Universidade Federal do Acre
Programa de Pós-Graduação em Ciência da Computação - PPgCC

Mineração de Dados

Regras de Associação

Prof. Dr. Daricélio Soares

Onde estamos?

- Tarefas Descritivas
  - Padrões e regras descrevem características importantes dos dados com os quais se está trabalhando.

- Mineração de Dados Indireta
  - Através de uma técnica de mineração, extraem-se padrões significativos que serão posteriormente avaliados.
  - O resultado da mineração complementa o conhecimento do especialista e deverá ser examinado e avaliado por este.

Regras de Associação (Transacionais)

- Uma regra de associação representa um padrão de relacionamento entre itens de dados de um domínio de aplicação que ocorre com uma determinada freqüência na base de dados.
- Seja I = {i1, i2, ..., in} o conjunto de itens do domínio da aplicação.
- Uma regra de associação R definida sobre I é uma implicação da forma

$$\mathbf{X \rightarrow Y}$$

- onde $X \subset I$, $Y \subset I$, $X \neq \emptyset$, $Y \neq \emptyset$ e $X \cap Y = \emptyset$. X é o antecedente da regra e Y é o conseqüente.

Regras de Associação (Transacionais)

$$X \rightarrow Y$$

$$X_1 \wedge X_2 \wedge \ldots \wedge X_n \rightarrow Y_1 \wedge Y_2 \wedge \ldots \wedge Y_m$$

{candidíase} → {pneumonia}

{café, leite} → {pão, manteiga, queijo}

A primeira regra indica, com um determinado grau de certeza, que se o paciente contraiu candidíase, então também teve pneumonia.

Medidas de Interesse

- Seja R a regra X → Y.
- Seja T o número de transações consideradas.
- Seja T (XUY) o número de transações que incluem os elementos de X U Y.

Suporte(R) = T(XUY) / T

Medidas de Interesse

- Seja R a regra $X \rightarrow Y$.
- Seja T_X o número de transações que incluem os elementos de X.
- Seja T_{XUY} o número de transações que incluem os elementos de X U Y.

$$\textbf{Confiança(R)} = T_{XUY} / T_X$$

Mineração de Regras de Associação

Input

Base de dados de transações (?);

Suporte mínimo;

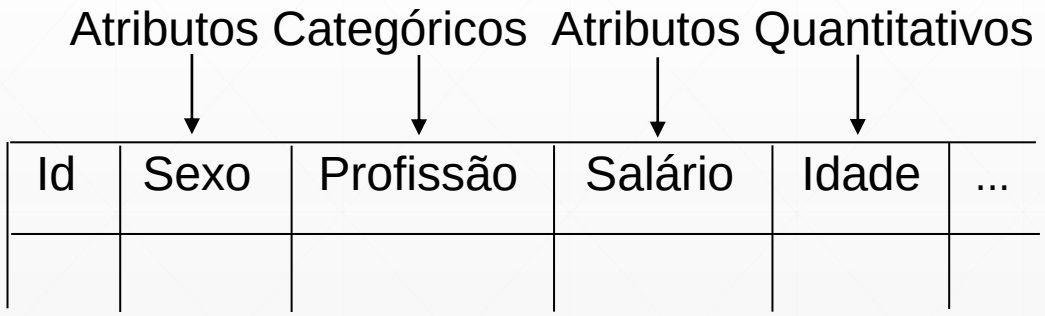
Confiança mínima.

Output

Todas as regras de associação que possuem suporte e confiança maiores ou iguais ao suporte e à confiança mínimos.

Regras de Associação Multidimensionais

- Regras de associação quantitativas são utilizadas quando se deseja minerar padrões em bases de dados relacionais (formadas por atributos quantitativos e atributos categóricos).



Regras de Associação Multidimensionais

- Exemplo (base de dados sobre a AIDS):

$$\textbf{(sexo=“M”)} \wedge \textbf{(20}\leq\textbf{idade}\leq\textbf{30)} \wedge \textbf{(opção-sex=“heterossex”)} \rightarrow \textbf{(usuário-drogas=“S”)}$$

- Esta regra indica, com confiança **C**, que pacientes aidéticos heterossexuais, entre 20 e 30 anos, do sexo masculino têm C**%** de chance de serem usuários de drogas.

