1 Topic

Mining Association Rules from Transactions File.

Association rule learning is a popular method for discovering interesting relations between variables in large databases. It was often used in market basket analysis domain. But in fact, it can be implemented in various areas where we want to discover the associations between variables. The association is described by a "IF THEN" rule. The IF part is called "antecedent" of the rule; the THEN part correspond to the "consequent" e.g. IF onions AND potatoes THAN burger (<u>http://en.wikipedia.org/wiki/Association_rule_learning</u>) i.e. if a customer buys onions and potatoes then he buys also burger.

It is possible to find co-occurrences in the standard attribute - value tables that are handled with the most of the data mining tools. In this context, the rows correspond to the baskets (transactions); the columns correspond to the list of all possible products (items); at the intersection of the row and the column, we have an indicator (true/false or 1/o) which indicates if the item belongs to the transaction. But this kind of representation is too naive. A few products are incorporated in each

transaction	produit
1	В
1	E
1	н
2	A
2	В
2	E
2	F
3	В
3	С
3	F
3	Н

basket. Each row of the table contains a few 1 and many o. The size of the data file is unnecessarily excessive. Therefore, another data representation, says "transactions file", is often used to minimize the data file size. In this tutorial, we treat a special case of the transactions file. The principle is based on the enumeration of the items included in each transaction. But in our case, we have only two values for each row of the data file: the transaction identifier, and the item identifier. Thus, each transaction can be listed on several rows of the data file.

For our example, the items B, E and H are incorporated into the transaction 1; the items A, B, E and F into the transaction 2; etc. This data

representation is more space-saving. Only the items included in each transaction are listed.

This data representation is quite natural considering the problem we want to treat. It also has the advantage of being more compact since only the items really present in each transaction are enumerated. However, it appears that many tools do not know manage directly this kind of data representation. We observe curiously a distinction between professional tools and the academic ones. The first ones can handle directly this kind of data file without special data preparation. This is the case of **SPAD 7.3** and **SAS Enterprise Miner 4.3** that we study in this tutorial. On the other hand, the academic tools need a data transformation, prior the importation of the dataset. We use a small program written in VBA (Visual Basic for Applications) under Excel to prepare the dataset. Thereafter, we perform the analysis with **Tanagra 1.4.37** and **Knime 2.2.2** (*Note: a reader told me that we can transform the dataset with Knime without the utilization of external program. This is true. I will describe this approach in a separate section at the end of this tutorial).*

Attention, we must respect the original specifications i.e. focus only on rules indicating the simultaneous presence of items in transactions. We must not, consecutively to a bad "presence - absence" coding scheme, to generate rules outlining the simultaneous absence of some items. This may be interesting in some cases may be, but this is not the purpose of our analysis.

2 Dataset

The «<u>transactions.txt</u>» data file describes 10.000 baskets, 8 goods are referenced. Of course, it is an artificial dataset intended for the evaluation the association rule learning tools. We have already used this dataset (<u>http://data-mining-tutorials.blogspot.com/2008/11/association-rule-learning-from.html</u>). The originality here, at least with SPAD and SAS, is that we treat directly the data organized in the transactions file format.

3 Association Rule Mining with SPAD 7.3



<u>SPAD</u> is a popular tool for the French data miners. Its modus operandi is consistent with the standards of the domain. A data mining process is represented as a stream. The nodes represent the operations performed on the data.

3.1 Creating a new diagram

When we launch SPAD, a dialog setting allows us to create a new diagram.

3.2 Importing the dataset

We add the "Fichier Texte Délimité" tool into the stream canvas. We set the appropriate settings (PARAMETERS contextual menu) to



import the "transactions.txt" data file.

There are 39,000 rows and 2 columns as specified by SPAD.

3.3 Extracting the rules

The "Règles d'Association" tool enables to extract the rules. We link the data access component to it. We set the following parameters (PARAMETERS contextual menu):

