Les Règles d'Association

MARKET DATA ANALYSIS
ou
L'analyse du panier de la ménagère

Ricco RAKOTOMALALA

Plan

1. Type de données traitées – Finalité de l'analyse
2. Recherche des « itemsets » fréquents
3. Construction des règles
4. Mesure d'évaluation des règles
5. Règles d'association et logiciels
6. Plus loin : recherche des motifs séquentiels

Type de données traitées

Finalité de l'extraction des règles d'association

Données de transaction (I)

Analyse des tickets de caisse

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			

Commentaires: ou un ticket de caisse...

>> Une observation = Un caddie

>> Ne tenir compte que de la présence des produits (pas de leur quantité)

>> Nombre variable de produits dans un caddie

>> La liste des produits est immense !

! Objectifs :

(1) Mettre en évidence les produits achetés ensemble

(2) Transcrire la connaissance sous forme de règle d'association

Si antécédent Alors conséquent

listes de produits

Ex. Si pastis et martini Alors saucisson et chips

Rq : on ne traite (capte) pas les règles négatives. ex : s'il n'a pas acheté de vin alors il va acheter du soda...

Données de transaction (II)
Tableau de transactions → Tableau binaire 0/1

Autre représentation des données de transactions

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

ex. de regroupement en familles de produits : s'il achète un ouvre-boîte, alors il va acheter une boîte de conserve

Selon la granularité choisie, le nombre de colonnes peut être immense.
(ex. détail par marques ou regroupement en familles → boîtes de cassoulet)

Données de transaction (III)

Tableau individus x variables → Tableau binaire 0/1

Codage disjonctif complet

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

Dès que l'on peut se ramener à des données 0/1
Il est possible de construire des règles d'association

Il s'agit de détecter les cooccurrences des modalités (attribut = valeur)
Certaines associations sont impossibles par construction (ex. on ne peut pas être « petit » et « grand » en même temps)

cependant

Critères d'évaluation des règles d'association

Support et confiance

Soit la règle d'association

R1 : Si p1 alors p2

Données

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

SUPPORT : Un indicateur de « fiabilité » de la règle

en termes absolus

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

en termes relatifs

CONFIANCE : Un indicateur de « précision » de la règle

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

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

« Bonne » règle = règle avec un support et une confiance élevée

Extraction des règles d'association
Démarche globale

Paramètres : Fixer un degré d'exigence sur les règles à extraire
>> Support min. (ex. 2 transactions)
>> Confiance min. (ex. 75%)

→ L'idée est surtout de contrôler (limiter) le nombre de règles produites

Démarche : Construction en deux temps
>> recherche des itemsets fréquents (support >= support min.)
>> à partir des itemsets fréquents, produire les règles (conf. >= conf. min.)

Quelques définitions
>> item = produit
>> itemset = ensemble de produits (ex. {p1,p3})
>> sup(itemset) = nombre de transactions d'apparition simultanée des produits (ex. sup{p1,p3} = 4)
>> card(itemset) = nombre de produits dans l'ensemble (ex. card{p1,p3} = 2)
>> itemset fréquent = itemset dont le support est ≥ à support min

Recherche des itemsets fréquents

Idée : extraire tous les « itemsets fréquents » en minimisant les calculs, notamment le nombre d'accès à la base de données

pour que le calcul soit réalisable sur de très grandes bases de données....

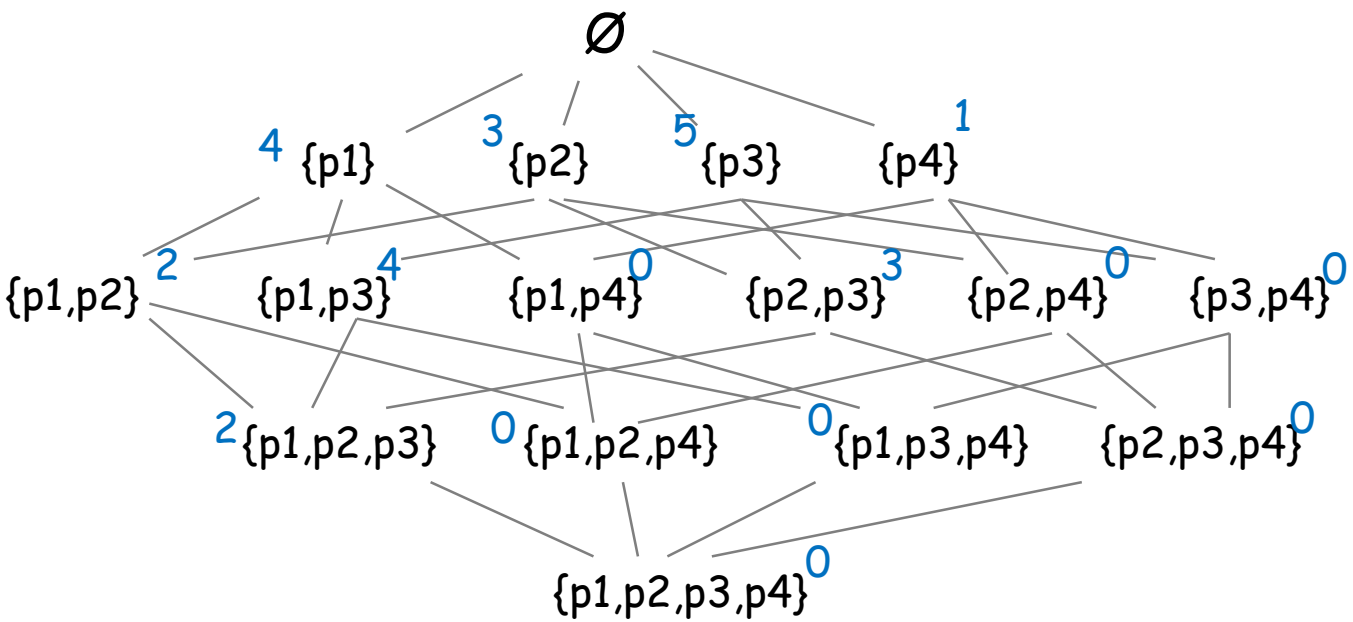
La recherche des itemsets fréquents peut être une finalité en elle-même c.-à-d. détecter les produits qui sont achetés simultanément