Meta-Regras
Mineração Baseada em Restrições

- Permitem a especificação do tipo de regras que se deseja minerar.
- Podem funcionar como restrições definidas pelo usuário.
- Podem representar hipóteses a serem confirmadas

Meta-Regra

$$idade(X,[30,39]) \wedge renda(X,[4K,6K]) \rightarrow compras(Y,\text{“educational software”})$$

Relembrando

- Regras Transacionais

$$\{\text{Strogonoff de Frango (caixa)}\} \rightarrow \{\text{Lasanha (caixa)}\}$$

$$\{\text{Milho Verde em Conserva}\} \rightarrow \{\text{Ervilhas em Conserva}\}$$

- Regras Multidimensionais

$$(\text{Escolaridade} = \text{“Analfabeto”}) \wedge (\text{Sexo} = \text{“F”}) \wedge (\text{Idade} > 50) \rightarrow (\text{Ocupação} = \text{“Doméstica”})$$

Suporte e Confiança (*Agrawal et al, 1993*) Problemas?

- O número de regras gerado costuma ser extremamente volumoso. Identificar as regras realmente úteis e interessantes torna-se uma tarefa difícil.

- O modelo gera regras redundantes, ilusórias ou até mesmo contraditórias.

Exemplo

Grupo I: antecedente e conseqüente muito populares

R: Cenoura → Batata Sup: 70,38% Conf: 91,38%
Sup(Cenoura) = **77,01%**
Sup(Batata) = **81,75%**

Grupo II: antecedente pouco freqüente e conseqüente muito freqüente

R: Filé → Açúcar Refinado Sup: 7,58% Conf: 86,49%
Sup(Filé) = **8,77%**
Sup(Açúcar Refinado) = **86,49%**

Grupo III: antecedente e conseqüente pouco freqüentes

R: Strogonoff de Frango → Lasanha Sup: 3,32% Conf: 77,78%
Sup(Strogonoff) = **4,27%**
Sup(Lasanha) = **14,45 %**

Dependência entre Itens

Id	Regra de Associação	Sup_X	Sup_Y	Sup	Conf
R1	Filé → Açúcar Refinado	8,77%	86,49%	7,58%	**86,49%**

A confiança da regra indica que **86,49%** dos clientes que compram filé de viola, também compram açúcar refinado.

A probabilidade de qualquer cliente comprar açúcar refinado é de 86.49%.

Os dois produtos são **independentes**.

Sup(Y) = Conf(X → Y)

Dependência entre Itens

Id	Regra de Associação	Sup_X	Sup_Y	Sup	Conf
R2	Banana Nanica → Banana Prata	12,09%	**76,07%**	7,35%	**60,78%**

A confiança da regra indica que **60,78%** dos clientes que compram banana nanica, também compram banana prata.

A probabilidade de qualquer cliente comprar banana prata é de 76.07%. Portanto clientes que compram banana prata têm menor chance de comprar banana nanica.

Os dois produtos possuem **dependência negativa**

$$\mathbf{Sup(Y) > Conf(X \rightarrow Y)}$$

Dependência entre Itens

Id	Regra de Associação	Sup_X	Sup_Y	Sup	Conf
R3	Milho Verde em Conserva ⇒ Ervilhas em Conserva	32,94%	**37,91%**	27,01%	**82,01%**

A confiança da regra indica que **82,01%** dos clientes que compram milho verde, também compram ervilhas.

A probabilidade de qualquer cliente comprar ervilhas é de **37.91%.** Portanto clientes que compram milho verde têm uma chance muito maior de comprar ervilhas.

Os dois produtos possuem **dependência positiva**.

$$\mathbf{Sup(Y) < Conf(X \rightarrow Y)}$$

Medidas de Interesses Objetivas

Lift (X →Y) : indica o quanto mais freqüente torna-se B quando A ocorre.

Lift(X →Y) = Conf (X→Y) / Sup(Y)

Ex 1: Filé → Açúcar Refinado
Lift(X →Y) = 0,8649 **/** 0,8649 = **1**

Ex 2: R: Banana Nanica → Banana Prata
Lift(X →Y) = 0,6078 **/** 0,7607 = **0,80**

Ex 3: R: Milho Verde em Conserva → Ervilhas em Conserva
Lift(X →Y) = 0,8201 / 0,3791 = **2,21**

Medidas de Interesses Objetivas

Rule Interest(X →Y) : diferença entre o suporte real e o suporte esperado da regra.

$$RI(X \rightarrow Y) = Sup(X \rightarrow Y) - (Sup(X) \times Sup(Y))$$

Ex 1: R: Filé → Açúcar Refinado
RI(R) = 0,0758 – (0,0877 x 0,8649) = **0**

Ex 2: R: Banana Nanica → Banana Prata
RI(R) = 0,0735 – (0,1209 x 0,7607) = **-0,06**

Ex 3: R: Milho Verde em Conserva → Ervilhas em Conserva
RI(R) = 0,2701 – (0,3294 x 0,3791) = **0,14**

O RI também é conhecido como *Leverage* ou *OS*.

Questões?