SPAD 7.3 - Exploration Regles Association	
Projet Edition Affichage Diagramme Dessin Outils ?	7 %
🝞 Projet 💷 🗶 🔚 🛄 Diagramme 🗕 🗙	
Exploration Regles Association 🛛 😤 🗻 🛃 🔹 👘 🗅 🗈 👗 🗙 🔷 🔂 📾 🖓 🖓 🖓 🖓 🖓 👘 👘	🖌 🎲 🥮
Diagrammes de traitements	
Diagramme	
🛉 Méthodes 💷 🎉 🗐 🔊	
Personnaliser transactions	
🕀 🗀 Classifications - Typologies	
🕑 🗀 Amado - Graphiques de Bertin Associat 🔚 Paramétrer F4 🐂	
Scoring et Moduliation	
Arbres de decis Arbres de decis Arbres de decis	
Type de données	
🗉 🗀 Tableaux multi 🔿 Données standard 💿 Données transactionnelles (1)	
🗄 🗁 Text Mining 🛛 Index Nom Stockage Rôle Sélectionner les variables repré	sentant les transactions et les iten
Modèles struct 0 transaction Réel Continue	
Industrialisatio I produit Chaîne Nominale ID transaction : transaction	(2) •
Er	(3) -
🕑 🗀 Beta	
Paramétres	
Support minimal :	(4) 0.01 🗧
Support maximal :	
Confiance minimale :	(O) 0.75
Cardinal maximal des règles :	
Cardinal maximal du conséquent des règles :	(0) 3 💭
Afficher les erreurs	Ok Annuler

We use the transactions file format (1). The identifier of the transaction is the first column "transaction" (2). The items are described in the second column "produit" (3). We set the minimum support to 0.01 i.e. a rule must cover at least (0.01 x 10000) 100 transactions (4). The minimum confidence is 0.75 (5). Last, the max cardinal of a rule (i.e. the max number of items authorized in a rule) is set to 4; we want rules up to 3 items in the consequent (6). The calculations are started as soon as we validate the parameters: 136 rules are extracted. We visualize them by clicking on the RESULTATS / VISUALISATION DES REGLES D'ASSOCIATION contextual menu.

-ichier	Onérations	Aide										
ichici	operations	Aluc								_		
136 filtrées sur 136 règles disponibles												
Liste des règles												
Num	. Antécedent	Cons	Lon	Sup	Supp	Support.Règle	Confiance	Sensibil	- AL			
1	A&C&E	D	4	0.050	0.500	0.050	1.000	0.100	2.000	1		
2	D&L&E	À	+	0.050	0.500	0.050	1.000	0.100	2.000			
3	A&C&E	G	4	0.050	0.500	0.050	1.000	0.100	2.000			
4	G & C & E	A	4	0.050	0.500	0.050	1.000	0.100	2.000			
5	A&C&E	В	4	0.050	0.550	0.050	1.000	0.091	1.818			
6	B&A&C	E	4	0.050	0.400	0.050	1.000	0.125	2.500			
	A&C&E	F	4	0.050	0.550	0.050	1.000	0.091	1.818			
7												

For instance, the rule n°1 means

If the consumer buys (A, C and E) then it buys also (D)

The description of the rule is completed with some interestingness measures which outline the relevance of the rule. We have described some of them in a previous tutorial (<u>http://data-mining-tutorials.blogspot.com/2009/02/interestingness-measures-for.html</u>).

Each measure has its specificity. If we want to sort the rules according to the LIFT indicator (in a decreasing order, we click on the header of the corresponding column (if we click again, the rules are ordered in an increasing order).

्तु Visuali	Visualisation des règles d'association												
Fichier	Fichier Opérations Aide												
136 filtrées sur 136 règles disponibles													
Liste des règles													
Num	Antécédent	Cons	Lon	Sup	Supp	Support.Règle	Confiance	Sensibil	🔸 Lift	-			
105	D&A	F&G	4	0.350	0.350	0.350	1.000	1.000	2.857				
112	F&G	D&A	4	0.350	0.350	0.350	1.000	1.000	2.857				
6	B&A&C	E	4	0.050	0.400	0.050	1.000	0.125	2.500				
12	B & D & C	E	4	0.050	0.400	0.050	1.000	0.125	2.500				
106	G & A	F & D	4	0.350	0.400	0.350	1.000	0.875	2.500				
110	F&D	G&A	4	0.400	0.350	0.350	0.875	1.000	2.500				
70	F & B & H	С	4	0.050	0.450	0.050	1.000	0.111	2.222	Ŧ			
٠				III					- F				
-													

3.4 Interactive exploration of the rules

Additional tools enable to better identify the content of the results. We click on the menu OPERATIONS / INFORMATIONS SUR LES ITEMS. We obtain the list of items.

