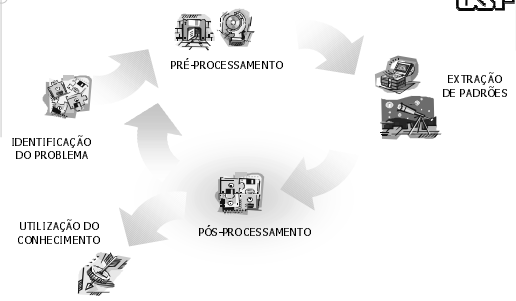
Data Mining (Pós-Processamento do Conhecimento)

Solange Oliveira Rezende
USP São Carlos / ICMC
Departamento de Ciências de Computação e Estatística
Laboratório de Inteligência Computacional
http://www.icmc.usp.br

Etapas do processo de *Data Mining*



O que é ?

- Métodos para investigar a precisão dos algoritmos, a representação do modelo, a complexidade e a dificuldade de entendimento do conhecimento extraído.
- O conhecimento extraído pode ser simplificado, avaliado, visualizado ou simplesmente documentado para o usuário final.

Motivação

- Os algoritmos podem extrair uma quantidade enorme de padrões, muitos dos quais podem não ser importantes, relevantes ou interessantes para o usuário.
- Fornecer ao usuário uma grande quantidade de padrões descobertos não é produtivo.
- O usuário procura por uma pequena lista de padrões interessantes.

Tipos de Pós-processamento

Filtragem do Conhecimento

- Mecanismos de pós-poda (árvore de decisão) e truncagem para regras de decisão.
- Aplicado em casos que os algoritmos geram árvores de decisão com muitas folhas ou regras de decisão muito específicas cobrindo poucos exemplos (*overfitting*).

Tipos de Pós-processamento (cont.)

Interpretação e Explanação

- O conhecimento pode ser sumarizado, documentado, visualizado ou modificado de modo a tornar-se compreensível ao usuário.
- O conhecimento pode ser comparado ao conhecimento preexistente para verificação de conflitos e conformidades.

Tipos de Pós-processamento (cont.)

Avaliação do Conhecimento

- O conhecimento é avaliado por meio de medidas e critérios que podem ser objetivo ou subjetivo.
- Algumas medidas são: interessabilidade, compreensibilidade, complexidade computacional, precisão.

Tipos de Pós-processamento (cont.)

Integração do Conhecimento

- Sistemas tradicionais de apoio à decisão são dependentes de uma única técnica, estratégia e modelo.
- Sistemas novos e sofisticados possibilitam combinar ou refinar resultados de várias técnicas e modelos, buscando uma maior precisão e um melhor desempenho.
- Exemplo: combinação de classificadores (*ensembles*).

Avaliação do Conhecimento

- Tabela de Contingência
- Medidas de Avaliação de Conhecimento Derivadas da Tabela de Contingência
- Outras Medidas de Avaliação de Conhecimento:
  - Medidas de Desempenho
  - Medidas de Qualidade

Tabela de Contingência

Regra: Body (B) -> Head (H)

	H	$\overline{H}$	
B	n(BH)	n(B$\overline{H}$)	n(B)
$\overline{B}$	n($\overline{B}$H)	n($\overline{BH}$)	n($\overline{B}$)
	n(H)	n($\overline{H}$)	*N*

Medidas de Avaliação de Conhecimento Derivadas da Tabela de Contingência

Framework com 16 medidas [Lavrac N., P. Flach & R. Zupan (1999)] :

1 – Precisão	9 – Precisão Relativa
2 – Confiança Negativa	10 – Confiança Negativa Relativa
3 – Sensibilidade	11 – Sensibilidade Relativa
4 – Especificidade	12 – Especificidade Relativa
5 – Cobertura	13 - Precisão Relativa Ponderada
6 – Suporte	14 - Confiança Negativa Relativa Ponderada
7 – Novidade	15 - Sensibilidade Relativa Ponderada
8 - Satisfação	16 - Especificidade Relativa Ponderada

Medidas de Avaliação de Conhecimento Derivadas da Tabela de Contingência (cont.)

