Les Règles d'association

HAMID NECIR | COURS FOUILLE DE DONNÉES

Définition

- La fouille de données (data mining) vise à découvrir, dans les grandes quantités de données, les informations précieuses qui peuvent aider à comprendre les données ou à prédire le comportement des données futures.

- Le datamining utilise depuis sont apparition plusieurs outils de statistiques et d'intelligence artificielle pour atteindre ses objectifs.

Introduction

- Rechercher des régularités dans les données
- Trouver des schémas fréquents, des associations, des corrélations parmi des ensembles d'articles dans des bases de données de transactions
- Comprendre les habitudes d'achat des clients en trouvant des associations et corrélations entre les différents achats que les clients placent dans leur "panier"

Applications :

- Rayonnage, mailing,...
- Analyse des données du panier,
- marketing croisé,
- conception de catalogue,
- analyse de journal Web, détection de fraude

Etapes de l'approche

Approche en deux étapes:

1. **Extraction des itemsets fréquents (Motifs)**

- Extraction de tous les ensembles d'éléments qui apparaissent fréquemment

2. Génération de règles

- Générez des règles de confiance élevées à partir de chaque ensemble d'éléments fréquents. déjà extrait

Etape 1: génération des itemsets fréquents

Etape 1: génération des itemsets fréquents

Conceptsde base

Soit I = {i1,i2,..., in} un ensemble d'éléments appelés items.

- Une base de données D consiste en un ensemble de transactions Ti tel que: Ti ⊆ I
- On dit que T contient X si X ⊆ T
- Item (motif) : Correspond à une valeur d'un attribut dans la base de transactions,
- Itemset : un itemset X est un ensemble d'items X ⊆ I
- K-itemset : est un itemset de longueur k (formé de k items).

- Exemple :

Pour les items {A, B, C} on a:

les 1-itemset: A B C
les 2-itemset: AB AC BC
les 3-itemsets: ABC

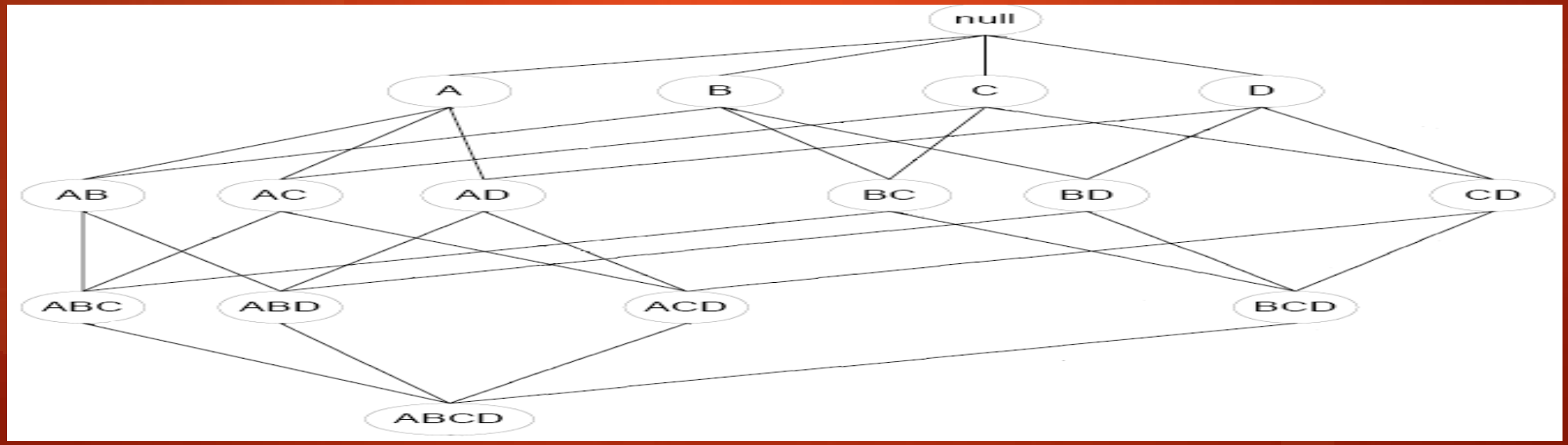
Problèmes

Générer tous les itemsets est trés complexe.

