CENTRO UNIVERSITÁRIO UNINORTE
CURSO DE PÓS-GRADUAÇÃO EM: Pós Graduação em Gerência de Banco de Dados.
DISCIPLINA: Mineração de Dados

Regras de Associação

Prof.º: Manoel Limeira
juniorlimeiras@gmail.com

Regras de Associação

- **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 frequência na base de dados.
- Seja D = $\{i_1, i_2, ..., i_n\}$ o conjunto de itens do domínio da aplicação.
- Uma regra de associação R definida sobre D é uma implicação da forma

$$\mathbf{X \Rightarrow Y}$$

onde $X \subset D$, $Y \subset I$, $X \neq \emptyset$, $Y \neq \emptyset$ e $X \cap Y = \emptyset$.

X é o antecedente da regra e Y é o consequente.

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

- Regras de associação possuem índices que indicam sua **relevância** e a **validade**.
- O **fator de suporte** de uma regra X ⇒ Y é definido pela porcentagem de transações que incluem todos os itens do conjunto X U Y.
- Representa a fração das transações que satisfazem tanto o antecedente quanto o consequente da regra.
- O suporte de uma regra tenta indicar sua relevância.

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

TID	Itens Comprados
101	leite, pão, suco
792	leite, suco
1130	leite, ovos
1735	pão, biscoito, café

Suporte({leite} ⇒ {suco}) = 2/4 = 50%

Suporte({suco} ⇒ {leite}) = 50%

Suporte({pão} ⇒ {suco}) = ?

Suporte({pão} ⇒ {ovos}) = ?

Suporte({pão,café} ⇒ {biscoito}) = ?

Medidas de Interesse

Seja X um conjunto de itens.

Seja T_X o número de transações que incluem os elementos de X.

Suporte(X) = T_X / T

TID	Itens Comprados	
		Suporte({leite}) = 3/4 = 75%
101	leite, pão, suco	Suporte({leite,suco}) = 2/4 = 50%
792	leite, suco	Suporte({pão,suco}) = ?
1130	leite, ovos	Suporte({pão,ovos}) = ?
1735	pão, biscoito, café	Suporte({pão,café,biscoito}) = ?

Medidas de Interesse

- O **fator de confiança** de uma regra $X \Rightarrow Y$ é definido pela porcentagem de transações que incluem os itens X e Y em relação a todas que incluem os itens de X.
- Representa o grau de satisfatibilidade do consequente, em relação às transações que incluem o antecedente.
- A confiança tenta indicar a **validade** da regra.

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.

Confiança(R) = T_{XUY} / T_X

Id-T.	Itens Comprados
101	leite, pão, suco
792	leite, suco
1130	leite, ovos
1735	pão, biscoito, café

Confiança({leite} ⇒ {suco}) = 2/3 = 67%
Confiança({suco} ⇒ {leite}) = 2/2 = 100%
Confiança({pão} ⇒ {suco}) = ?
Confiança({pão} ⇒ {ovos}) = ?
Confiança({pão,café} ⇒ {biscoito}) = ?

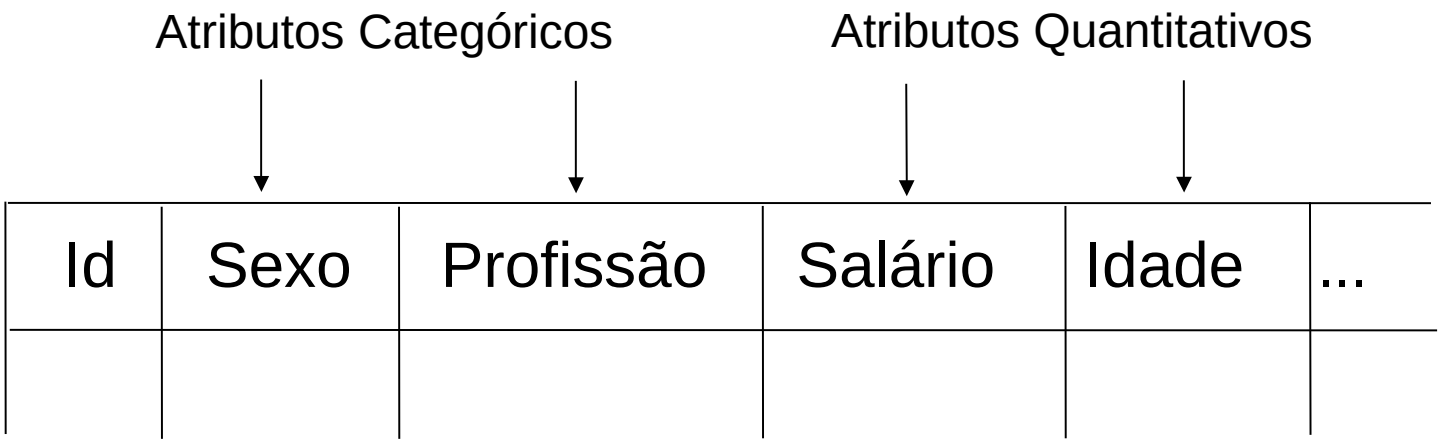
Mineração de Regras de Associação

- **Entrada:**
  - Base de dados de transações
  - Suporte mínimo
  - Confiança mínima
- **Saída:**
  - 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).

Tabelas = Relações
Atributos = Dimensões



Regras de Associação Multidimensionais

- Exemplo (base de dados sobre a AIDS):

(sexo="M") ^ (20≤idade≤30) ^ (opção-sex="heterossex") ⇒ (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.

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 consequente muito populares
  - R: Cenoura → Batata
  - Sup(Cenoura) = 77,01%
  - Sup(Batata) = 81,75%

Sup: 70,38%
Conf: 91,38%

- Grupo II: antecedente pouco frequente e consequente muito frequente
  - R: Filé → Açúcar Refinado
  - Sup(Filé) = 8,77%
  - Sup(Açúcar Refinado) = 86,49%

Sup: 7,58%
Conf: 86,49%

- Grupo III: antecedente e consequente pouco frequentes
  - R: Strogonoff de Frango → Lasanha
  - Sup(Strogonoff) = 4,27%
  - Sup(Lasanha) = 14,45 %

Sup: 3,32%
Conf: 77,78%

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**.

$$Sup(Y) = Conf(X \rightarrow Y)$$

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**

$$Sup(Y) > Conf(X \rightarrow Y)$$

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**.

$$Sup(Y) < Conf(X \rightarrow Y)$$

Medidas de Interesses Objetivas

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

$$Lift(X \rightarrow Y) = Conf(X \rightarrow Y) / Sup(Y)$$

- Ex 1: Filé → Açúcar Refinado

$$Lift(X \rightarrow Y) = 0{,}8649 / 0{,}8649 = \mathbf{1}$$

- Ex 2: R: Banana Nanica → Banana Prata

$$Lift(X \rightarrow Y) = 0{,}6078 / 0{,}7607 = \mathbf{0{,}80}$$

- Ex 3: R: Milho Verde em Conserva → Ervilhas em Conserva

$$Lift(X \rightarrow Y) = 0{,}8201 / 0{,}3791 = \mathbf{2{,}21}$$

CENTRO UNIVERSITÁRIO UNINORTE
CURSO DE PÓS-GRADUAÇÃO EM: Pós Graduação em Gerência de Banco de Dados.
DISCIPLINA: Mineração de Dados

Regras de Associação

Prof.º: Manoel Limeira
juniorlimeiras@gmail.com