Association Rule Learning

MARKET BASKET ANALYSIS

Ricco RAKOTOMALALA

Market basket transactions – Transactional format (I)

N° transaction (Caddie)	Contenu du caddie			
1	pastis	martini	chips	saucisson
2	martini	chips		
3	pain	beurre	pastis	
4	saucisson			
5	pain	lait	beurre	
6	chips	pain		
7	confiture			

>> one row = one record = one **transaction**
>> only the presence of the products [**items**] is relevant (not their quantity)
>> Variable number of items in a transaction
>> Very high number of possible items

Goals:
(1) Highlight the relationship between the items (the products that are bought together)
(2) Represent the knowledge in the form of **association rules**

Set of items (**itemset**)

IF antecedent THEN consequent

Ex. IF (a customer purchases) pastis and martini THEN (he purchases also) saucisson and chips

Market basket transactions → Tabular data (II)

Another representation of the transaction data

N° transaction (Caddie)	Contenu du caddie		
1	p1	p2	p3
2	p1	p3	
3	p1	p2	p3
4	p1	p3	
5	p2	p3	
6	p4		

Caddie	p1	p2	p3	p4
1	1	1	1	0
2	1	0	1	0
3	1	1	1	0
4	1	0	1	0
5	0	1	1	0
6	0	0	0	1

The number of columns can be very high. We have a very sparse data. Some columns can be merged if we want to handle families of products.

Standard tabular data (III)

From attribute-value dataset to binary dataset.

Dummy coding

Observation	Taille	Corpulence
1	petit	mince
2	grand	enveloppé
3	grand	mince

Observation	Taille = petit	Taille = grand	Corpulence = mince	Corpulence = enveloppé
1	1	0	1	0
2	0	1	0	1
3	0	1	1	0

Once the data can be transformed to binary data, we can learn association rules.

We want to detect the co-occurrence of modalities (values of the variables). Some associations are not possible by nature e.g. an individual cannot be tall (grand) and short (petit) at the same time.

Dataset

Caddie	p1	p2	p3	p4
1	1	1	1	0
2	1	0	1	0
3	1	1	1	0
4	1	0	1	0
5	0	1	1	0
6	0	0	0	1

R1: IF p1 THEN p2

SUPPORT: Proportion of transactions which contains the itemset

sup(R1) = 2 or sup(R1) = 2/6 = 33%

in absolute terms in relative terms

CONFIDENCE: Estimate of the probability that the consequent is true if the antecedent is true

$$conf(R1) = \frac{\sup(R1)}{\sup(antecedent\ R1)}$$

$$= \frac{\sup(p1 \rightarrow p2)}{\sup(p1)} = \frac{2}{4} = 50\%$$

"Interesting" rule = rule with both high support and high confidence

Extraction of association rule (I)

Basic algorithm (based on the Zaki's ECLAT approach)

Settings: set constraints on support and confidence
» MIN Support (ex. 2 transactions)
» MIN Confidence (ex. 75%)

→ The aim is to generate only interesting rules
→ The aim is also to control the number of rules extracted

Process: Extraction in two major steps
» Frequent Itemset Generation (itemset for which support ≥ support min.)
» From frequent itemset, rule generation (confidence ≥ conf. min.)

Some definitions:
» item = product
» itemset = set of products (ex. {p1,p3})
» sup(itemset) = Number of transactions where the products are simultaneously present (ex. sup{p1,p3} = 4)
» card(itemset) = Number of products into the itemset. (ex. card{p1,p3} = 2)

Extraction of association rules (II)