N items ⟹ $2^N - 1$ itemsets possible

- **Comment garder en mémoire un nombre important d'itemsets ?**
  - **100 items => 2100 -1 itemsets possibles !!!!**
- **Comment calculer le fréquence d'apparition d'un nombre important d'itemsets dans une grande base de données (100 million de transactions) ?**

Exemple : Les itemsets possibles pour pour 3 items {A, B, C}.



Mesures utilisées

- **Support :** facteur de fiabilité de la règle .

Le support d'un itemset X noté Sup(X) est la proportion de transaction de D contenant X.

$$Sup(X) = Freq(X) / |D|$$

où Freq (X): nombre des transactions dans D qui contiennent l'itemset X. $Sup(X) \in [0,1]$

- **Itemset fréquent :** On appelle X un itemset fréquent si:

$$Sup(X) >= MinSup$$

où MinSup est le seuil minimal transmit par l'utilisateur. $MinSup \in [0,1]$

Propriétés de l'algorithme Apriori

1) **Propriété de monotonie:** ***tous les sous ensembles*** **d'un itemset fréquent sont aussi** ***fréquents.***

Exemple:

si ABCD est un d'itemset fréquents, alors, les sous ensembles :ABC,ABD, BCD, AB,AC,BC,BD,CD,A,B,C,D les sont aussi.

2) **Propriété d'anti monotonie:**

tous les sur ensembles **d'un itemset infréquent sont aussi** ***infréquents.***

Exemple:

si AB est un itemset infréquents, alors, les sur ensembles :ABC,ABD les sont aussi.

Algorithme Apriori

Entrée: *K* : Contexte d'Extraction, Minsup

Sortie: Ensemble des itemsets fréquents

1: Initialiser l'ensemble de candidats de taille 1

2: **tant que** ensemble de candidats est non vide **faire**

1) Calculer le support des candidats
2) Élaguer l'ensemble de candidats par rapport à minsup

Étape d'élagage (ou de test)

1) Construire l'ensemble des candidats pour l'itération suivante

Étape de construction

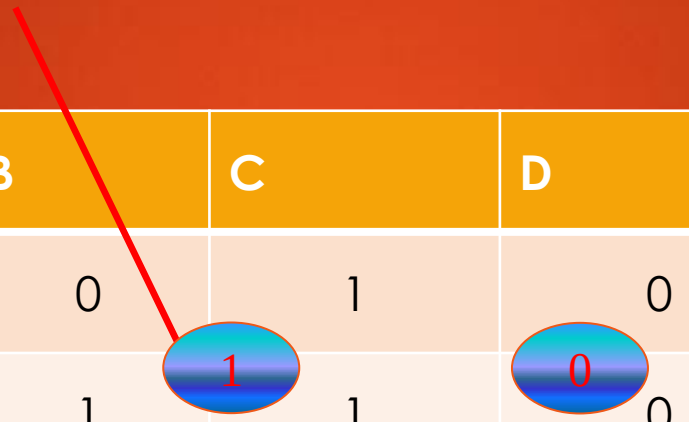
5: fin tant que

6: retourner Ensemble des itemsets fermés fréquents

Exemple

Soit D la base de transactions contenant un ensemble de transaction décrivant des achats de produits dans l'ensemble {A, B, C, D, E}

	A	B	C	D	E
T1	1	0	1	0	1
T2	1	1	1	0	0
T3	1	0	0	1	1
T4	0	1	1	0	1
T5	0	1	1	0	0



Exemple

NECIR HAMID

Les étapes d'extraction des itemsets fréquents par l'algorithme Apriori sont:
MinSup=2/5

1-itemset	Sup
A	3/5
B	3/5
C	4/5
D	1/5
E	3/5

infréquent

2-itemset	Sup
AB	1/5
AC	2/5
AE	2/5
BC	3/5
BE	1/5
CE	2/5

infréquent

infréquent

3-itemset	Sup
ACE	1/5

infréquent

Etape 2: génération des règles d'association

Règle d'association

NECIR HAMID

Règle d'association : Indication de précision de la règle. Implication de la forme:

$$A \rightarrow B$$

où A B ⊆ I, B ⊆ I et A ∩ B = Ø

A est dit Antécédant et B est dit Conséquent.

