Associações
&
Análises de Itens Freqüentes

Eduardo R. Hruschka

Baseado no curso de Gregory Piatetsky-Shapiro, disponível no sítio http://www.kdnuggets.com

Visão Geral:

- Transações
- Itens freqüentes
- Regras de Associação
- Aplicações

Exemplos de transações:

ID	Produto
1	leite, pão, ovos
2	pão, açúcar
3	pão, cereal
4	leite, pão, açúcar
5	leite, cereal
6	pão, cereal
7	leite, cereal
8	leite, pão, cereal, ovos
9	leite, pão, cereal

Exemplo de base de dados de transações:

T	Produtos
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

Exemplos (*Instances*) = transações

Itens:

A = leite
B= pão
C= cereal
D= açúcar
E= ovos

Exemplo de Banco de dados de transações:

T	Produtos
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

Produtos convertidos em atributos binários:

T	A	B	C	D	E
1	1	1	0	0	1
2	0	1	0	1	0
3	0	1	1	0	0
4	1	1	0	1	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	0	1	0	0
8	1	1	1	0	1
9	1	1	1	0	0

Definições:

- Conjunto de itens I : um subconjunto de itens possíveis.
  - Exemplo: I = {A,B,E} (ordem não é importante)
- Transação (T): conjunto de itens.
- Suporte (I) = nº de transações t que contém I.
  - Na base de dados anterior temos que: sup({A,B,E})=2, sup({B,C})= 4
- Conjunto de itens freqüentes: sup(I)≥*sup_mín,* onde *sup_mín* é o suporte mínimo, definido pelo usuário.

Propriedade do subconjunto:

- **Todo subconjunto de um conjunto freqüente é também freqüente.**
- Por quê?
- Exemplo: suponhamos que {A,B} seja freqüente. Dado que cada ocorrência de {A,B} inclui A e B, então A e B tem de ser eles próprios freqüentes.
- Quase todos os algoritmos para extrair regras de associação são baseados nesta propriedade.

Regras de Associação:

- Regra de associação *R* :
  - Conjunto de itens *1* $\Rightarrow$ Conjunto de itens *2.*
  - Conjuntos de itens disjuntos, e conjunto de itens 2 é não vazio.
  - *Tradução:* se determinada transação inclui o conjunto de itens 1, então esta também inclui o conjunto de itens 2.
- Exemplos:
  - A,B $\Rightarrow$ E,C
  - A $\Rightarrow$ B,C

Como obter regras de associação?

- *Dado um conjunto de itens freqüentes {A,B,E}, quais são as possíveis regras de associação?*
- *Exemplos:*
  - A $\Rightarrow$ B; B $\Rightarrow$ A; A $\Rightarrow$ E; etc.
  - A $\Rightarrow$ {B, E}
  - {A, B} $\Rightarrow$ E
  - {A, E} $\Rightarrow$ B
  - B $\Rightarrow$ {A, E}
  - __ $\Rightarrow$ A,B,E (regra vazia).

Classificação X Regras de Associação:

Regras de classificação

- Focada no *campo alvo* (atributo meta);
- Especificam as classes em todos os casos;
- Medida: precisão (*Accuracy*)

Regras de associação

- Muitos *campos alvos* (combinações de atributos meta)
- Medidas: em geral suporte e confiança.

Suporte e Confiança:

- Suponhamos que *R: $I \Rightarrow J$* seja uma regra de associação.
  - sup(R) = sup(I $\cup$ J)
    - A união dos conjuntos I $\cup$ J define o suporte.
  - conf (R) = sup(R) / sup(I) é a confiança de R
    - Número de transações que possuem I e J dividido pelo número de transações que possuem I.
- Regras de associação com suporte mínimo são às vezes chamadas de "regras fortes".

Exemplo de Regras de Associação formadas por três itens:

- ***Dado um conjunto de itens freqüentes {A,B,E}, quais regras de associação possuem sup_mín=2 e conf_mín=50% ?***

  A, B $\Rightarrow$ E : conf=2/4 = 50%

  A, E $\Rightarrow$ B : conf=2/2 = 100%

  B, E $\Rightarrow$ A : conf=2/2 = 100%

  E $\Rightarrow$ A, B : conf=2/2 = 100%

Confiança é menor do que a requerida:

A $\Rightarrow$ B, E : conf=2/6=33% < 50%

B $\Rightarrow$ A, E : conf=2/7=28% < 50%

__ $\Rightarrow$ A,B,E : conf: 2/9 = 22% < 50%

T	Lista de itens
1	**A, B, E**
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	**A, B, C, E**
9	A, B, C

Exemplo (Freitas & Lavington):

Se “café” então “pão”; Suporte=0,3 / Confiança=1.

Se “café” então “manteiga”; Suporte=0,3 / Confiança=1.

Se “pão” então “manteiga”; Suporte=0,4 / Confiança=0,8.

Se “café E pão” então “manteiga”; Suporte=0,3 / Confiança=1.