1 – Precisão: É uma medida do quanto uma regra é específica para o problema.

$$Acc(B \rightarrow H) = P(H \mid B) = \frac{P(HB)}{P(B)}$$

2 – Confiança Negativa: Corresponde a precisão para os exemplos que não são cobertos pela regra.

$$Neg\,\mathrm{Re}l(B \rightarrow H) = P(\overline{H} \mid \overline{B}) = \frac{P(\overline{HB})}{P(\overline{B})}$$

Medidas de Avaliação de Conhecimento Derivadas da Tabela de Contingência (cont.)

3 – Sensibilidade: É a probabilidade condicional de que B é verdadeiro dado que H é verdadeiro.

$$Sens(B \rightarrow H) = P(B \mid H) = \frac{P(BH)}{P(H)}$$

4 – Especificidade: Corresponde a completeza para os exemplos que não são cobertos pela regra.

$$Spec(B \rightarrow H) = P(\overline{B} \mid \overline{H}) = \frac{P(\overline{BH})}{P(\overline{H})}$$

Medidas de Avaliação de Conhecimento Derivadas da Tabela de Contingência (cont.)

5 – Cobertura: Mede o número de exemplos cobertos pelo corpo da regra (B).

$$Cov(B \rightarrow H) = P(B) = \frac{n(B)}{N}$$

6 – Suporte: Mede o número de exemplos cobertos pela regra.

$$Sup(B \rightarrow H) = P(BH) = \frac{n(BH)}{N}$$

Medidas de Avaliação de Conhecimento Derivadas da Tabela de Contingência (cont.)

7 – Novidade: Identifica o quanto uma regra é inovadora, interessante ou não usual.

$$Nov(B \rightarrow H) = P(BH) - P(B)P(H)$$

8 – Satisfação: Mede o aumento relativo na precisão entre a regra *B -> verdade* e a regra *B -> H*

$$Sat(B \rightarrow H) = \frac{P(\overline{H}) - P(\overline{H} \mid B)}{P(\overline{H})}$$

Medidas de Avaliação de Conhecimento Derivadas da Tabela de Contingência (cont.)

9 – Precisão Relativa: Representa o ganho de precisão obtido em relação a uma regra padrão *verdade -> H*.

$$RAcc(B \rightarrow H) = P(H \mid B) - P(H)$$

10 – Confiança Negativa Relativa: Análogo a precisão relativa para os exemplos que não são cobertos pela regra.

$$RNeg\ \mathrm{Re}l(B \rightarrow H) = P(\overline{H} \mid \overline{B}) - P(\overline{H})$$

Medidas de Avaliação de Conhecimento Derivadas da Tabela de Contingência (cont.)

11 – Sensibilidade Relativa: Mede o ganho relativo de sensibilidade obtido em relação à sensibilidade de uma regra padrão *B -> verdade*.

$$RSens(B \rightarrow H) = P(B \mid H) - P(B)$$

12 – Especificidade Relativa: Análogo a sensibilidade relativa para os exemplos que não são cobertos pela regra.

$$RSpec(B \rightarrow H) = P(\overline{B} \mid \overline{H}) - P(\overline{B})$$

Medidas de Avaliação de Conhecimento Derivadas da Tabela de Contingência (cont.)

13 - Precisão Relativa Ponderada: Avalia de maneira balanceada a generalidade e a precisão relativa de uma regra.

$$WRAcc(B \rightarrow H) = P(B)(P(H \mid B) - P(H))$$

14 - Confiança Negativa Relativa Ponderada: Avalia de maneira balanceada a generalidade e a confiança negativa relativa de uma regra.

$$WRNeg\ \mathrm{Re}l(B \rightarrow H) = P(\overline{B})(P(\overline{H} \mid \overline{B}) - P(\overline{H}))$$

Medidas de Avaliação de Conhecimento Derivadas da Tabela de Contingência (cont.)

15 - Sensibilidade Relativa Ponderada:
Estabelece um balanceamento entre a sensibilidade relativa e a generalidade de uma regra.