recherche de co-occurence simple, sans s'intéresser à la causalité

Extraction des itemsets fréquents

Il s'agit de parcourir un treillis et de calculer les supports associés à chaque combinaison

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



$C_4^1 = 4$ # itemsets de card = 1

$C_4^2 = 6$ # itemsets de card = 2

$C_4^3 = 4$

$C_4^1 = 1$

$\Sigma = 15 = 2^4 - 1$

Le nombre de configuration est très vite très élevé

Chaque configuration nécessiterait un scan de la base de données

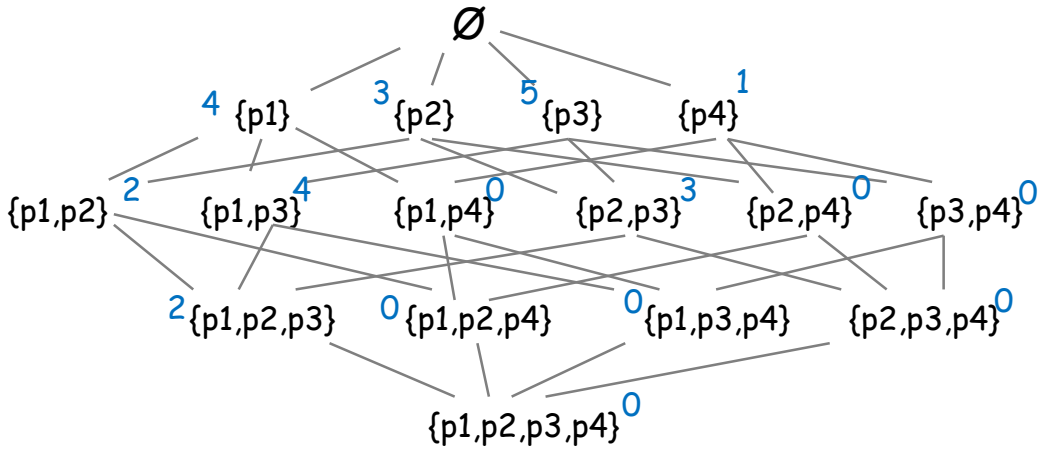
➔ Il faut s'appuyer sur le paramètre « support.min »

➔ Et les propriétés des itemsets

Pour réduire le nombre de configuration à évaluer réellement

Extraction des itemsets fréquents

Quelques définitions



Sup.min = 2

Itemset : un ensemble d'items

Superset : B est un superset de A si card(A) < card(B) et $A \subset B$
➔ sup(B) ≤ sup(A)

Itemset fréquent : itemset dont le support est ≥ à sup.min

Si un itemset n'est pas fréquent, tous ses supersets ne le seront pas non plus.

⊃

Itemset fréquent fermé : itemset fréquent dont aucun de ses supersets n'a un support identique (ex. {p1,p3} est fermé, {p1,p2} ne l'est pas)

⊃

Itemset fréquent maximal : itemset fréquent dont aucun de ses supersets n'est fréquent (ex. {p1,p2,p3} est maximal)

On prend usuellement comme point de départ les itemsets fréquents pour générer **toutes** les règles d'association.

Prendre comme point de départ les itemsets fréquents fermés pour générer les règles permet de réduire le nombre de règles redondantes
Ex. (p1 → p2) et (p1→p2,p3) auront la même confiance, la 1ère est redondante par rapport à la 2nde.

Donne la représentation la plus compacte possible de la liste des itemsets. Ex. si on sait que {p1,p2,p3} est fréquent, on sait que {p1,p2}, {p1,p3} et {p2,p3} le sont également (mais on ne connaît pas leur support)