Discovering the frequent itemsets

Potentially: (2^J-1) candidate itemsets

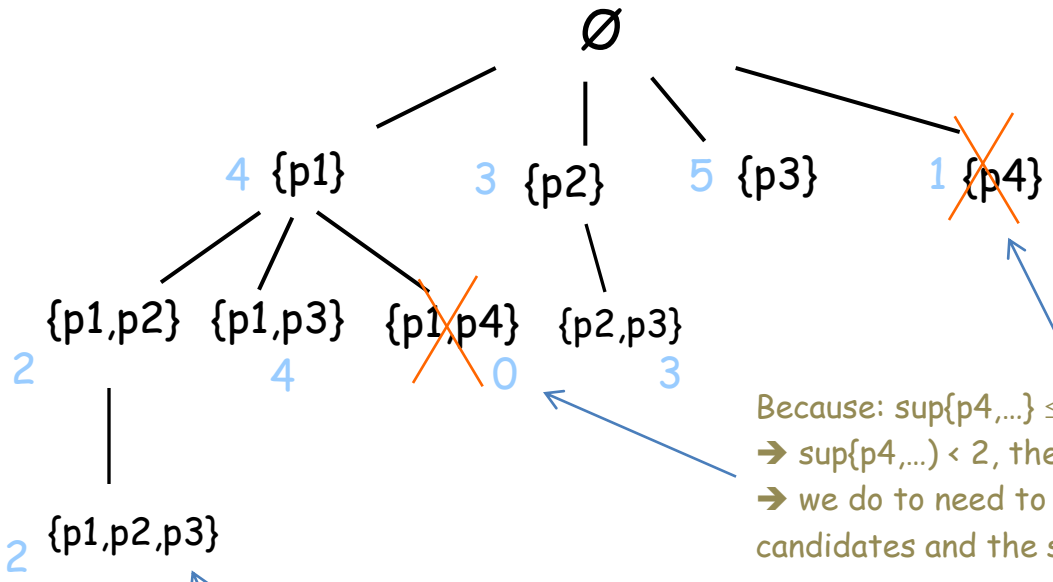
>> Amount of calculations not tractable
>> Each calculation requires the accessing of the database

$C_4^1 = 4$ ← Itemsets with cardinal = 1
$C_4^2 = 6$ ← Itemsets with card = 2
$C_4^3 = 4$ ← Itemsets with card = 3
$C_4^4 = 1$
$\Sigma = 15 = 2^4 - 1$ ...

Reduce the search space by eliminating straightaway some combinations

Dataset

Caddie	p1	p2	p3	p4
1	1	1	1	0
2	1	0	1	0
3	1	1	1	0
4	1	0	1	0
5	0	1	1	0
6	0	0	0	1



What happens if we set sup.min = 3?

Extraction of association rules (III)

Extracting the rules from itemset with card = 2

We need to check all the combinations. 2 calculations for each itemset.

Dataset

Caddie	p1	p2	p3	p4
1	1	1	1	0
2	1	0	1	0
3	1	1	1	0
4	1	0	1	0
5	0	1	1	0
6	0	0	0	1

{p1,p2}
- p1→p2 : conf. = 2/4 = 50% (refused)
- p2→p1 : conf. = 2/3 = 67% (refused)

{p1,p3}
- p1→p3 : conf. = 4/4 = 100% (accepted)
- p3→p1 : conf. = 4/5 = 80% (accepted)

{p2,p3}
- p2→p3 : conf. = 3/3 = 100% (accepted)
- p3→p2 : conf. = 3/5 = 60% (refused)

What happens if we set conf. min. = 55 %?

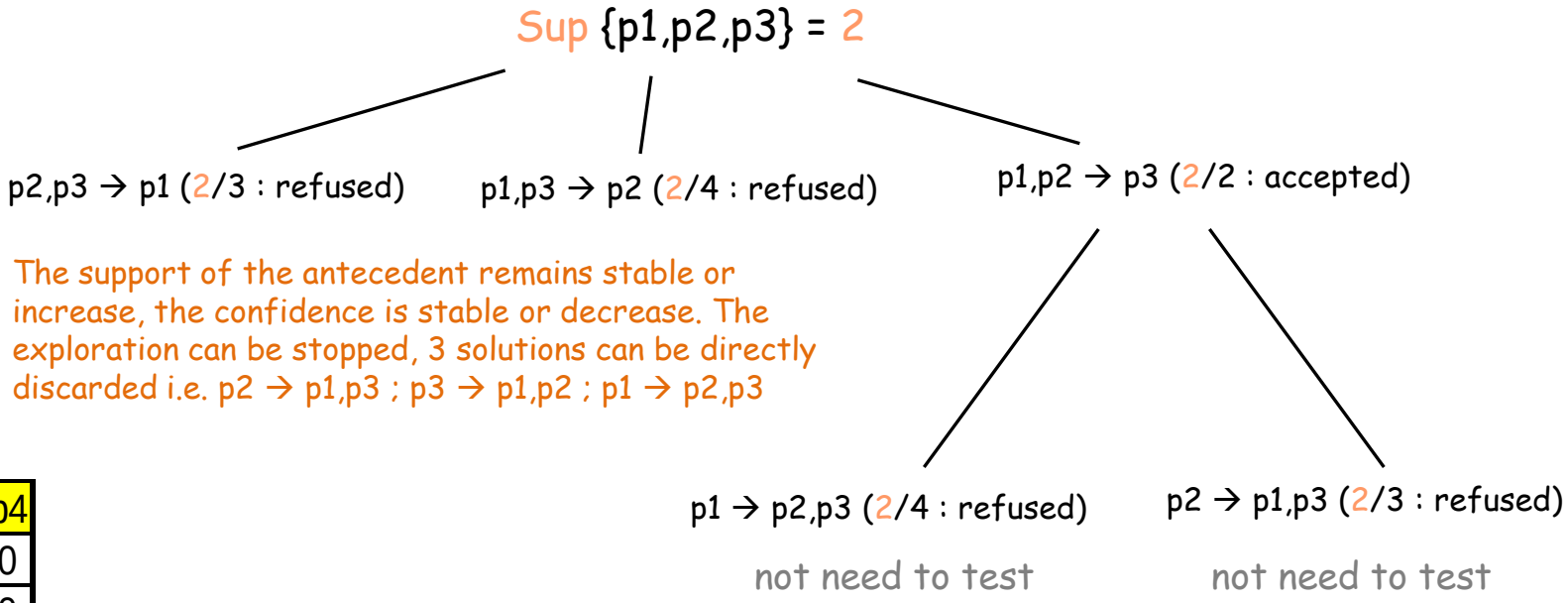
Extraction of association rules (IV)