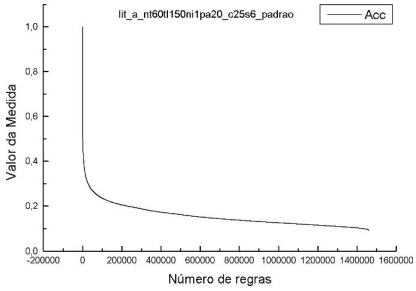
$$WRSens(B \rightarrow H) = P(H)(P(B \mid H) - P(B))$$

16 - Especificidade Relativa Ponderada:
Estabelece um balanceamento entre a especificidade relativa e a generalidade de uma regra.

$$WRSpec(B \rightarrow H) = P(\overline{H})(P(\overline{B} \mid \overline{H}) - P(\overline{B}))$$

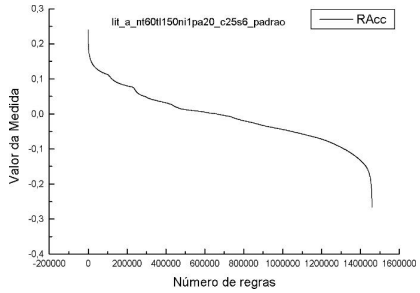
Exemplo de Comportamento de algumas Medidas

Precisão



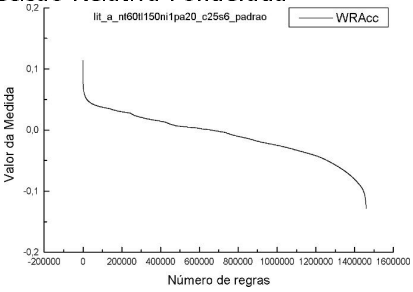
Exemplo de Comportamento de algumas Medidas (cont.)

Precisão Relativa



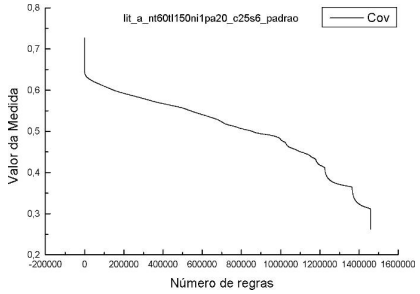
Exemplo de Comportamento de algumas Medidas (cont.)

Precisão Relativa Ponderada



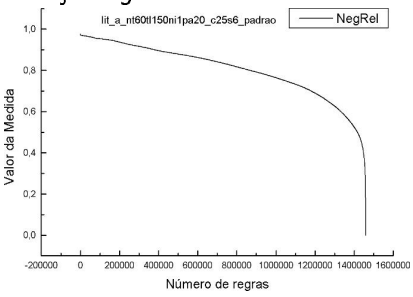
Exemplo de Comportamento de algumas Medidas (cont.)

Cobertura



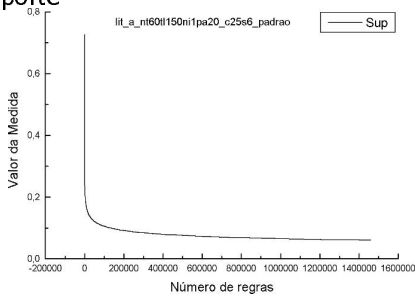
Exemplo de Comportamento de algumas Medidas (cont.)

Confiança Negativa



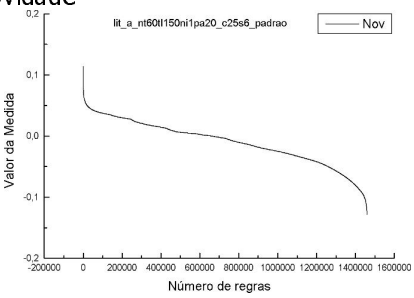
Exemplo de Comportamento de algumas Medidas (cont.)

Suporte



Exemplo de Comportamento de algumas Medidas (cont.)

Novidade



Outras Medidas de Avaliação de Conhecimento

Medidas de Desempenho

- Desempenho pode ser visto como o conjunto de características ou de possibilidades de atuação de um processo, tais como velocidade, capacidade, agilidade, autonomia, rendimento, etc.
- Os resultados dependem de como o experimento foi realizado.

Medidas de Desempenho (CONT.)

- Precisão
- Erro Verdadeiro
- Erro Amostral
- Variância
- Consistência
- Completude
- Velocidade
- Tempo de Aprendizado

Medidas de Desempenho (CONT.)

- Precisão
  - Grau de confiabilidade do conhecimento.
  - Geralmente, é representado pela taxa de erro, embora em alguns casos existam erros mais sérios que outros.

Medidas de Desempenho (CONT.)