Extraction des itemsets fréquents

Une approche très simple

$C_4^1 = 4$ ← Itemsets de card = 1

$C_4^2 = 6$ ← Itemsets de card = 2

$C_4^3 = 4$ ← Itemsets de card = 3

$C_4^4 = 1$

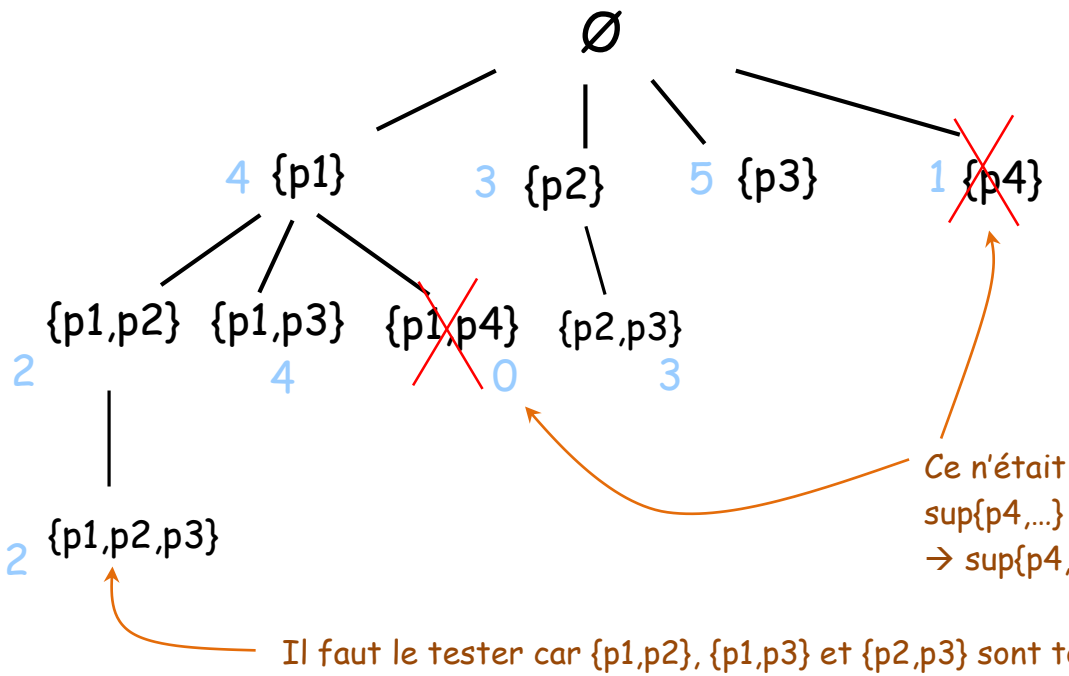
$\Sigma = 15 = 2^4 - 1$

...

Réduire l'exploration en éliminant d'emblée certaines pistes via le support min (sup.min = 2 ici) et les propriétés des itemsets

Données

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



Que se passerait-il si nous avions sup. min. = 3 ?

Extraction des règles

Idée : déduire les règles à partir des itemsets fréquents

On limite la prolifération des règles en utilisant le critère confiance min.

éviter les règles redondantes....

Extraction des Règles d'Association

Recherche des règles pour les itemsets de card = 2

Il faut tester toutes les combinaisons : 2 tests par itemset

Tous les supports sont dispos dans le treillis, pas besoin de scanner la base

Données

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% (refusé)
- p2→p1 : conf. = 2/3 = 67% (refusé)