🔄 Visua	lisation des rè	gles d'ass	ociation	1		\sim	-		
Fichier	Opérations	Aide							
	🔄 Filtrer			/	_				
	Afficher t	outes les r	ègles		Liste des items			E	
					5 501 T	h	nformations gé	énérales	
Lis 🗌	Informati	ons sur les	s items			Nombre d'items	8		
Num	📝 Options			[Supp	Nombre de transactions	10000		
105	D&A	F&G	- 4	0.350	0.3				
112	F&G	D&A	4	0.350	0.3		Items		
6	B&A&C	E	4	0.050	0.4	Item		Supp.Absolu	Supp.Relatif
12	B&D&C	E	4	0.050	0.4	A		5000	0.500
106	G&A	F&D	4	0.350	0.4	В		5500	0.550
110	F&D	G&A	4	0.400	0.1	C		4500	0.450
70	F&B&H	C	4	0.050	0.4	F		4000	0.500
10		-	-			F		5500	0.550
						G		5000	0.500
						н		4500	0.450
						li			

SPAD provides tools for filtering the rules. By clicking on the OPERATIONS / FILTRER menu, a dialog box appears. We can filter the rules according the presence of some items into the rules, or according numerical criteria.

Let us visualize the rules containing the item A into the consequent part. We drag the item A from the list of the items and drop it into the "Consequent" branch. We click on the "Appliquer" button: 30 rules are highlighted.

📆 Visua	lisation des règ	les d'associati	on					-	l	0	x
Fichier	Opérations A	lide									
	Filtrer			1							
	Afficher to	utes les règles		Itrées su	r 136 règle	s disponibl	es				
Lis	🐻 Information	ns sur les item	s								
Numé	😡 Options			Supp	Supp	Supp	Confi	Sersi.	Lift	Valeu	<u>د</u> م
2	D&C&E	A	4	0.050,	0.500	0.050	1.000	A 10	0 2.000	22 940	100
4	G & C & E	A	4	0.050	Paramètre	s de filtrag	e				
8	F&C&E	A	4	0.050	Critères la	aiauna I c.a		1			
22	D & H & E	A	4	0.050	Cilleres io	giques Litt	eres numeriq	ues			
25	F & H & E	A	4	0.050		Liste des	items		Conditions s	sur les règles	
32	D & E	G&A	4	0.200	A			i	····· Règle		_
34	G & E	D&A	4	0.200	B				- Antécédent		
35	G & D & E	A	4	0.150					E- Conséquen	t	
1				1	E				Α		
		^			F						
3					G						
					1						
										Арр	liquer

Now, we want to filter this subset of rules by retaining only those rules with a LIFT upper than 1.8 (and less than 10, but all the rules anyway have a LIFT less than 10).

In the "Critère numérique" tab, we set the appropriate bounds. Then, we activate these new filtering parameters.

	Fichier	alisation des rè Opérations	gles d'associa Aide	tion						10000			x
		ste des règles	۵	24	filtrées s	ur 136 règle	es disponible	es			Ţ		
	Numé	éro Antécédent	Conséqu.	L	Supp	Supp	Supp	Co	nfi	Sensi	Lift	Valeu	<u>[^]</u>
	2	Paramètres de fil		000	0.100	2.000	22.940	= =					
	4		and ge	1		000	0.100	2.000	22.940	_			
	22	Critères logiques	Uniteres numeriqu	ues		000	0.100	2.000	22.940	- 11			
	25			000	0.100	2,000	22.940	- 11					
	32	Mesure	Actif	Val.Min	1	/al.Max			750	0.429	2.143	41.929	- 11
	34	# Antécédent	0		1	4			750	0.429	2.143	41.929	
	35	# Conséquent	٢		1	4		- 11	000	0.300	2.000	42.006	-
	1	Longueur	9		5	5							F
		Support.Antécéc	٢	0.	000	1.000							_
		Support.Conséqu	٢	0.	000	1.000							11
1		Support.Règle	٢	0.	000	1.000						-	
		Confiance	9	0.	000	1.000							
		Sensibilité		0.	000	1.000							
		Lift	٢		1.8	10.000							
		Valeur test	0	-10	000	10.000		- 11					
		Déploiement	9	0.	000	1.000							
		U					Appliq	uer					

We click on the "Appliquer" button. We obtain 24 rules.

Like this, we can highlight the subset of rules which is the most relevant according to the goals and the constraints of our analysis.

4 Association Rule Mining with SAS EM 4.3

I am not an expert in SAS, even less with regard "Enterprise Miner". The description proposed here can appear simplistic sometimes. The most important thing is to obtain the desired result.

