobrir regras de exceção que sejam surpreendentes a partir de bases de dados. Em primeiro lugar, é formalizado o problema em termos de limites menos rígidos na confiança da regra de exceção. Este artigo propõe modificações no algoritmo proposto anteriormente pelos mesmos autores PEDRE. No algoritmo PEDRE, o problema consiste em identificar pares de regras das quais uma representa a regra de exceção associada a regra de senso comum, na forma:

$r(\mu, \nu)$ º $A_\mu \rightarrow c$

$A_\mu \wedge B_\nu \rightarrow c'$

Onde $A\mu \equiv a_1 \wedge a_2 \wedge ... \wedge a_n$ $B\mu \equiv b_1 \wedge b_2 \wedge ... \wedge b_n$ sendo º $A\mu \rightarrow c$, $A\mu \wedge B\nu \rightarrow c'$ regras de senso comum e $Bn \rightarrow c'$ regra de exceção.

Um critério para avaliação de surpresa é a Intensidade da Implicação, o qual representa o grau de surpresa que uma regra $A \rightarrow c$ tenha poucos contra-exemplos. Sejam *U* e *V* conjuntos relacionados aleatoriamente cujos números de exemplos |U| e |V| em um conjunto de dados D são iguais aos números de exemplos em D cobertos pelo conseqüente e antecedente da regra: |U = |c|, |V| = | Aμ |

A Intensidade da Implicação *I* para a regra é:

I = 1 – Pr (|VU'| ≤ |Ac'|)

Desta forma serão descobertas as k regras excepcionais com maior valor de intensidade de implicação.

O emprego da Intensidade da Implicação resolve uma limitação apresentada pela probabilidade condicional, ou seja, a probabilidade de uma regra $X \rightarrow Y$ é invariável em relação ao número de exemplos onde Y ocorre (g(Y)), bem como, o número total de exemplos do conjunto (E). Porém, $X \rightarrow Y$ ocorre mais freqüentemente quando o tamanho de g(Y) aumenta ou quando o tamanho de E diminui, além disso, esta implicação será mais significativa quando o tamanho de todos estes conjuntos aumenta na mesma proporção.

Piatetsky-Shapiro (1991) apresenta a função *rule-interest* (*RI*) que quantifica a correlação entre os atributos de uma regra de classificação. A função RI é dada pela seguinte expressão:

$$RI = |X \cap Y| - \frac{|X||Y|}{N}$$

onde N é o número total de exemplos, |X| e |Y| são os números de exemplos que satisfazem as condi-

ções |X| e |Y|, respectivamente, $|X \cap Y|$ é o número de exemplos que satisfazem $X \rightarrow Y$, e |X||Y| / N é o número de exemplos esperados sendo X e Y independentes (i.e. não associados).

Quando RI = 0, ou seja $|X \cap Y| = |X|\,|Y| / N$, então o X e Y são estatisticamente independentes e a regra não é interessante. Quando o RI > 0 (RI < 0), então o X é positivamente correlacionado (negativamente) correlacionado ao Y.

3 Conclusão e trabalhos futuros

Este artigo procura sintetizar vários trabalhos da literatura que propõem e descrevem métodos para pós-processar o conhecimento descoberto, a partir de algoritmos de *Data Mining*, com o objetivo de tornar mais interessante e/ou facilitar a interpretação por parte do usuário, durante o processo decisório. Os artigos tratados fazem parte de dois grupos: os métodos ditos subjetivos e aqueles caracterizados como objetivos.

Da mesma forma que não existe a indicação que um determinado algoritmo de *Data Minig* seja o mais apropriado para qualquer domínio, a mesma afirmação se aplica para os algoritmos que pós-processam o conhecimento descoberto. O sucesso de um experimento de descoberta de conhecimento, que subsidie o processo decisório, está na habilidade de identificar os algoritmos que serão utilizados, não apenas para a etapa de descoberta, mas também para a etapa de refinamento deste conhecimento.