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

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

Que se passerait-il si nous avions conf. min. = 55 %

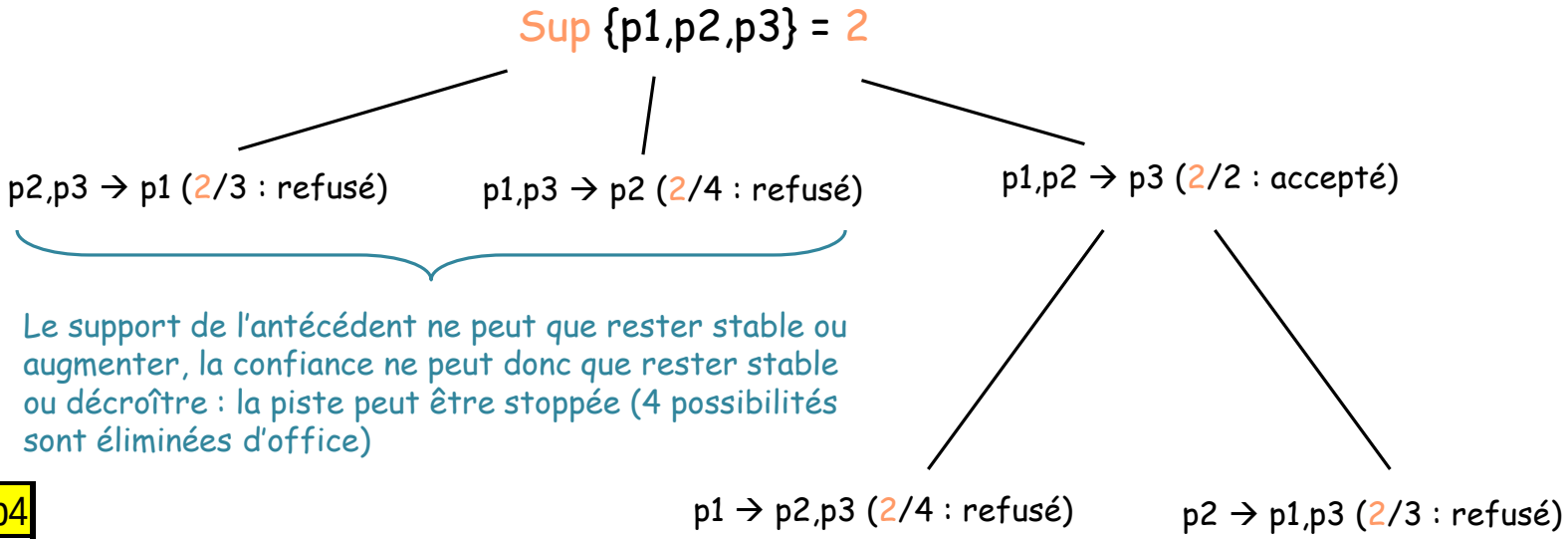
Extraction des Règles d'Association

Recherche des règles pour les itemsets de card = 3 et plus…

$C_3^1 = 3$ ← Règles avec conséquent de card = 1

$C_3^2 = 3$ ← Règles avec conséquent de card = 2

Réduire l'exploration en éliminant d'emblée certaines pistes
Le support de la règle ne change jamais = support de l'itemset
On peut jouer sur le support de l'antécédent



Données

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

Que se passerait-il si nous avions conf. min. = 55 %

Mesures d'évaluation des règles

Aller au-delà du support et de la confiance

Support et confiance

En termes probabilistes

R : p3→p1

Données

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

Support (en termes relatifs)

Proba. conjointe

$sup(R) = 4/6 = P(p_3 p_1)$

Confiance

Proba. conditionnelle

$conf(R) = sup(\{p1,p3\})/sup(\{p3\}) = 4 / 5 = P(p_3/p_1)$

P(p1/p3)

Mais

Une règle peut avoir d'excellents supports et confiance sans être pour autant « intéressante »

Si Sexe = Masculin Alors Cerveau = présent

Support = 50%
Confiance = 100%

Il faut un critère – une mesure d'intérêt – qui caractérise une forme de causalité c.-à-d. l'idée « la connaissance de l'antécédent amène de l'information (supplémentaire) sur la connaissance du conséquent »

Un indicateur de pertinence des règles

Dépasser le support et la confiance avec le LIFT

R : Antécédent → Conséquent

On sait que

$$\sup(A) = P(A)$$ ← Support de l'antécédent

$$\sup(C) = P(C)$$ ← Support du conséquent

$$\sup(A \rightarrow C) = \sup(AC) = P(AC)$$

$$conf(A \rightarrow C) = P(C / A)$$

Le LIFT → $$lift(A \rightarrow C) = \frac{P(C / A)}{P(C)}$$

Ex. LIFT(fumer → cancer) = 3% / 1% = 3

Rapport de probabilité - Surcroît d'occurrence du conséquent quand l'antécédent est présent

Lift = 1 → La règle ne sert absolument à rien…

~ fumer multiplie par 3 la survenue du cancer (~ proche de la notion de risque relatif)

Le LIFT ne peut être calculé qu'après coup pour filtrer les règles. Nous ne pouvons pas l'utiliser pour guider l'apprentissage

Remarque :

Le LIFT peut se lire également comme un rapport de vraisemblance ➔ sous H0 : A et B sont indépendants

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

Autres indicateurs de pertinence des règles

Données

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

Le point de départ est un tableau croisé

R : p3→p1

	Antécédent	Non(Antécédent)	Total
Conséquent	$n_{ac} = 4$ exemples		$n_c = 4$
Non(conséquent)	contre-exemples		
Total	$n_a = 5$		$n = 6$

Quelques mesures

Mesure	Formule
Support	$\frac{n_{ac}}{n}$
Confiance	$\frac{n_{ac}}{n_a}$
Lift	$\left(\frac{n_{ac}}{n_a}\right) / \left(\frac{n_c}{n}\right)$
Leverage	$\frac{n_{ac}}{n} - \frac{n_a}{n} \times \frac{n_c}{n}$
Importance	$\ln\left[\left(\frac{n_{ac}}{n_a}\right) / \left(\frac{n_c - n_{ac}}{n - n_a}\right)\right]$
Conviction	$\frac{n_a \times (n - n_c)}{n \times (n_a - n_{ac})}$
Surprise	$\left(\frac{n_{ac}}{n} - \frac{n_a - n_{ac}}{n}\right) / \left(\frac{n_c}{n}\right)$

développé par Sergey Brin →

- Il faut pouvoir les interpréter.
- Mis à part la confiance et le support, elles interviennent uniquement après coup pour trier les règles.
- Dans certaines configurations (ex. contre-exemples = 0), les mesures ne sont pas calculables. On peut s'en sortir avec l'estimation laplacienne des probabilités.
- Aucune n'est vraiment décisive parce que la notion de « règle intéressante » est difficile à situer.