4.1 Importing the data file

First step, we want to import the "transactions.txt" data file into the WORK data library (*Note: it is more appropriate to create a specific data library, but we use the most basic way in this tutorial*). To do this, after we launch SAS, we click on the File / Import Data menu. We select the "Tab Delimited File (*. txt)" format. We set the TRANSACTIONS as the name of the dataset into the WORK library.

4.2 SAS Enterprise Miner

We click on the SOLUTIONS / ANALYSE / ENTERPRISE MINER menu to launch the Data Mining module of SAS. Usually a project has already been created with a default diagram. We set "Induction of Assoc Rule" as the name of the diagram.

The "Input Data Source" component enables to import the dataset. We add it into the workspace. We open the dialog settings by clicking on the OPEN menu.

🔀 SAS - [SAS Enterprise Miner - Test Assoc Rule [Induction o	f Assoc Rule]]
Fichier File Edit View Options Actions Help	Fenêtre _ & ×
 Comparison Comparison	< 🗈 🖨 🛍 📍 🥘
■I• ■ ■ * ※ ½ &	* 😹 🏂 🖹 💂
Test Assoc Rule	Open Run
	Copy Delete Clone About
Diagrams Tools Reports ()	Connect items Move items Move and connect *
Sortie - (Sans titre)	Editeur - sans titre1
	C:\Users\Maison

We select TRANSACTIONS from the WORK library as "Source Data". Into the VARIABLES tab, we specify the role of the columns using the SET MODEL ROLE contextual menu: TRANSACTION is the ID; PRODUIT is the TARGET.



Then we add the ASSOCIATION component into the stream canvas. We set the following settings.

😽 SAS - [SAS Enterprise Mir	ner - Test Assoc Rule [Induction of Assoc Rule]]
M Fichier File Edit Vi	ew Options Actions Help Fenêtre
~	🗸 🖆 🔛 🗙 🕴 🕄 🖨 🛍 🛛 ? 🥘
	* 💥 🖄 🏤 ≽ 🖹 🛃
Test Assoc Rule	Rule
T	Association
DiagramsTor	Output Selected Output Notes Data Variables General Sequences Time Constraints Sort (1) Analysis mode: CBy Context @Association C Sequences
Journal - (Sans titre	Minimum Transaction Frequency to Support Associations: 5% of largest single item frequency Specify as a percentage: 9 Specify a count: 100 (2)
	Maximum number of items in an association: 4
	Minimum confidence for rule generation: $75 \times (3)$

We want to obtain association rules (1). The minimum support of the rule is 100 (2). The maximum cardinal of the rules is 4. And the confidence minimum is 75% (3).

We confirm these choices. Then we click on the RUN menu. The rules are listed in a new visualization window.

🕱 SAS - [Res	ults - Assoc	ciation]					
🔄 Fichier	File Edit	View	Fenêtre				- 5 :
~				- 🖨 l	🖻 🕨 k k	? 🖑 🤇	💠 🕹 🖬 🖇
Rules	Frequen	cies	Code	Log N	lotes		
	Relations	Lift	Support(%)	Confidence(%)	Transaction Count	Rule	•
1	2	1.60	40.00	80.00	4000.0	G ==> D	
2	2	1.60	40.00	80.00	4000.0	D ==> G	
3	2	1.45	40.00	80.00	4000.0	D ==> F	
4	2	1.45	40.00	80.00	4000.0	A ==> F	
5	3	2.00	35.00	100.00	3500.0	G & F ==> D	
6	3	1.59	35.00	87.50	3500.0	G & D ==> F	
7	3	1.75	35.00	87.50	3500.0	F & D ==> G	
8	3	2.00	35.00	100.00	3500.0	G & F ==> A	
9	3	1.82	35.00	100.00	3500.0	G & A ==> F	
10	3	1.75	35.00	87.50	3500.0	F & A ==> G	
11	3	1.75	35.00	87.50	3500.0	G & D ==> A	
12	3	2.00	35.00	100.00	3500.0	G & A ==> D	-
*							•
Journal	- (Sans titre)	诺 Editeu	r - sans titre1	😼 Result	s - Associatio	on
					C:\User	s\Maison	

Like SPAD, 136 rules are extracted. We can sort them also according to the LIFT criterion.