- Erro Verdadeiro
  - É a taxa de erro da hipótese sobre toda a desconhecida distribuição *D* de exemplos.
  - O erro verdadeiro de uma hipótese *h* em relação a função objetivo *f* e uma distribuição *D*, é a probabilidade de *f* mal predizer um exemplo *x* extraído aleatoriamente da distribuição *D*.

Medidas de Desempenho (CONT.)

- Erro Amostral ou Aparente
  - Refere-se à taxa de erro da hipótese sobre a amostra de dados disponível.

Medidas de Desempenho (CONT.)

- Variância
  - Mede quanto as suposições do algoritmo de aprendizado variam com respeito a outras, isto é, como é a flutuação para conjuntos diferentes de treinamento.

Medidas de Desempenho (CONT.)

- Consistência
  - É a medida do quanto um conhecimento é específico para o problema.
  - Quanto mais alta a consistência, mais precisamente é coberta a classe em questão.
  - A consistência é máxima quando a regra cobre somente os exemplos da classe e nenhum exemplo fora desta classe, isto é, quando não existem erros positivos.

Medidas de Desempenho (CONT.)

- Completude
  - É a medida que calcula o quanto do domínio do problema é coberto pela regra.
  - Quanto mais alta a completude, mais exemplos são cobertos pela regra.

Medidas de Desempenho (CONT.)

- Velocidade
  - Em algumas circunstâncias a velocidade de um preditor é um aspecto muito importante.

Medidas de Desempenho (CONT.)

- Tempo de Aprendizado
  - Pode se tornar um fator muito importante em ambientes que mudam rapidamente, pois pode ser necessário aprender um novo conhecimento rapidamente, ou ajustar um novo conhecimento ao já existente.

Medidas de Qualidade

- São medidas que permitem avaliar e, conseqüentemente, aprovar, aceitar ou recusar o conhecimento extraído.
- Algumas das medidas relatadas anteriormente também podem ser utilizadas para avaliar a qualidade do conhecimento obtido.

Medidas de Qualidade (CONT.)

- Algumas abordagens consideram fatores como:
  - a possível existência de um grande número de regras dificultando a análise por parte do usuário;
  - a possível complexidade ao avaliar a regra, por exemplo, o número de condições ou a estrutura de representação.
- As medidas de avaliação de qualidade estão mais relacionadas com os conceitos de Compreensibilidade e Interessabilidade.

Medidas de Qualidade (CONT.)

Compreensibilidade:

- Está ligado a conceitos subjetivos relacionados, principalmente, a facilidade de compreensão do modelo ou conhecimento por um ser humano leigo, e a capacidade exploratória dos padrões extraídos no processo.

Medidas de Qualidade (CONT.)

Compreensibilidade:

- Regra:
  - número de atributos na parte condicional;
  - número de atributos na parte conclusiva;
  - número de condições;
  - número de regras;
  - número de termos da função.

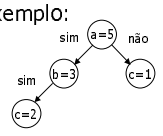
Exemplo: a=5 **and** b=3 **->** c=2

Medidas de Qualidade (CONT.)

Compreensibilidade:

- Árvore:
  - número de nós;
  - profundidade;
  - largura.

Exemplo:



Geralmente, é mais fácil compreender o conhecimento representado por uma árvore do que por um conjunto de regras.

Medidas de Qualidade (CONT.)

LABIC

USP

Interessabilidade:

- É uma maneira de avaliar qualidade tentando capturar o conhecimento interessante (ou inesperado).
- Estão baseadas em vários aspectos, principalmente na utilidade que o conhecimento representa para o usuário.
- Interessa ao usuário identificar os padrões que tenham erros mínimos e que consigam generalizar um grande número de exemplos.

Medidas de Qualidade (CONT.)

LABIC

USP

Interessabilidade:

- **Objetiva:** está relacionada somente com a estrutura dos padrões e do conjunto de dados de teste. Não levam em consideração fatores específicos do domínio.
- **Subjetiva:** depende também da classe de usuários que examinam os padrões. Consideram fatores específicos do conhecimento do domínio e o interesse do usuário para selecionar um conjunto de regras interessantes.

Medidas de Qualidade (CONT.)