Extracting the rules for itemset with card ≥ 3

$C_3^1 = 3$ ← Rules with card consequent = 1

$C_3^2 = 3$ ← Rules with card consequent = 2

Reduce the search space by eliminating some solutions



Dataset

Caddie	p1	p2	p3	p4
1	1	1	1	0
2	1	0	1	0
3	1	1	1	0
4	1	0	1	0
5	0	1	1	0
6	0	0	0	1

What happens if we set conf. min. = 55 %?

Alternative measures of interestingness

e.g. LIFT

Support is high
Confidence = 100%

What about the following rule?

IF hair = brown THEN brain = present

The confidence in probabilistic terms

$$conf(A \rightarrow C) = \frac{\sup(AC)}{\sup(A)} = \frac{P(AC)}{P(A)} = P(C/A)$$

LIFT

$$lift(A \rightarrow C) = \frac{P(C/A)}{P(C)} = \frac{P(AC)}{P(A) \times P(C)}$$

P(.) = Support in relative terms

Ratio between of the observed support to that expected if A and C were independent
Lift ≤ 1 → Negative correlation between the antecedent and the consequent

Interpretation : LIFT(smoke → cancer) = 3% / 1% = 3

When we smoke, the risks for cancers occurring is multiplied by 3.

The LIFT measure can be computed afterwards for filtering or sorting of rules.
It cannot be used during the search of solutions.

Many other measures are proposed in literature, none really emerged.

From association rules to sequential pattern mining

The values are delivered in sequence (time series analysis is a special case)

Can we extract this kind of rule?

IF «wrecking of vehicle » and « full reimbursement » Then « purchase of new car»

Step 1 Step 2 Step 3

Transactional data

Timed data (at least sequence of values)

Clients	Achat 1	Achat 2	Achat 3	Achat 4
C1	(1, 2, 3)	(4, 2, 5)	(1, 6, 2)	(4, 1)
C2	(1, 3, 2)	(1, 2, 3)	(6, 3, 2)	
C3	(4, 8)	(1, 3, 7)	(5, 8)	(1, 4)
C4	(5, 2, 3)	(1, 2, 3)	(1, 2, 8)	(1, 6, 2)

Itemsets and rules

Support < (1, 3) (2) (6, 2) > = 3 (or $\frac{3}{4}$ = 75%)

If (1, 3) Then (2) (6, 2) → confidence = $\frac{3}{4}$ = 75%

If (1, 3) (2) Then (6, 2) → confidence = 3/3 = 100%

! The calculations are not easy. Few tools incorporates this approach.

References

Wikipedia, "Association rule learning".

M. Zaki, S. Parthasaraty, M. Ogihara, W. Li, "New Algorithms for Fast Discovery of Association Rules", in Proc. of KDD'97, p. 283-296, 1997.

P.N. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining", Addison-Wesley, 2006 ; Chap.6 "Association Analysis: Basic concepts and Algorithms".

TANAGRA Tutorials about "Association Rules".

Wikipedia, "Sequential pattern mining".