7 0	Des	elect All					• ×
	Prin	nt Preview		Fenêtre			- 5
~	Prin	nt	/		- 🕘 🖣	D N N? 🖑 🤇 🗢 🖫	a No 1
Rı	Sor	t 👘		Code	Descendir	ig	
			\checkmark		Descendir	g Within Relations	
		Relations	Lift	Support(%)	Ascending	Within Relations	~
L	1	4	2.86	35.00		,	
L	2	4	2.86	35.00	Ascending	3	
L	3	4	2.50	5.00	100.00	500.00 C & B & A ==> E	E
L	4	4	2.50	35.00	100.00	3500.0 G & A ==> F & D	_
L	5	4	2.50	35.00	87.50	3500.0 F & D ==> G & A	
	6	4	2.50	5.00	100.00	500.00 D & C & B ==> E	
	7	4	2.22	5.00	100.00	500.00 H & G & B ==> C	
[8	4	2.22	5.00	100.00	500.00 G & E & B ==> C	
[9	4	2.22	5.00	100.00	500.00 H & F & B ==> C	
[10	4	2.19	35.00	87.50	3500.0 F & A ==> G & D	
[11	4	2.19	35.00	87.50	3500.0 G & D ==> F & A	
[12	4	2.14	15.00	75.00	1500.0 E & D ==> G & F	-
	•						•
	Journal	- (Sans titre)		Editeur	r - sans titre1	Results - Association	
						C:\Users\Maison	

Into the FREQUENCIES tab, we have the list of the items and their count.

😽 SAS - [Re	esults - Associa	tion]				• X
😼 Fichier	File Edit	View Fen	être			- 8 ×
~			- 6	🏐 🖻 🕨 🖎	k? 🖉 🤇	🔶 🖁 🗧
Rules	Frequenc	ies) Co	de Log	Notes		
	Count	Item				
1	5500	F				
2	5500	В				
3	5000	G				
4	5000	D				E
5	5000	A				
6	4500	H				
7	4500	С				
8	4000	E				
						-
Journa	al - (Sans titre)	₹ E	diteur - sans ti	trel 😼 Re	sults - Associat	ion
				C:\Users\	Maisc	

Like SPAD again, we can display only a part of the rule set using logical or numerical criteria. We click on the VIEW / SUBSET TABLE for that.

Left Hand Side:		Right Hand Side.	Excitide her	
Transaction Count Rule		Transaction Count	Rule	^
5500.0 F		5500.0 F		
5500.0 B	_	5500.0 B		
5000.0 G	E	5000.0 G		=
5000.0 D	_	5000.0 D		-
5000.0 A	_	5000.0 A		
4500.0 H		4500.0 H		
4500.0 C		4500.0 C		
4000.0 E	-	4000.0 E		-
4	· .	*		+
Туре:		Type:		
1			Ū	

5 Association Rule Mining with Tanagra and Knime

In this section we show how to transform the dataset prior to the importation into Tanagra and Knime. Again, the objective of the analysis is to extract only the positive rules (the items cooccurrence in the transactions). We create an intermediate data file where we use a o/1 coding scheme to specify the presence of the items into the transactions.

We have already shown how Tanagra handles this kind of format for the extraction of association rule (<u>http://data-mining-tutorials.blogspot.com/2008/11/association-rule-learning-from.html</u>). In

this tutorial, we propose a very simplistic program (in VBA) to transform the transactions file into this format. Thereafter, we show how to treat the transformed data file with Tanagra and Knime. *A priori*, the obtained rules must be the same as the ones extracted with SPAD and SAS.

5.1 Transforming the data file

We use the following program to transform the dataset under Excel.