LABIC

USP

Interessabilidade Objetiva:

- Modelos de regras: interessam somente as regras que "casam" com os modelos definidos com o especialista/usuário e o analista.
- Cobertura de Regras Mínima: um conjunto mínimo de regras que cobrem o maior número de exemplos do conjunto de dados são apresentadas ao usuário.
- Medida PS: leva em consideração a questão do desbalanceamento das classes favorecendo regras que predizem a classe minoritária.

Medidas de Qualidade (CONT.)

LABIC

USP

Interessabilidade Subjetiva:

- Inesperabilidade: avalia se as regras são excepcionais (não observadas) frente ao conhecimento do especialista do domínio. Pode ser interpretada no sentido estatístico, como uma alta casualidade que aparece sob uma suposição ou hipótese independente ou sob algum *bias* que seja contrário às crenças do usuário.
- Acionabilidade: avalia se o usuário pode utilizar as regras para obter vantagem na aplicação da regra, ou seja, as regras são consideradas interessantes, desde que o usuário possa obter vantagens ao utilizá-las.

Medidas de Qualidade (CONT.)

LABIC

USP

Algumas outras medidas também utilizadas em interessabilidade são:

- Interessabilidade dos Atributos;
- Custo de Classificação Incorreta e Tamanho do Disjunto;
- Extensão da Medida PS;
- Grau de Surpresa de Pequeno Disjuntos;
- Grau de Surpresa dos Atributos Individuais de Regras.

Pós-Processamento em Classificação

LABIC

USP

Classificação

Consiste na generalização de exemplos com suas respectivas classes conhecidas em um modelo capaz de reconhecer a classe de um novo exemplo:

Regras de Produção

```
a=5 and b=3 -> c=2
```

Árvores de Decisão

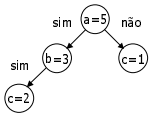


Tabela de Contingência para Classificação

	Positivos	Negativos	
Preditos como positivos	*p*	*fp*	*p + fp*
Preditos como negativos	*fn*	*n*	*fn + n*
	p + fn	*fp + n*	*N*

Pós-Processamento em Classificação (cont.)

Medidas de Qualidade

- Estatística de Cohen
- Estatística IMAFO
- Estatística de Coleman
- SKIB1 e SKIB2
- Uso da Informação

Pós-Processamento em Classificação (cont.)

- Estatística de Cohen
  - Busca encontrar um nível de associação na diagonal principal da matriz de contingência, para servir como medida de qualidade de uma regra.
  - O cálculo é feito através da combinação dos valores de consistência e completude de uma regra com o número total de exemplos e o número de exemplos de uma classe cobertos pela regra.

Pós-Processamento em Classificação (cont.)

- Estatística IMAFO
  - Apresenta um valor real entre 0 e 10 para medir a qualidade de uma regra.
  - Embora sejam esperados bons resultados dessa estatística, na prática existem dificuldades quanto a sua interpretação.

Pós-Processamento em Classificação (cont.)

- Estatística de Coleman
  - É uma medida da "combinação" entre a primeira coluna e cada linha da tabela de contingência.
  - Os valores negativos significam um relacionamento inverso entre a regra e a classe por ela predita.
  - Essa estatística não extrai a completude da regra. Isso a torna incapaz de medir o efeito de falsos negativos na qualidade da regra.

Pós-Processamento em Classificação (cont.)

- SKIB1 e SKIB2
  - São combinações das estatísticas de Cohen e Coleman (aproveitando as melhores características de ambas).
  - São medidas de combinação e associação, isto é, elas extraem valores de toda a tabela de contingência para retornar a qualidade da regra.
  - Lidam bem com muitos dos problemas de distribuição de classes que causam confusão em outras medidas.

Pós-Processamento em Classificação (cont.)

- Uso da Informação
  - A teoria da informação é também uma área que está estritamente relacionada a estatística, e pode oferecer alguma ajuda para medições de qualidade.
  - Assim, o resultado deve ser usado como um limite mínimo de qualidade da regra. Tal como a estatística de Coleman, essa medida falha ao incorporar a completude da regra gerando as mesmas conseqüências indesejáveis.

Pós-Processamento em Regressão

Regressão

- Consiste em obter um modelo, baseado em uma amostra de exemplos, que descreva uma relação entre um conjunto de atributos de entrada e um atributo-meta contínuo (função de Regressão):

$$y = f(x_1, x_2, \ldots, x_d)$$

- O modelo é utilizado para predizer o valor de um atributo-meta contínuo de novos exemplos.

