DATA EXPLORATION

Méthodes descriptives : Les règles d’association

OBJECTIF

Il s'agit de regrouper des objets qui vont naturellement ensemble pour définir des règles d'association du type

SI *condition* ALORS *résultat*

Les règles d'association sont traditionnellement liées au secteur de la distribution, avec comme principale application «l'analyse du panier de la ménagère», c-a-d la recherche d'associations entre produits sur les tickets de caisse. La méthode recherche quels produits tendent à être achetés ensemble.

Les données forment un ensemble de *m* transactions

Ticket 1	Ticket 2	Ticket 3	Ticket 4
Farine	Œufs	Œufs	Œufs
Sucre	Farine	Sucre	Chocolat
Lait	Sucre	Chocolat	Thé
Œufs	Chocolat		

A partir de ces données, on cherche des règles du type

Farine → Sucre / Sucre → Farine

Œufs, Sucre → Chocolat

La méthode peut être appliquée à tout autre secteur d'activité.

CONFIANCE ET SUPPORT

Chaque transaction est constitué d'items. Ces items forment un ensemble de *n* éléments I={Farine, Sucre, Lait, Chocolat, Œufs,Thé}.

On construit un tableau récapitulatif des items par transaction

	Farine	Sucre	Lait	Chocolat	Œufs	Thé
Ticket 1	X	X	X		X	X
Ticket 2	X	X		X	X	
Ticket 3		X		X	X	
Ticket 4				X	X	X

Un tel tableau permet de déterminer la fréquence d'un produit,

$$Fréquence = \frac{nb\ de\ \ fois\ \ acheté}{nb\ \ d'achats\ \ total}$$

ou bien la fréquence à laquelle deux ou plusieurs produits se rencontrent dans un

Le support d'une règle X ⟶ Y est la fréquence à laquelle les items X et Y apparaissent simultanément dans les transactions.

Ex. Sup(Farine ⟶ Sucre) = 50% = Sup (Sucre ⟶ Farine)
Les produits Farine et Sucre apparaissent dans 50% des transactions

La confiance d'une règle X ⟶ Y est le rapport entre son support et la fréquence de X

Ex. Conf(Farine ⟶ Sucre) = 100% / Conf (Sucre ⟶ Farine)=67%
100% des transactions contenant le produit Farine contiennent aussi le produit Sucre
67% des transactions contenant le produit Sucre contiennent aussi le produit Farine

LE *LIFT* OU AMÉLIORATION

Dans l'exemple précédent, on calcule le support et la confiance de la règle :
{Sucre, Œufs} $\rightarrow$ Chocolat

X		Y	support	Confiance
Sucre	Œufs	Chocolat	50%	67%

Donc si Sucre et Œufs apparaissent dans un ticket alors il y a 67% de chance de voir aussi Chocolat.

Cependant Chocolat apparait dans 75% des tickets, donc la règle n'est d'aucune utilité!

Le lift ou amélioration permet de comparer la prédiction d'un résultat avec une règle ou sans règle,

Lift=confiance/support(résultat)

- Lift >1, la règle améliore la prédiction
- Lift <1, la règle ne sert à rien

Ex. Lift({Sucre, Œufs} $\rightarrow$ Chocolat) =0,89
Lift({Farine, Sucre} $\rightarrow$ Lait) =4

Remarques :

$$Lift(X \rightarrow Y) = \frac{Conf(X \rightarrow Y)}{freq(Y)} = \frac{freq(X,Y)}{freq(X)freq(Y)}$$

$$Lift(X \rightarrow Y) = Lift(Y \rightarrow X)$$

Algorithme de Construction des règles

Un algorithme naïf de construction consiste à procéder de la façon suivante :

- Déterminer les sous-ensembles de I (l'ensemble des items) de support supérieur à S_0, appelés sous-ensembles fréquents
- Construire les règles ayant un niveau de confiance supérieur à C_0
- Garder les règles ayant un lift supérieur à 1

Si l'ensemble I contient *n* éléments, le nombre de sous-ensembles possibles est $\mathbf{2^n}$!!!