0		(* •) =	1		tran	sactions.xlsm -	Microsoft Excel		_				X
r@	Accueil	Insertion	Mise en pa	<i>.</i>	Formules Don	nées Dévisi	on Affichage	e Développ	peur Comp	léments	Acrobat	0 -	⊐ x
Vis Ba	ual Macros	Enregistrer u Utiliser les ré Sécurité des r Çade	ne macro Ferences relat macros	tives I	nsérer Mode • Création	Propriétés Visualiser le c Exécuter la bo Contrôles	ode oîte de dialogue	Source	Propriétés du m Kits d'extension Actualiser les de XML	nappage 📑 onnées	Importer Exporter		
	C4	- (0	fx										×
		0	C	D	-		G			-			-
1	transaction	produit		A	B	C	D	E	F	G		н	Ê
2	1	в		0	1	0	0	1	0	0		1	
3	1	E		C **									
4	1	н			transactions.xlsm -	Module1 (Cod	e)				-		X
5	2	A		(0	Général)			▼ Trans	ToGrid				-
6	2	в			Public Sub	TransToGri	.d ()						
7	2	E			'activer la	bonne feu	ille						_
8	2	F			Sheets("trans2bin").Activate Dim i As Long, numProd As Long, id As Long, prevId As Long, i As Long								
9	3	В			'désactiver le rafraîchissement de l'écran								
10	3	С			Application.ScreenUpdating = False								
11	3	F			'initialisa	tion							
12	3	н			'passer en	revue les	couples "tr	ansaction-	-produit"				
13	4	A			For $i = 2$ I	o 39001 St	ep 1						
14	4	D			'id de	transactio	n						
15	4	F			ia = Ce	r la ligne	de 0 si no	uvelle lic	me				
16	4	G			If (id	> prevId)	Then						
17	5	В			For	; j = 4 To	11 Step 1						
18	5	D			New	Cells(id	+ 1, j).Val	.ue = 0					
19	5	E			'ac	tualiser 1	a transacti	on traitée					
20	5	F			pre	vId = id							
14 4	+ H trans	2bin / bin2t	rans 🖉 🞾		End If	de mondord	-						
Prê	: 🛅				numero	ae produi = Asc(Cel	.t ls(i, 2).Va	lue) - 64					
					'inscri	re la vale	ur	140) 01					
					Cells (i	d + 1, num	Prod + 3).V	alue = 1					
					Next i	vafraichi	agamant de	Llágran					
					Application	.ScreenUpd	lating = Tru	e ecran					
					End Sub								
				20									
													<u> </u>

The program is rather simple. It relies on two elements to generate the binary table: the transaction ID indicates the row that must be completed; the ASCII code of the product name (this solution is highly specific to our data representation) to detect the column to fill. All the blanks cells are filled with zeros. We export the new data table into the "transactions_binary.txt" text file. Here are the first rows of the data file.

Fichier	Edition	Format						
la.		1 Olillar	Affichage	?				
-	в	С	D	E	F	G	н	
0	1	0	0	1	0	0	1	
1	1	0	0	1	1	0	0	
0	1	1	0	0	1	0	1	
1	0	0	1	0	1	1	0	
0	1	0	1	1	1	0	0	
1	1	Ō	1	0	1	1	Ō	
0	0	1	1	0	0	1	1	
1	Ō	1	1	Ō	1	1	ō	
0	1	1	0	0	0	1	1	
	-	_		-		_	_	Ŧ
4)	i.

5.2 Performing the analysis with Tanagra

We launch Tanagra. Then, we create a new diagram by clicking on the FILE / NEW menu. We select the "transactions_binary.txt" data file.

Choos	e your dataset and start downloa	d				
TANAGRA 1.4.37						
File Diagram Winc	Diagram title :					
	Default title					
New						
C On an	Data mining diagram file name	:				
Open inin	D:\Temp\Exe\default.tdm					
Save				\sim 1		
Save as	Dataset (*.txt,*.arff,*.xls) :			<u> </u>		
	tasets\aassociation_rule_datas	et\test_assoc_ru	le\transactions_bin	ary.txt 🛃		
Close		Tanagra				×
Evit		Terrogra				
		Regarder dans :	test_assoc_rule		👻 🎯 🤌 🗁 🛄	•
		-	Nom Date	a da m	Taille Moto-cl	ár -
			Nom Date	e de m Type	Tallie Wots-ci	6
		Emplacements	transactions tot			
		récents	transactions bin	harv.bt		
		_				
Data visualization	Statistics					
		Bureau				
Feature construction	Feature selection					
PLS	Clustering	100				
		100 A				
Spy learning assessment	Scoring	Maison				
A Correlation scatternlot	Scatterplot					
Export dataset	Scatterplot with lab	200				
		Ordinateur		K		
			Nom du fichier :	transactions binary b	t 🔹	Ouvrir
					-	
			Types de tichiers :	Text file	•	Annuler

We must specify the role of the columns. We use the DEFINE STATUS component. We set the 8 variables as INPUT.

🐨 File Diagram Component Window Help
Define attribute statuses
Data visualization Statistic: Feature construction Feature sele PLS Clusterin Spv learning assessment Scoring Correlation scatterplot Scatterplot Scatterplot Scatterplot Scatterplot OK

We insert the A PRIORI component (ASSOCIATION tab) into the diagram. We set the following settings (PARAMETERS contextual menu): (Support min = 0.01; Confidence min = 0.75; Max Cardinal = 4; Lift min = 0.0).