Como trabalhos futuros pode-se propor a experimentação destes diversos métodos, considerando distintos domínios, problemas, etc. tentando identificar em que situação os mesmos podem ser mais indicados. Inclusive numa situação de experimentação poderia ser interessante contar com a avaliação de usuários para qualificar os resultados produzidos por esses métodos.

Referências bibliográficas

CAGO, P.; BENTO, C. (1998). *A metric for selection of the most promising rules*. PKDD-1998, p. 19-27.

FABRIS, C. C.; FREITAS, A. A. (1999). "Discovering surprising patterns by detecting occurrences of Simpson's paradox". In: *Research and Development in Intelligent Systems XVI* (Proc. ES99, The 19th SGES Int. Conf. on Knowledge-Based Systems and Applied Artificial Intelligence), 148-160. Springer-Verlag.

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

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

GEBHARDT, F. (1991). "Choosing among competing generalizations". *Knowledge* Acquisition 3, p. 361-380.

HSU, W.; LEE, M. L.; LIU, B.; LING, T.W.. (2000). *Exploration Mining in Diabetic Patients Databases*: Findings and Conclusions. KDD – 2000, p. 430-436.

HUSSAIN, F.; LIU, H.; LU, H. (2000). *Exception Rule Mining with a Relative Interestingness Measure*. PAKDD-2000, LNAI 1805, p. 86-96.

LIU, B.; HSU, W.; MA, Y. (2001). "Identifying Non-Actionable association Rules". *International Conference on Knowledge Discovry & Data Mining* KDD-2001.

LIU, B.; HSU, W.; MA, Y.; LEE, H.Y.; CHEN S. (1999). "Mining Interesting Knowledge Using DM-II". ACM SIGKDD *International Conference on Knowledge Discovry & Data Mining* (KDD-99).

LIU, B.; HSU, W.; MUN, L.F.; LEE, H.Y. (1999). "Finding Interesting Patterns Using User Expectations". IEEE *Transactions on Knowledge and Data Engineering*. Vol. 11 (6) p. 817-832.

LIU, B.; HU, M.; HSU, W. (2000). *Multi-Level Organization and Summarization of the Discovered Rules*. KDD-2000 p. 208-217.

MATHEUS, C.J.; PIATETSKY-SHAPIRO, G.; MCNeill, D. (1994). "An Application of Keffir to analysis of healthcare information". In *Proc. Of the AAAI-94 Workshop on Knowledge Discovery in Data Bases.*

MITCHELL, T. M. (1997). *Machine Learning*. MacGraw-Hill, USA.

MONARD, M. C.; BARANAUSKAS, J.A. (2003). "Indução de regras e Árvores de Decisão". In *Sistemas Inteligentes*. REZENDE, S. O. Editora Manole Ltda. p. 115-140.

PIATETSKY-SHAPIRO, G. (1991). "Discovery, analysis and presentation of strong rules". In *Knowledge Discovery in Databases*. AAI/MIT-Press. p.229-248.

ROMAO, W.; FREITAS, A.A.; PACHECO, R.C.S. (2002). "A Genetic Algorithm for Discovering Interesting Fuzzy Prediction Rules: applications to science and technology data". To appear in *Proc. Genetic and Evolutionary Computation Conf.* (GECCO-2002).

SILBERSCHATZ, A.; TUZHILIN, A. (1995). "On subjective measures of interestingness in knowledge discovery". In *Proc. Of the First International Conference on Knowledge Discovery & Data Mining*. p 828-834.

_____. (1996). "What Makes Patterns Interesting in Knowledge Discovery Systems". IEEE Tran. *Knowledge & Data Mining*, 8(6).

SUZUKI, E.; KODRATOFF, Y. (1998). *Discovery of Surprising Exception Rules Based on Intensity of Implication*. PKDD-1998.