

Regras de Associação

Prof. Sandro Jerônimo

Regras de Associação

Objetivo

Encontrar padrões frequentes, associações, correlações entre conjunto de itens ou objetos de um banco de dados transacional, banco de dados relacional ou outro repositório de informação.

Aplicações

- Análise de cestas de compras, marketing, portfólio de produto/serviços, etc.
- Quais subsequentes compras após ter comprado um PC?
- Qual tipo de DNA é sensível a uma nova droga?
- Turistas que optam em viajar para os EUA também aceitariam viajar para a Europa

Regras de Associação - Definições

Formato das Regras

- $A \rightarrow B$ $A \text{ e } B \rightarrow C$ $A \rightarrow B \text{ e } C$ $B \rightarrow A$...
- $A \rightarrow B$ significa: se **A** acontece em uma transação então B também acontece

Exemplos

- Quem compra computador (**A**) compra software (**B**) $A \rightarrow B$
- Quem compra computador (**A**) e software (**B**) compra webcam (**W**) $A \text{ e } B \rightarrow W$
- Quem opta por ir ao cinema (**C**) também compra refrigerante (**R**) e pipoca (**P**) $C \rightarrow R \text{ e } P$

Regras de Associação - Estatísticas

As regras possuem medidas estatísticas

- **Suporte:** probabilidade dos elementos acontecerem

$$\text{sup}(A \rightarrow B) = \frac{\text{número de transações com A e B}}{\text{número total de transações}}$$

- **Confiança:** confiança em que um elemento implica a outro





$$\text{conf}(A \rightarrow B) = \frac{\text{número de transações que suportam } (A \cup B)}{\text{número de transações que suportam A}}$$

- **EXEMPLO:** suponha a informação de clientes que compram computadores e tendem a comprar softwares de edição de texto. Na análise as medidas foram : suporte 2% e confiança 60%. O suporte significa que 2% de todas transações analisadas mostram que computador e software foram comprados juntos. Confiança de 60% indica que 60% dos clientes que compram computador também compram software.

*As medidas refletem a utilidade e o grau de certeza das regras. As regras que satisfazem aos dois valores mínimos são consideradas **fortes**.*

Identificando Regras Fortes



ID Transação	Itens das compras	
2000		A, B, C
1000		A, C
4000		A, D
5000		B, E, F

Usuário deve definir o suporte e confiança mínima. Exemplo:

- Suporte: 50%
- Confiança: 50%

Teremos as seguintes regras

- $A \rightarrow C$
Suporte: $2/4 = 0.5 = 50\%$
Confiança: $2/3 = 0.66 = 66.6\%$
- $C \rightarrow A$
Suporte: $2/4 = 0.5 = 50\%$
Confiança: $2/2 = 1 = 100\%$

Produto	Núm. do Produto
Pão	1
Açúcar	2
Leite	3
Papel Higiênico	4
Manteiga	5
Fralda	6
Cerveja	7
Refrigerante	8
Iogurte	9
Suco	10

Exemplo - Compras

Num transação	Itens comprados
T1	{1,3,5}
T2	{2,1,3,7,5}
T3	{4,9,2,1}
T4	{5,2,1,3,9}
T5	{1,8,6,4,3,5}
T6	{9,2,8}



Um mercado planeja criar espaços para oferta “casada” de produtos. Quais de produtos fariam parte desses espaços assumindo um suporte de 50%?

Alguns *Itemset* selecionados

Num transação	Itens comprados
T1	{1,3,5}
T2	{2,1,3,7,5}
T3	{4,9,2,1}
T4	{5,2,1,3,9}
T5	{1,8,6,4,3,5}
T6	{9,2,8}

Produto	Núm. do Produto
Pão	1 <i>(apareceu 5x)</i>
Açúcar	2 <i>(apareceu 4x)</i>
Leite	3 <i>(apareceu 4x)</i>
Manteiga	5 <i>(apareceu 4x)</i>
logurte	9 <i>(apareceu 3x)</i>

<i>Itemset</i>	Suporte ($\geq 50\%$) e Confiança
{Pão, Leite} {1,3}	Suporte = $4/6 = 66,6\%$ Confiança (Pão \rightarrow Leite) = 80% Confiança (Leite \rightarrow Pão) = 100 %
{Pão, Manteiga} {1,5}	Suporte = $4/6 = 66,6\%$ Confiança (Pão \rightarrow Mant.) = 80% Confiança (Mant. \rightarrow Pão) = 100%
{Leite, logurte} {2,9}	Suporte = $3/6 = 50\%$ Confiança (Leite \rightarrow logurte) = 75% Confiança (logurte \rightarrow Leite) = 100%
{Pão, Leite, Manteiga} {1, 3, 5}	Suporte = $3/6 = 50\%$ Confiança (Pão \rightarrow Leite, Mant) = 60% Confiança (Leite \rightarrow Pão, Mant) = 100% Confiança (Mant \rightarrow Pão, Leite) = 60% Confiança (Pão, Leite \rightarrow Mant) = 75% Confiança (Pão, Mant \rightarrow Leite) = 75% Confiança (Leite, Mant \rightarrow Pão) = 75%

Algoritmo *Apriori*

O algoritmo mais usado para encontrar regras de associação. Baseia-se no fato usar conhecimento já obtidos dos *itemsets* anteriores.

- **Fase I** - Descubra todos os conjuntos de itens com suporte maior ou igual ao mínimo suporte especificado pelo usuário.
- **Fase II** - A partir dos conjuntos de itens frequentes, descubra regras de associação com fator de confiança maior ou igual ao especificado pelo usuário.

Algoritmo Apriori – Ilustração

Suporte mínimo = 2 (50%)

Base de Dados D

TID	Itens
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D

itemset	Sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

X

L_1

Itemset	Sup.
{1}	2
{2}	3
{3}	3
{5}	3

L_2

itemset	Sup.
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_2

itemset	Sup.
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

X

X

Scan D

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

C_3

itemset
{2 3 5}

Scan D

L_3

itemset	Sup.
{2 3 5}	2

Existem outras medidas de interesse

Lift: indica a força de uma regra sobre a coocorrência aleatória de seus antecedentes e consequentes.

$$lift(A \rightarrow B) = \frac{sup(A \rightarrow B)}{sup(A) \times sup(B)}$$

Convicção: Assim como a confiança, é sensível à direção da regra.

$$conv(A \rightarrow B) = \frac{1 - sup(B)}{1 - conf(A \rightarrow B)}$$

Ganho: Ganho é calculado baseado em um valor theta (θ) dado. Usualmente $\theta = 2.0$

$$ganho(A \rightarrow B) = sup(A \cup B) - \theta * sup(A)$$

Laplace: Laplace é calculado baseado em um parâmetro k . Usualmente $k = 1.0$.

$$laplace(A \rightarrow B) = \frac{sup(A \cup B) + 1}{sup(A) + k}$$

Piatetsky-Shaprio (P-S):

$$ps(A \rightarrow B) = sup(A \cup B) - sup(A) * sup(B)$$



PUC Minas Virtual