Les règles d'association dans les logiciels

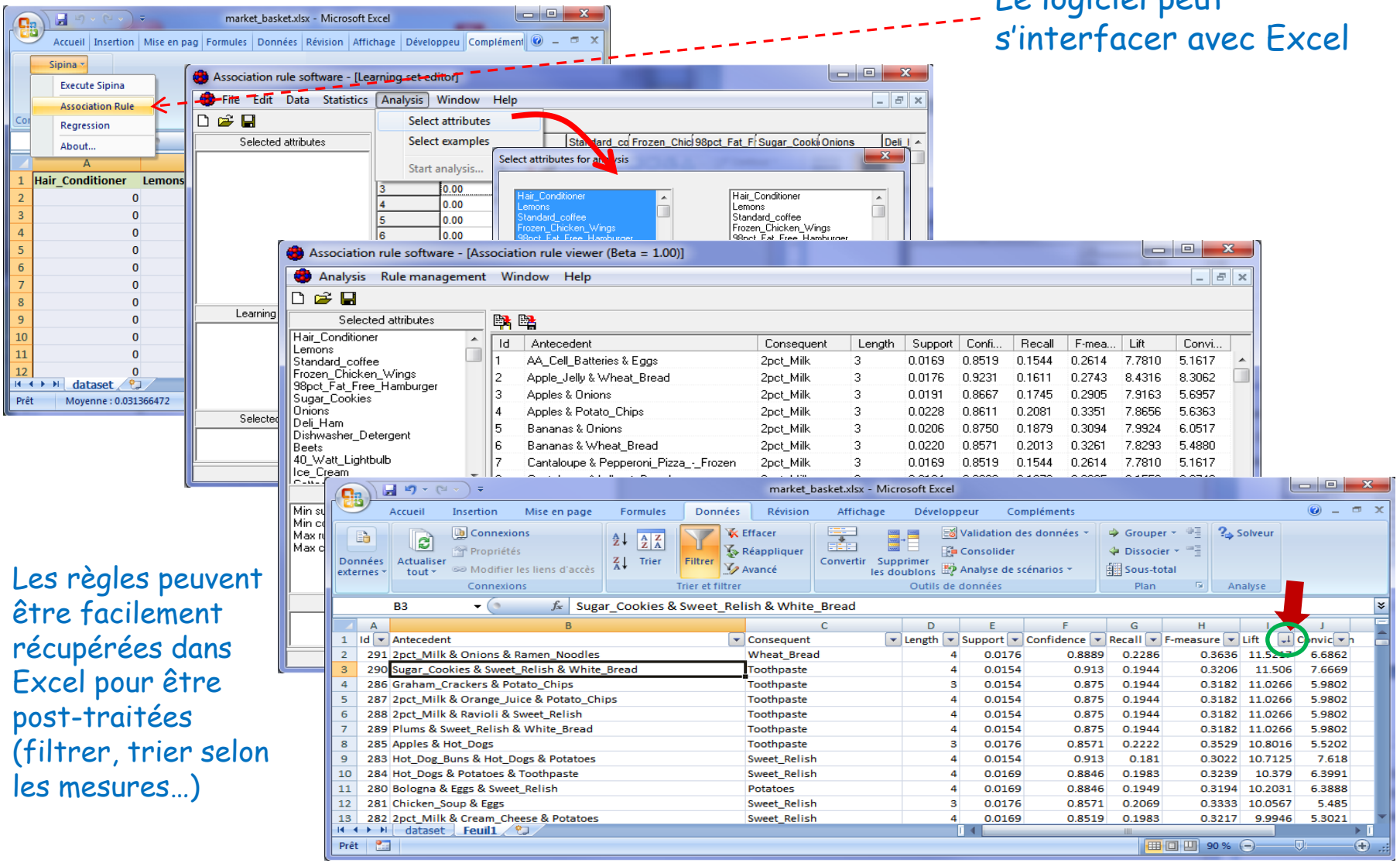
Les logiciels s'appuient sur différents algorithmes
A PRIORI, ECLAT, FP-GROWTH, etc.

Méthode présente dans tous les outils estampillés
« Data Mining »