TANAGRA 1.4.37 - [Dataset (transactions_binary.txt)]					
Tile Diagram Component Window Help	_ <i>8</i> ×				
	Association rule parameter				
Default title Define status 1 Database : D:\DataMinin Parameters Execute View Com	Parameters Support: 0.01 Confidence: 0.75 Max card itemsets: 4				
Data visuadzation Statistics Nonpa	3				
Feature construction Feature selection PLS Clustering Spv learning assessment Scoring Starting Spv learning Starting Scoring Starting Scoring	OK Cancel Help				

We confirm and we click on the VIEW menu to launch the calculations.

File Diagram Comp	ori 1j onent Window Help			- Carrier		-	8 >
🗅 📾 🔚 🎎 Default t	itle						
Dataset (transactio	ons binary.txt)	ULES					
E Ta Define status 1			Number of ru	iles : 136			
	N	Antecedent	Consequent	Lift	Support (%)	Confidence (%)	
	Parameters 1	"D=true" - "A=true"	"F=true" - "G=true"	2.85714	35.000	100.000	
	Execute 2	"F=true" - "G=true"	"D=true" - "A=true"	2.85714	35.000	100.000	
	View 3	"F=true" - "D=true"	"G=true" - "A=true"	2.50000	35.000	87.500	
	4	"B=true" - "D=true" - "C=true"	"E=true"	2.50000	5.000	100.000	
	5	"B=true" - "A=true" - "C=true"	"E=true"	2.50000	5.000	100.000	-
		Components					
Data visualization	Statistics	Nonparametric statistics	Instance sele	ection Fe	Feature construction		
Feature selection	Regression	Factorial analysis	PLS		Clustering		
Spv learning	Meta-spv learning	Spv learning assessment	Scoring		Association		
📽 A priori 🛛 📽 A priori MR 📲 A priori PT 👘 Assoc Outlier 🅞 Spv Assoc Rule 😓 Spv Assoc Tree							

We obtain 136 rules, like SPAD and SAS. We note that Tanagra generates only the positive rules which highlight the simultaneous presence of the items in transactions. The rules are sorted according the LIFT criterion (decreasing order). Unlike SPAD or SAS, we cannot handle interactively the rules.

5.3 Performing the analysis with Knime

We create a diagram into Knime. We import the data file using the FILE READER component.

File Edit View Search Run Node Help	
📑 - 🔛 🔞 🛷 - 👘 - 💱 - 🤛 -	Q, - <> \> 100% - & Q, \c ፼ = ▷ ① ⊕ ③ ② ④
🚵 Workflow Projects 🖉 🗘 🐼 🖪 😫	🗢 🗖 🗘 0: Association 🕼 *2: Assoc Rules Binary Dat 🛛 🗖 🚺 Node Description 🕅 🦷 🗖
Assoc Rules Binary Dataset Association Heat reporting Multithreading Decision Tree Multithreading Decision Tree Note Repository	File Reader Configure Node Configure Execute File Reader File Reader This node can be used to read data from an ASCII file or URL location.
▼	It can be configured to read various
	Dialog - 2:1 - File Reader
 File Reader ARFF Reader CARFF Reader Table Reader Table Reader Model Reader Write Artificial Data Cache Database Data Manipulation Data Views Statistics 	File Settings Enter ASCII data file location: (press 'Enter' to update preview) valid URL: //association_rule_dataset/test_assoc_rule/transactions_binary.txt • Basic Settings read row IDs Column delimiter: // read column headers Java-style comments Single line comment:
	Click column header to change column properties (* = name/type user settings)
	Row ID I A I B I C I D I E I F I G Row 0 0 1 0 0 1 0 0
	OK Apply Cancel

Another transformation is needed to generate association rules. We use the BITVECTOR GENERATOR component for this. It transforms each transaction into a vector of bits. We set the following parameters (CONFIGURE menu)

A KNIME		
File Edit View Search Run Node H	Help	
📑 - 🖬 💼 🛷 - 🖢 - 🖗 -	🍫 💁 - 🤣 😂 100%	🛆 Dialog - 2:2 - Bitvector Convertor
🚵 Workflow Projects 📃 🗆	▲ 0: Association ▲ *2: Assoc	File
(File Reader	Default Settings Memory Policy
Assoc Rules Binary Dataset	Bitvector Generator	Numeric input (many columns)
Association		Threshold: 1
		Use percentage of the mean:
🛕 Node Repository 📃 🗆	Node 1	Percentage: 100 📩
- ☆ ↔ ▽	Node 2	Exclude Select Indude
📒 Database 🔷	(C	Column(s): Search add >> Column(s): Search
🗟 Data Manipulation	B	Select all search hits
Column		
🔁 Binning 📰	9	
🗁 Convert & Replace	8	
🗁 Filter	,	< <re>enve</re>
🗁 Split & Combine 👘	QQ	I E
🔁 Transform	E Outline 🖾	<< remove all
HiLite Collector		I G
Row		· · ·
🗁 Filter	The second secon	Bits from string column
Transform		Parse bitvectors from strings (one column)
Bitvector Generator		
Concatenate 👻	69	String column to be parsed
4	×	Kind of string representation: BIT 🚽
	Q	General
		Remove column(s) used for bit vector creation
		OK Apply Cancel

