Universidade Técnica de Lisboa

INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO

Informática e Sistemas de Informação Aplicados em Economia

Descoberta de Conhecimento em Bases de Dados.

Associações

Descoberta de Conhecimento em Bases de Dados. Classificação e Associações

- **Associações**
- **Conjuntos frequentes**
- **Algoritmos para a descoberta**

A Pesquisa de Associações

- **Regras de Associação**
- **Espaços para aplicação**
- **Frequências**
- **Descoberta de Subconjuntos Frequentes**
- **Geração de Regras de Associação**

A Pesquisa de Associações

A **pesquisa de associações** ou **análise de afinidades** é a procura de padrões e condições que descrevem como vários items se agrupam juntos ou acontecem juntos em séries de eventos ou transacções.

Uma **regra de associação** ou **afinidade** tem a forma:

Quando *Item1* ***Também*** *Item2*

O problema considerado é encontrar regras de associação a partir de dados binários que por seu lado terão sido obtidos a partir de ficheiros de transacções organizados de acordo com os modelos usuais, nomeadamente o modelo relacional.

Regras de Associação

Assuma-se que temos um **conjunto $R = \{A_1,\ldots,A_p\}$** de atributos binários, isto é, o domínio de cada A_i é $\{0,1\}$.

Uma **relação** $r = \{t_1,\ldots t_n\}$ no esquema R é uma matriz com colunas R and n linhas, sendo cada linha um vector de comprimento p cujos elementos são 0 e 1.

Uma ***regra de associação*** em r é uma expressão da forma $X \rightarrow B$, onde $X \subseteq R$ e $B \in R \setminus X$. O significado intuitivo da regra é que se uma linha da matriz r tem um 1 em cada coluna de X, então a linha tende a ter um 1 também na coluna B.

Exemplos de Espaços para Regras de Associação

- Uma base de dados de estudantes numa universidade: as linhas correspondem a estudantes, as colunas a cursos, e um 1 na posição *(e, c)* indica que o estudante *e* frequentou o curso *c*.
- Dados recolhidos a partir de leitores de código de barras em supermercados: as colunas correspondem a produtos, e cada linha corresponde ao conjunto de produtos comprado uma vez.
- Uma base de dados sobre o IDE: as linhas correspondem a investidores, as colunas a sectores e um 1 na posição *(i,s)* significa que houve investimento de *i* no sector *s*.

Frequências

- Dado $W \subseteq R$, representamos por *s(W, r)* a *frequência* de *W* em *r*: a fracção de linhas de *r* que tem um 1 em cada coluna de *W*.
- A *frequência da regra* $X \rightarrow B$ em *r* é definida por $s(X \cup \{B\}, r)$, e a *confiança da regra* é $s(X \cup \{B\}, r) / s(X, r)$.

Na ***descoberta de regras de associação****, a tarefa é encontrar todas as regras* $X \rightarrow B$ *tais que*

- *a frequência da regra seja pelo menos um valor dado* σ *e*
- *a confiança seja pelo menos igual a outro valor* θ.

Limitações

Não há limites predefinidos ao número de atributos do lado esquerdo X duma regra de associação $X \rightarrow B$

- isto é importante para que associações não esperadas não sejam desprezadas antes de o processamento se iniciar.
- espaço de pesquisa tem um tamanho exponencial no número de elementos da relação de input o que requer algum cuidado com os algoritmos de tratamento.

Subconjuntos Frequentes

Um subconjunto $X \subseteq R$ é *frequente em r*, se $s(X, r) \geq \sigma$.

Uma vez conhecidos todos os conjuntos frequentes de r, encontrar as regras de associação é fácil.

Concretamente, para cada conjunto frequente X e cada $B \in X$ verificar se a regra $X \setminus \{B\} \rightarrow B$ tem uma confiança suficientemente alta.

A descoberta de todos os conjuntos frequentes pode ser feita de muitos modos diferentes. Uma abordagem típica é usar o facto de que todos os subconjuntos de um conjunto frequente são também frequentes.