Tabela de Contingência para Regressão

- Valor predito é um valor numérico ou uma equação, o que envolve um grau de incerteza quanto à classe predita.
- Dificuldade na identificação do número de elementos classificados correta e incorretamente para cada categoria.
- O cálculo do valor do número de exemplos falsos e verdadeiros quanto ao atributo meta (ou seja, a cobertura da regra) se torna mais complicado por ser contínuo.

Tabela de Contingência para Regressão (cont.)

	B	$\overline{B}$
H	n(HB)	n(H $\overline{B}$)
$\overline{H}$	n($\overline{H}$ B)	n($\overline{HB}$)

- B denota o conjunto de exemplos para o qual o corpo da regra é verdadeiro
- $\overline{B}$ denota o seu complemento (o conjunto de exemplos para o qual o corpo da regra é falso)
- De forma similar H e $\overline{H}$

Pós-Processamento em Regressão (cont.)

Medidas de Desempenho

Precisão para problemas de regressão:

- MAD
- MSE
- RMSE

Pós-Processamento em Regressão (cont.)

- MAD (*Mean Absolute Deviation* ou Média da Diferença Absoluta)
  - É uma medida de erro que quantifica o erro do modelo pela média dos desvios absolutos de suas predições, isto é, consiste na média da diferença (em módulo) entre os valores reais e preditos para um atributo-meta.

Pós-Processamento em Regressão (cont.)

- MSE (*Mean Squared Error* ou Média dos Erros ao Quadrado)
  - Consiste na média do quadrado da diferença entre os valores reais e preditos para um atributo-meta.
  - Essa medida, muitas vezes, é utilizada para minimizar o erro quanto à predição dos valores.

Pós-Processamento em Regressão (cont.)

- RMSE (*Relative Mean Squared Error* ou MSE Relativa)
  - Essa medida fornece um valor relativo para o erro.
  - Um valor entre zero e um indica que a regra está se saindo melhor que apenas predizer o valor médio y.

Pós-Processamento em Regressão (cont.)

Medidas de Qualidade

- GanhoMAD
- LC
- Q

Pós-Processamento em Regressão (cont.)

- GanhoMAD
  - Quantifica o ganho entre duas regras de regressão diferenciadas pela medida MAD.
  - Para a utilização dessa medida é necessário que a MAD esteja normalizada. Essa normalização é necessária para evitar a ocorrência de um ganho negativo.

Pós-Processamento em Regressão (cont.)

- LC (*Lost Coverage*)
  - É uma medida de interessabilidade que quantifica a perda de cobertura dos exemplos entre duas regras R e R'.
  - O valor de LC para a regra original é o próprio número de exemplos.

Pós-Processamento em Regressão (cont.)

- Q
  - Com o auxílio das medidas GanhoMAD e LC, uma medida de qualidade Q da regra R pode ser calculada a partir de constantes (pesos) atribuídas a GanhoMAD e a LC.
  - Se o valor de w_{ganho} for alto, de acordo com a medida Q, as regras mais específicas terão uma maior interessabilidade. Por outro lado, se o valor de w_{ganho} for baixo, as regras mais gerais serão as escolhidas, diminuindo-se a precisão.

Pós-Processamento em Regras de Associação

Regras de Associação

- Caracteriza o quanto a presença de um conjunto de atributos nos registros de uma Base de Dados implica na presença de algum outro conjunto distinto de atributos nos mesmos registros:

$$LHS \rightarrow RHS$$

- Objetiva encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados.

Tabela de Contingência para Regras de Associação

	RHS	$\overline{RHS}$	
LHS	n(LHS RHS)	n(LHS $\overline{RHS}$)	n(LHS)
$\overline{LHS}$	n($\overline{LHS}$ RHS)	n($\overline{LHS}$ $\overline{RHS}$)	n($\overline{LHS}$)
	n(RHS)	n($\overline{RHS}$)	*N*

Pós-Processamento em Regras de Associação(cont.)