Il est impossible de tous les lister. On procède donc par élimination

- On supprime de I tous les items ayant une fréquence inférieure à S_0
- On construit tous les sous-ensembles de cardinal 2 et on ne retient que les sous-ensembles fréquents pour l'étape suivante
- On construit tous les sous-ensembles de cardinal 3 et on ne retient que les sous-ensembles fréquents pour l'étape suivante, etc …
- A partir des sous-ensembles fréquents retenus dans les étapes précédentes, on construit les règles d'associations et on retient celles qui ont une confiance supérieure à C_0
- On calcule le lift sur les règles restantes

APPLICATION DE L'ALGORITHME

En général, on note

- C_k l'ensemble des sous-ensembles d'items *candidats* de cardinal k
- F_k l'ensembles des sous-ensembles d'items *fréquents* de cardinal k

Application de l'algorithme avec S_0=30%

Item 1	Fréq.
Sucre	¾
Farine	½
~~Lait~~	~~¼~~
Œufs	1
Chocolat	¾
~~Thé~~	~~¼~~

Item 1	Item 2	Fréq.
Sucre	Farine	½
Sucre	Œufs	¾
Sucre	Chocolat	½
Farine	Œufs	½
~~Farine~~	~~Chocolat~~	~~¼~~
Œufs	Chocolat	¾

Item 1	Item 2	Item 3	Fréq.
~~Sucre~~	~~Farine~~	~~Œufs~~	~~¼~~
Sucre	Chocolat	Œufs	½

C_1={{Sucre},{Farine},{Lait},{Œufs},{Chocolat},{Thé}}
F_1={{Sucre},{Farine},{Œufs},{Chocolat}}
C_2={{Sucre,Farine},{Sucre,Œufs},{Sucre,Chocolat},{Farine,Œufs},{Farine,Chocolat},{ Œufs,Chocolat}}
F_2={{Sucre,Farine},{Sucre,Œufs},{Sucre,Chocolat}, {Farine,Œufs},{ Œufs,Chocolat}}
C_3={{Sucre,Farine,Œufs},{Sucre,Œufs,Chocolat}}
F_3= {{Sucre,Œufs,Chocolat}}

On a éliminé beaucoup de cas mais il reste tout de même toutes les règles suivantes :

X	Y	Freq X	Freq Y	Freq X&Y	Conf.	Lift
Sucre	Chocolat	¾	¾	½	0.67	0.89
Chocolat	Sucre	¾	¾	½	0.67	0.89
Sucre	Oeufs	¾	1	¾	1	1
Œufs	Sucre	1	¾	¾	0.75	1
Chocolat	Œufs	¾	1	¾	1	1
Œufs	Chocolat	1	¾	¾	0.75	1
…						

X	Y	Freq X	Freq Y	Freq X&Y	Conf.	Lift
Sucre, Chocolat	Œufs	½	1	½	1	1
Œufs	Sucre, Chocolat	1	½	½	0.5	1
Sucre, Œufs	Chocolat	¾	¾	½	0.67	0.89
Chocolat	Sucre, Œufs	¾	¾	½	0.67	0.89
Chocolat, Œufs	Sucre	¾	¾	½	0.67	0.89
Sucre	Chocolat, Œufs	¾	¾	½	0.67	0.89

ALGORITHME APRIORI

Algorithme naïf Apriori

```
Initialisation de L1
Pour ( k=2; Lk-1≠∅;k++) faire
        Créer Ck
        Calculer le support des items de Ck
        En déduire Lk
Fin
Retourner les Lk
```

Complexité de l'algorithme Apriori

Pour chaque k :

- La création de C_k nécessite $O(card^2(L_{k-1}))$
- La construction de L_k nécessite la lecture complète du tableau des données

Il existe des améliorations de l'algorithme de construction des règles d'association permettant de travailler sur de grosses bases de données

R. Agrawal et R. Srikant. Fast algorithms for mining association rules in large databases. In proceedings of the 20th international conference on Very Large Data Bases (VLDB'94) , pages 478-499, 1994.

Avantages

- ✓ Résultats clairs : les règles d'association sont faciles à interpréter. Elles sont faciles à utiliser pour des utilisations concrètes.
- ✓ Apprentissage non supervisé : la méthode ne nécessite pas d'autre information qu'une classification en items et la donnée d'une liste d'items pour extraire les règles.
- ✓ Simplicité de la méthode : la méthode et les calculs sont élémentaires, elle peut être programmée sur un tableur.

Inconvénients

- ✓ Coût de la méthode : la méthode est coûteuse en temps de calcul. Le regroupement d'articles permet de diminuer les calculs mais on peut alors éliminer malencontreusement des règles importantes
- ✓ Les items rares : la méthode est plus efficace pour les items fréquents
- ✓ La qualité des règles : la méthode peut produire des règles triviales ou inutiles.

Avez-vous des questions?

Documents ayant servis à la rédaction des slides et TD :

- *Extraction de règles d'association à partir de données, Y. Lechevallier, E. Diday, Y. Chevaleyre*
- *Groupement par similitudes, R. Chelouah, H. de Milleville*