Règles d'association dans la distribution SIPINA

L'outil d'extraction de règle – Interaction avec Excel

Le logiciel peut s'interfacer avec Excel



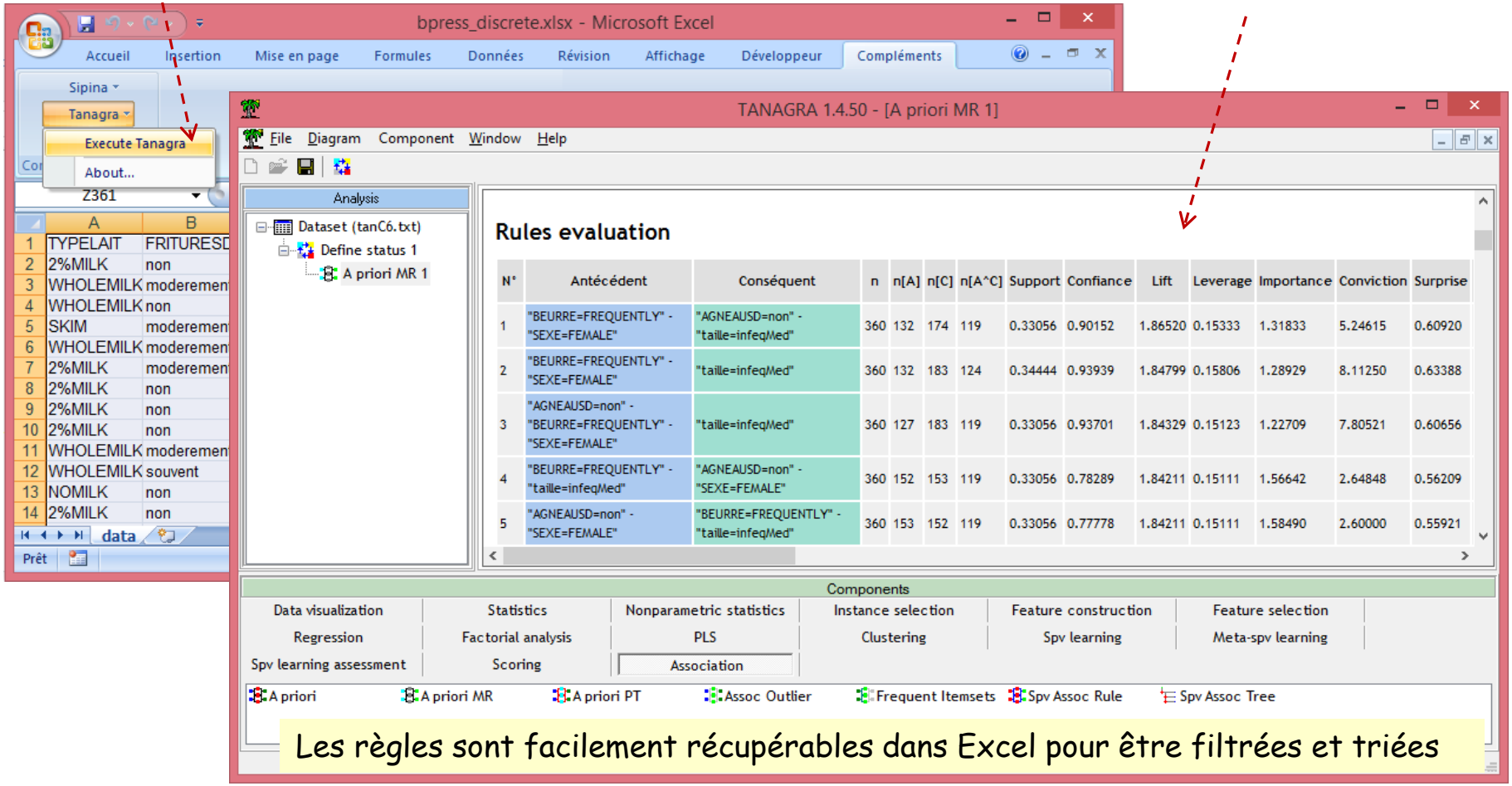
Les règles peuvent être facilement récupérées dans Excel pour être post-traitées (filtrer, trier selon les mesures…)

Tutoriel Tanagra, « Associations dans la distribution SIPINA », avril 2013.

Règles d'association avec TANAGRA (I)

Interaction avec Excel

Plusieurs mesures sont disponibles, un tutoriel spécifique y est consacré



N°	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importance	Conviction	Surprise
1	"BEURRE=FREQUENTLY" - "SEXE=FEMALE"	"AGNEAUSD=non" - "taille=infeqMed"	360	132	174	119	0.33056	0.90152	1.86520	0.15333	1.31833	5.24615	0.60920
2	"BEURRE=FREQUENTLY" - "SEXE=FEMALE"	"taille=infeqMed"	360	132	183	124	0.34444	0.93939	1.84799	0.15806	1.28929	8.11250	0.63388
3	"AGNEAUSD=non" - "BEURRE=FREQUENTLY" - "SEXE=FEMALE"	"taille=infeqMed"	360	127	183	119	0.33056	0.93701	1.84329	0.15123	1.22709	7.80521	0.60656
4	"BEURRE=FREQUENTLY" - "taille=infeqMed"	"AGNEAUSD=non" - "SEXE=FEMALE"	360	152	153	119	0.33056	0.78289	1.84211	0.15111	1.56642	2.64848	0.56209
5	"AGNEAUSD=non" - "SEXE=FEMALE"	"BEURRE=FREQUENTLY" - "taille=infeqMed"	360	153	152	119	0.33056	0.77778	1.84211	0.15111	1.58490	2.60000	0.55921