Medidas Objetiva de Avaliação

- Suporte e Confiança [Agrawal R. & R. Srikant (1994)];
- *Lift* e Confiança Eperada [Rathjens D. (1996)];
- Medidas J1 e J2 [Wang K., S. H. W. Tay & B. Liu (1998)];
- *Framework* [Lavrac N., P. Flach, & R. Zupan (1999)];
- Convicção [Adamo J.M. (2001)].

Pós-Processamento em Regras de Associação(cont.)

Suporte

Quantifica a incidência da regra no conjunto de dados. É equivalente a probabilidade de que LHS e RHS ocorram juntos no conjunto de dados :

$$p(LHS\ RHS) = \frac{n(LHS\ RHS)}{N}$$

N – número total de transações (exemplos) consideradas.

n(LHS RHS) – número de transações na qual LHS e RHS ocorrem juntos.

Pós-Processamento em Regras de Associação(cont.)

Confiança

Indica a freqüência com que LHS e RHS ocorrem juntos em relação ao número total de registros em que LHS ocorre. É equivalente a probabilidade condicional p(LHS | RHS) :

$$p(LHS \mid RHS) = \frac{n(LHS\ RHS)}{n(LHS)}$$

n(LHS) – número de transações na qual LHS ocorre.

Pós-Processamento em Regras de Associação(cont.)

Medidas Subjetivas de Avaliação

[Liu B., W. Hsu, S. Chen & Y. Ma (2000)]

- Identificação de conformidade;
- Conseqüente inesperado;
- Antecedente inesperado;
- Antecedente e conseqüente inesperados.

Pós-Processamento em Regras de Associação(cont.)

Alguns conceitos importantes

- Impressão Geral : informar uma relação que o especialista acredita existir entre os itens especificados.
- Conhecimento Impreciso : informar um conhecimento (uma regra LHS -> RHS) que o especialista supõe ser verdadeiro.

Pós-Processamento em Regras de Associação(cont.)

Identificação de Conformidade

Identifica e classifica as regras que estão em conformidade com uma impressão geral ou um conhecimento impreciso fornecido pelo usuário do domínio.

Pós-Processamento em Regras de Associação(cont.)

Conseqüente Inesperado

Permite avaliar se o conseqüente da regra é inesperado. Essa medida identifica e classifica as regras que são contrárias a impressão geral ou conhecimento impreciso fornecido pelo usuário do domínio.

Pós-Processamento em Regras de Associação(cont.)

Antecedente Inesperado

Identifica outros antecedentes que estão associados aos conseqüentes especificados na impressão geral ou no conhecimento impreciso.

Pós-Processamento em Regras de Associação(cont.)

Antecedente e Conseqüente Inesperado

Permite avaliar se o antecedente e o conseqüente da regra são inesperados. A medida recorda o usuário de que existem outras regras, na qual, o antecedente e o conseqüente não haviam sidos especificados na impressão geral ou no conhecimento impreciso.

Considerações Finais

- Após a análise do conhecimento, se os resultados não forem satisfatórios, o processo de extração pode ser reiniciado com o objetivo de se obter melhores resultados.
- No final do processo de Extração de Conhecimento, é interessante que todo o conhecimento adquirido seja disponibilizado em um ambiente adequado para facilitar sua interpretação e utilização.

Bibliografia

- Adamo J.M. (2001). Data Mining for Association Rules and Sequential Patterns. Springer-Verlag.
- Agrawal R. & R. Srikant (1994). Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Databases.
- Lavrac N., P. Flach & R. Zupan (1999). Rule Evaluation Measures: A Unifying View. In S. Dzeroski and P. Flach (Eds.), Proceedings of the Ninth International Workshop on Inductive Logic Programming (ILP-99), Volume 1634, pp. 174–185. Springer-Verlag. LNAI.
- Rathjens D. (1996). MineSet™ User's Guide. Silicon Graphics, Inc.
- Wang K., S. H. W. Tay & B. Liu (1998). Interestingness-Based Interval Merger for Numeric Association Rules. In R. Agrawal, P. E. Stolorz, and G. Piatetsky-Shapiro (Eds.), Proc. 4th Int. Conf. Knowledge Discovery and Data Mining, KDD, pp. 121–128. AAAI Press.
- Liu B., W. Hsu, S. Chen & Y. Ma (2000). Analyzing the Subjective Interestingness of Association Rules. IEEE Intelligent Systems & their Applications 15 (5), 47–55.