Descoberta de Subconjuntos Frequentes

- **Descobrir os conjuntos frequentes de tamanho 1** lendo os dados uma vez e registando o número de vezes que cada atributo *A* ocorre.
- **Formar conjuntos de tamanho 2** tomando todos os pares de atributos *{B,C}* tais que *{B}* e *{C}* sejam ambos frequentes.
- **Avaliar a frequência** dos conjuntos candidatos relativamente à base de dados.
- **Formar os candidatos de tamanho 3**: estes são conjuntos *{B,C,D}* tais que *{B,C}*, *{B,D}*, e *{C,D}* sejam todos frequentes.
- o processo continua até que não possam ser formados mais conjuntos candidatos.

Descoberta de Subconjuntos Frequentes

Transacções do Período

Número de transação	Código de Produto
1	A
1	B
1	C
2	A
2	C
3	A
4	C
4	B
4	D
4	A
5	A

Número de transação	Código de Produto
5	D
6	B
7	A
8	A
8	C
8	D
9	B
9	D
10	A
10	B
10	D

Descoberta de Subconjuntos Frequentes

Produtos Envolvidos em Transacções

A partir do ficheiro é possível

- definir o conjunto *R={A,B,C,D}* e
- a relação *r* que terá o seguinte aspecto

Número de transacção	A	B	C	D
1	1	1	1	0
2	1	0	1	0
3	1	0	0	0
4	1	1	1	1
5	1	0	0	1
6	0	1	0	0
7	1	0	0	0
8	1	0	1	1
9	0	1	0	1
10	1	1	0	1

Descoberta de Subconjuntos Frequentes

Frequência dos Produtos - 1

Conjunto	Frequência absoluta
{A}	8
{B}	5
{C}	4
{D}	5

Admitam-se os seguintes valores: σ = 0.1 e θ = 0.8. Com base nestes valores todos os subconjuntos são frequentes pelo que todos os formados por 2 elementos são candidatos a frequentes.

Descoberta de Subconjuntos Frequentes

Frequência dos Produtos - 2

Conjunto	Frequência absoluta
{A,B}	3
{A,C}	4
{A,D}	4
{B,C}	2
{B,D}	3
{C,D}	2

Descoberta de Subconjuntos Frequentes

Quadro de Regras Potenciais - 1

Regra	Confiança
A→B	3/8
A→C	4/8
A→D	4/8
B→C	2/5
B→D	3/5
C→D	2/4

$\sigma = 0.1$ e
$\theta = 0.8$

Regra	Confiança
B→A	3/5
C→A	**4/4**
C→B	2/4
D→A	**4/5**
D→B	3/5
D→C	2/5

Associações: C→A e D→A

Descoberta de Subconjuntos Frequentes

$C \rightarrow A$

Com confiança de 100%

Se Produto C

Também Produto A

$D \rightarrow A$

Com confiança de 80%

Se Produto D

Também Produto A

Descoberta de Subconjuntos Frequentes

Frequência dos Produtos - 3

Conjunto	Frequência absoluta
{A,B,C}	2
{A,C,D}	2
{A,B,D}	2
{B,C,D}	1

Como considerámos σ = 0.1 qualquer dos subconjuntos é frequente pelo que {A,B,C,D} também o é. Da tabela de frequências anterior podem deduzir-se as associações possíveis.

Descoberta de Subconjuntos Frequentes

Quadro de Regras Potenciais - 2

Regra	Confiança
A,B→C	2/3
A,C→B	2/4
A,B→D	2/3
A,D→B	2/4
A,C→D	2/4
A,D→C	2/4

Regra	Confiança
B,C→D	1/ 2
B,D→C	1/3
B,C→A	**2/2**
B,D→A	2/3
C,D→A	**2/2**
C,D→B	½

Associações: B,C→A e C,D→A

Descoberta de Subconjuntos Frequentes

$B, C \rightarrow A$

Com confiança de 100%

Se Produto B e Produto C

Também Produto A

$C, D \rightarrow A$

Com confiança de 100%

Se Produto C e Produto D

Também Produto A

Descoberta de Subconjuntos Frequentes

Quadro de Regras Potenciais - 3

Regra	Confiança
A,B,C→D	½
A,C,D→B	1/2
A,B,D→C	½
B,C,D→A	**1/1**

Com confiança de 100%

Se Produto B e Produto C e Produto D

Também Produto A