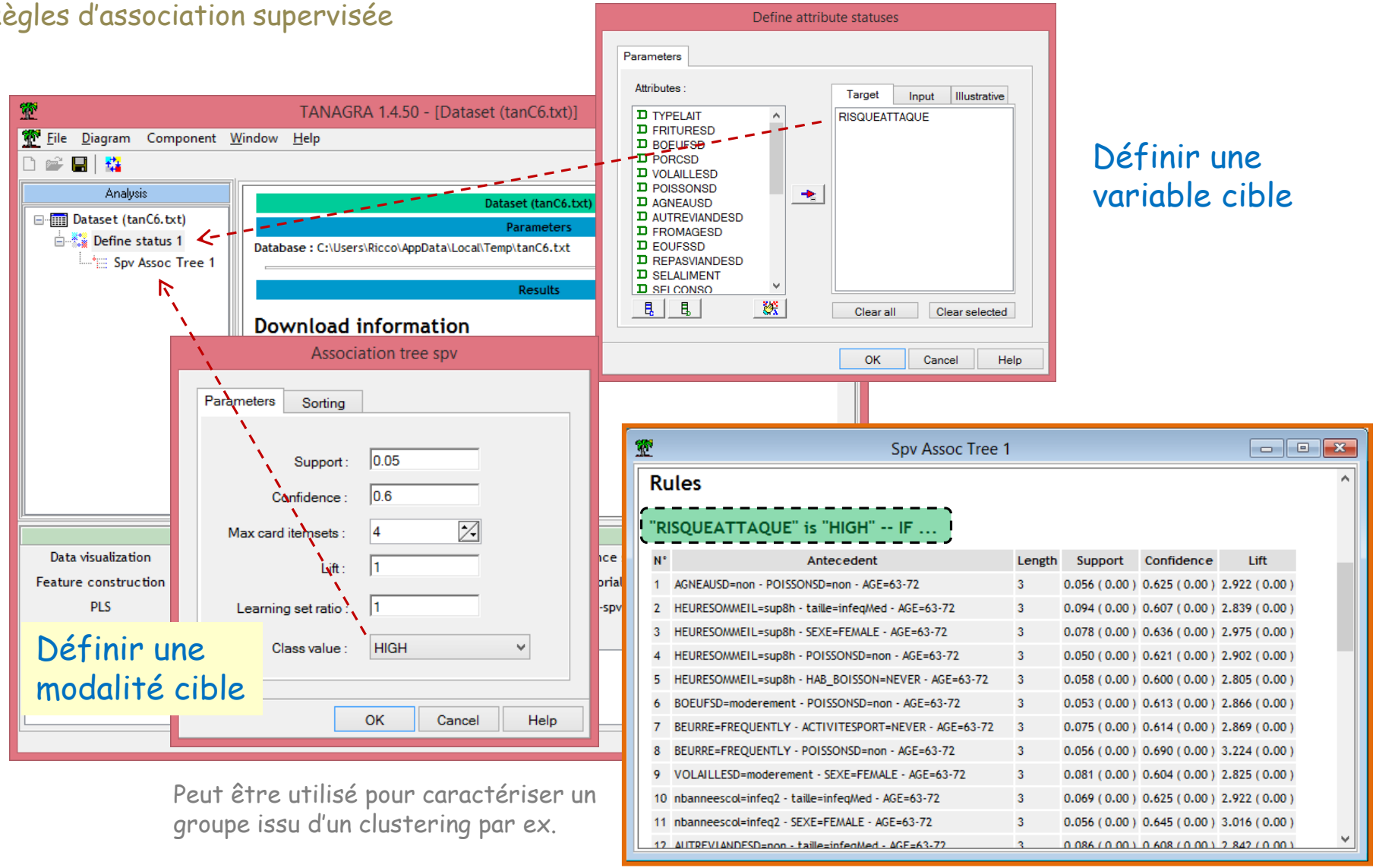
Les règles sont facilement récupérables dans Excel pour être filtrées et triées

Tutoriels

« Règles d'association – Orange, Tanagra et Weka », avril 2008.
« Règles d'association – Comparaison de logiciels », novembre 2008.
« Règles d'association avec APRIORI PT », avril 2008.

Règles d'association avec TANAGRA (II)

Règles d'association supervisée



Définir une variable cible

Peut être utilisé pour caractériser un groupe issu d'un clustering par ex.

Tutoriels

« Règles d'association prédictives », février 2009.
« Règles d'association "supervisées" », avril 2008.

Règles d'association avec R

Le package « arules »

(l'équivalent existe sous Python - http://tutoriels-data-mining.blogspot.com/2019/02/regles-dassociation-sous-python.html)

Extraction de différents types d'itemsets fréquents

itemset_mining.r

```
#clear the memory
rm(list=ls())
#importing the dataset
library(xlsReadWrite)
dataset <- read.xls(file="itemset_mining.xls",colNames=T,sheet=1)
print(dataset)
#loading arule library
library(arules)
#extracting the frequent itemsets
params <- list(supp = 0.2, minlen = 2, maxlen = 4, target="frequent itemsets")
result <- apriori(as.matrix(dataset), parameter = params)
inspect(result)
#extracting the closed itemsets
params <- list(supp = 0.2, minlen = 2, maxlen = 4, target="closed frequent itemsets")
result <- apriori(as.matrix(dataset), parameter = params)
inspect(result)
#extracting the maximally itemsets
params <- list(supp = 0.2, minlen = 2, maxlen = 4, target="maximally frequent itemsets")
result <- apriori(as.matrix(dataset), parameter = params)
inspect(result)
```

Extraction et visualisation des règles

assoc rule on german.r

```
#charger le package
library(arules)
#charger le fichier de données
german <- read.table(file="credit-german.txt",header=T,dec=".",sep="\t")
summary(german)
#transformer les données attributs-variables
#en données transactionnelles
german.trans <- as(german,"transactions")
summary(german.trans)
#extraction des règles
german.regles <- apriori(german.trans,parameter=
                  list(supp=0.25,conf=0.75,minlen=2,maxlen=10,target="rules"))
summary(german.regles)
#afficher les 10 premières règles trouvées
inspect(german.regles[1:10])
#afficher les 5 règles avec le lift le + élevé
regles.triees <- sort(german.regles,by="lift")
inspect(regles.triees[1:5])
```