Une règle d'association exprime le fait que les items de A tendent à apparaître avec ceux de B.

Mesure de qualité d'une règle d'association

NECIR HAMID

Confiance: La confiance d'une règle d'association A → B, représente la proportion (Pouvant être exprimé en pourcentage) de transactions couvrant A qui couvrent aussi B.

$$Confiance(A \rightarrow B) = Sup(AB)/Sup(A)$$

Règle solide : Une règle d'association A → B est solide si sa confiance dépasse un seuil donné, fixé a priori, appelé **MinConf**.

Règle forte : Une règle d'association A → B est forte si sa confiance est égale a 1 (100%).

Mesure de qualité d'une règle d'association

Lift: Mesure le caractère significatif de l'association.

$$Lift(A \rightarrow B) = Sup(AB) / (Sup(A).Sup(B))$$

- Un lift supérieur à 1 : Indique une corrélation positive
- Un lift de 1 indique une corrélation nulle
- Un lift inférieur à 1 : Indique une corrélation négative

Exemple

NECIR HAMID

Les étapes d'extraction des règles d'association. On considère MinConf=3/5=0,67

itemset	règles d'association	Lift	Confiance
AC	A→C	1,40	Conf(A→C)=0.67
	C→A	1.40	Conf(C→A)=0,5
AE	A→E	1,11	Conf(A→E)=0,67
	E→A	1,11	Conf(E→A)=0,67
BC	B→C	1,25	Conf(B→C)=1
	C→B	1,25	Conf(C→B)=0,75
CE	C→E	0,83	Conf(C→E)=0,5
	E→C	0,83	Conf(E→C)=0,67

Règle forte

L'algorithme CLOSE

- Proposé en 1998 par Pasquier et al.
- Algorithme itératif par niveau pour l'extraction des itemsets fermés fréquents.
- Durant chaque itération k de l'algorithme, un ensemble de k-générateurs candidats est considéré. Chaque élément de cet ensemble est constitué de trois éléments : le k-générateur candidat, sa fermeture, et leur support.
- À la fin de l'itération k, l'algorithme stocke un ensemble contenant les k-générateurs fréquents, leurs fermetures, qui sont des itemsets fermés fréquents, et leurs supports.

Algorithme CLOSE

Entrée: *K* : Contexte d'Extraction, minsup

Sortie: Ensemble des itemsets fermés fréquents

1: Initialiser l'ensemble de candidats de taille 1

2: **tant que** ensemble de candidats est non vide **faire**

1) Calculer le support des candidats
2) Élaguer l'ensemble de candidats par rapport à minsup
3) calculer les fermetures des candidats retenus

Étape d'élagage (ou de test)

1) Construire l'ensemble des candidats pour l'itération suivante
2) Élaguer cet ensemble en utilisant les propriétés des itemsets fermés.

Étape de construction

5: fin tant que

6: retourner Ensemble des itemsets fermés fréquents

Exemple

Soit D la base de transactions contenant un ensemble de transaction décrivant des achats de produits dans l'ensemble {A, B, C, D, E} . MinSup=2/5

A	B	C	D	E
1	1	1	1	1
1	1	0	0	0
0	0	1	0	1
1	1	0	1	1
1	0	1	1	0

Exemple

Les étapes d'extraction des itemsets fréquents par l'algorithme CLOSE sont:

1-Gen	Sup	Ferm
A	4/5	A
B	3/5	AB
C	3/5	C
D	3/5	AD
E	3/5	E

2-itemset	Sup	Ferm
AC	2/5	ACD
AE	2/5	ABDE
BC	1/5	ABCDE
BD	2/5	ABDE
BE	2/5	ABDE
CD	2/5	ACD
CE	2/5	CE
DE	2/5	ABDE

3-itemset	Sup	Ferm
ACE	1/5	ABCDE
CDE	1/5	ABCDE

Exemple

Les étapes d'extraction des règles d'association. On considère MinConf=3/5=0,67

règles d'association	Confiance	Lift
B→A	Confiance=1	1,25
D→A	Confiance=1	1,67
AC→D	Confiance=1	4,17
AE→BD	Confiance=1	6,25
BD→AE	Confiance=1	6,25
BE→AD	Confiance=1	4,17
CD→A	Confiance=1	3,13
DE→AB	Confiance=1	4,17