T	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
1	N	S	N	S	S	N	N
2	S	N	S	S	S	N	N
3	N	S	N	S	S	N	N
4	S	S	N	S	S	N	N
5	N	N	S	N	N	N	N
6	N	N	N	N	S	N	N
7	N	N	N	S	N	N	N
8	N	N	N	N	N	N	S
9	N	N	N	N	N	S	S
10	N	N	N	N	N	S	N

Encontrando regras de associação:

- Conforme visto, uma regra possui dois parâmetros: *minsup* e *minconf*;
  - sup(R)≥*sup_mín* & conf(R)≥*conf_mín*
- Problema:
  - Encontrar todas as regras que forneçam *sup_mín* e *conf_mín* .
- Inicialmente, encontrar todos os itens freqüentes.
  - Procurar por conjuntos formados por um item.
  - Procurar por conjuntos formados por dois itens entre aqueles obtidos no passo anterior; e assim sucessivamente (lembrar da propriedade dos subconjuntos).

Algoritmo Apriori (Agrawal & Srikant):

Idéia: usar conjuntos de um item para formar conjuntos de dois itens, usar conjuntos de dois itens para formar conjuntos de três itens, e assim por diante.

- Se {A,B} é um conjunto de itens freqüentes, então {A} e {B} também são conjuntos de itens freqüentes.
- Se X é um conjunto de itens freqüentes formado por *k* itens, então todos os (*k*-1) itens de X também são freqüentes.

$\Rightarrow$ Obter conjuntos formados por *k* itens pela união de conjuntos de (*k*-1) itens.

Um exemplo:

- Dados 5 conjuntos freqüentes de 3 itens:

  `(A B C), (A B D), (A C D), (A C E), (B C D)`

- Ordenação melhora a eficiência computacional;
- Conjunto candidato formado por 4 itens:

  `(A B C D)`

→ Pode ser freqüente, pois todos os seus subconjuntos de 3 itens são freqüentes.

→ E o que dizer sobre o conjunto `(A C D E)`?

→ Como (C D E) não é freqüente, o conjunto {A C D E} também não é freqüente.

Gerando regras de Associação:

- Processo de dois estágios:
  - Determinar os itens mais freqüentes;
  - Para cada subconjunto de itens freqüentes I fazer:
    - Para cada subconjunto J de I:
      - Determinar todas as regras de associação da forma:

        $(I\text{-}J) \Rightarrow J$
- Principal idéia usada em ambos os estágios: propriedade dos subconjuntos.

Continuando o exemplo anterior (Freitas & Lavington)...

Se “café” então “pão”; Suporte=0,3 / Confiança=1.

Se “café” então “manteiga”; Suporte=0,3 / Confiança=1.

Se “pão” então “manteiga”; Suporte=0,4 / Confiança=0,8.

Se “café E pão” então “manteiga”; Suporte=0,3 / Confiança=1.

R	Leite	Café	Cerveja	Pão	Manteiga	Arroz	Feijão
1	N	S	N	S	S	N	N
2	S	N	S	S	S	N	N
3	N	S	N	S	S	N	N
4	S	S	N	S	S	N	N
5	N	N	S	N	N	N	N
6	N	N	N	N	S	N	N
7	N	N	N	S	N	N	N
8	N	N	N	N	N	N	S
9	N	N	N	N	N	S	S
10	N	N	N	N	N	S	N

- FASE I: descobrir conjuntos de itens freqüentes - suporte $\geq$ sup_mín;
- FASE II: descobrir regras com alto fator de confiança – confiança $\geq$ conf_mín;

- Passo 1 – suporte p/ conjuntos com 1 item:
  - Leite=0,2; Café=0,3; Cerveja=0,2; Pão=0,5; Manteiga=0,5; Arroz=0,2; Feijão=0,2.
  - Considerando sup_mín=0,3 $\Rightarrow$ café, pão, manteiga seriam os itens freqüentes!

Passo 2 – suporte p/ conjuntos com 2 itens:

$\Rightarrow$ Procurar considerando somente os itens mais freqüentes: café, pão e manteiga.

Café, pão $\Rightarrow$ suporte=0,3;

Café, manteiga $\Rightarrow$ suporte=0,3;

Manteiga, pão $\Rightarrow$ suporte=0,4;

Conjunto de itens freqüentes para sup_mín $\geq$ 0,3:

{café, pão}, {café, manteiga}, {manteiga, pão}.

Passo 3 – suporte p/ conjuntos com 2 e 3 itens:

{Café, pão, manteiga} suporte=0,3;

Calcula-se, então, a confiança das regras candidatas:

Se "x" então "y": |x e y| ÷ |x|

a) {café, pão} :

Se "café" então "pão" – conf.=1,0;

Se "pão" então "café" – conf.=0,6.

b) {café, manteiga} :

Se "café" então "manteiga" – conf.=1,0;

Se "manteiga" então "café" – conf.=0,6.

c) {manteiga, pão} :

Se "manteiga" então "pão" – conf.=0,8;

Se "pão" então "manteiga" – conf.=0,8.

d) {café, manteiga, pão} :

Se "café, pão" então "manteiga" – conf.=1,0;

Se "café, manteiga" então "pão" – conf.=1,0;

Se "manteiga, pão" então "café" – conf.=0,75

Se "café" então "pão,manteiga" – conf.=1,0;

e assim por diante, escolhendo-se depois as regras que respeitam conf_mín.

"Filtrando" regras de associação:

- Problema: grandes bases de dados (e.g. supermercados) podem produzir um número elevado de regras de associação, mesmo com valores razoáveis para suporte e confiança...
- Um detalhe sobre a confiança:
  - Se todas as transações incluem Z, então:
  - Qualquer regra I $\rightarrow$ Z terá confiança igual a 100% !
- Filtrar regras ou pré-processar a base...
- Medidas de interesse objetivas e subjetivas: tópico efervescente de pesquisa!