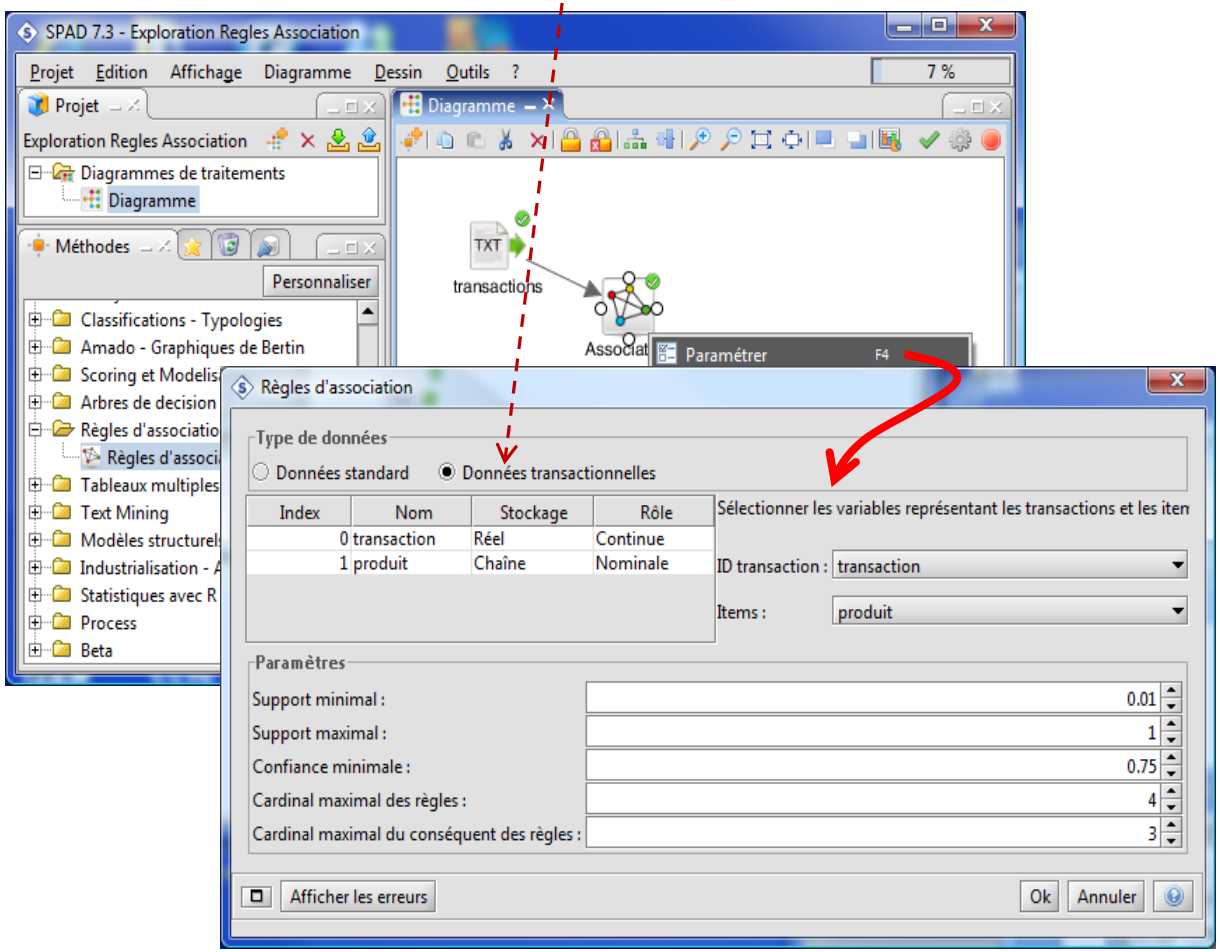
Règles d'association avec SPAD

Peut traiter indifféremment les bases « individus x variables » et les bases transactionnelles

Spécification du format de données pour le traitement

Format « transactions »

transaction	produit
1	B
1	E
1	H
2	A
2	B
2	E
2	F
3	B
3	C
3	F
3	H



Un outil interactif permet de filtrer et trier les règles

Extraction des motifs séquentiels

Intégrer des contraintes temporelles (succession)
dans la recherche des règles

Des règles d'association aux motifs séquentiels

Introduire la date des transactions (ou du moins tenir compte de leur succession)

Peut-on produire des règles du type ?

Si « destruction véhicule » et « remboursement intégral » Alors « achat nouveau véhicule »

Étape 1 Étape 2 Étape 3

Données de transactions

Datées (au moins succession d'achats)

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)

Itemset et règles

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

Si (1, 3) Alors (2) (6, 2) → confiance = $\frac{3}{4}$ = 75%

Si (1, 3) (2) Alors (6, 2) → confiance = 3/3 = 100%

libres

Calculs très complexes, très peu de logiciels proposent cette approche

http://himalaya-tools.sourceforge.net/Spam/

Références

Wan Aezwani Wab, « Apriori and Eclat Algorithms in Association Rule Mining », Slideshare, Avril 2014.

P. Tan, M. Steinbach, V. Kumar, « Association Analysis: Basic Concepts and Algorithms », chapitre 6 de l'ouvrage « Introduction to Data Mining », 2005.

Tutoriels Tanagra consacrés aux Règles d'Association (mise en œuvre, les différentes études possibles, comparaison des logiciels).