We use the INTERACTIVE TABLE component to visualize the new column. BITVECTORS corresponds to a vector of bits in a hexadecimal encoding scheme.



We can launch the extraction of the rules using the ASSOCIATION RULE LEARNER component. We set the following settings (CONFIGURE menu).



The transactions are described by the BITVECTORS column (1). The minimum support of the generated rules is 0.01 (2). The maximal itemset length is 4 (3). Last, we want to generate rules (4) with a confidence upper than 0.75 (5).

We click on the EXECUTE menu to launch the calculations. We can visualize the rules with the INTERACTIVE TABLE component.



Unlike the other tools, Knime generates only the rules with one item into the consequent. We obtain 94 rules.

We can filter the rules with the ROW FILTER component. We want to visualize the rules with a LIFT upper than 1.8 (and lower than 10.0). We parameterize the component as follows.



The filtering is based on the LIFT criterion (1). We set the bounds of the values (2).

We use another INTERACTIVE TABLE component in order to visualize the filtered rule set.



We can refine the filtering. For instance, based on this new set of rules, we can highlight the ones with the item A into the consequent. We add the ROW FILTER component. The settings are specified about the consequent of the rule here.



Now, we have only 16 rules.



5.4 Performing the analysis with Knime (again)

After the publication of the French version of this tutorial, Loïc LUCEL told me that it was possible to generate within Knime the o/1 data table. We can therefore avoid the priori recoding using an external program (in VBA here). The operation can be included directly into the stream diagram. Thank you very much for these indications Loïc.

Let's see this new approach. We create a new Workflow Project. Using the FILE READER component, we import the transactions file "transactions.txt". We note that Knime recognizes Knime automatically the transaction ID as integers [Integer], the products column as strings [String].

File Edit View Search Run Node Help	Dialog - 21 File Reader
📑 🖬 📓 🛷 🔹 🖗 🖛 🖓 🗣 🖗 🗸 🖗 🖗	File
Workflow Projects Workflow Projects Using Pivoting Component Using Pivoting Component Node Repository Io File Reader Io File Reader Stable Reader Write Model Reader Write Artificial Data Cache Image: Cache<	File Settings Enter ASCII data file location: (press 'Enter' to updata preview) valid URL: stasets/aassociation_rule_dataset/test_assoc_rule/transactions.txl • Basic Settings read row IDs Column delmiter: <ta> V read column headers V ignore spaces and tabs Java-style comments Single line comment: Preview Click column header to change column properties (* = name/type user settings) Row ID I transac § produit Row1 1 Row2 1 Row4 2</ta>
	OK Apply Cancel

Then, we add the PIVOTING component (DATA MANIPULATION / ROW / TRANSFORM branch).

				Dialog - 2:2 - Pivoting
				File
				<u> </u>
				Options Memory Policy
	File Rea	der		Pivot (column header)
			Pivoting	Pivot column: S produit
			> 🛐 Þ	Group (row header)
	Mode	1		Group column: 🚺 transaction 💌
			Node 2	Aggregation (table content)
				Count co-occurrences
				Count co-occurrences
				C Enable aggregation
				Aggregation column: 🗍 transaction 👻
				Aggregation method: SUM -
				Advance
				Enable hiliting
				☑ Ignore missing values
				OK Apply Cancel
		File Reader	Pinoti	Interactive Table
			<u> </u>	
		(in in initial contraction of the second sec	0.00	
		Node 1	Node	2 Node 3
ble View	v - 2:3 - Intera	ctive Table		
ninte N	B	E E	TH I	IA IF IC ID IG
	1	1	1 0	
	1	1	0	
	0	0	0	
	1	1	0 0	1 0 1 0
	1	0	0	
	v	v	-	v 1 1 7

Dialog - 2:2 - Pivoting

About the settings, we set PRODUIT as PIVOT, TRANSACTION as GROUP. Thus, we count the presence of each item for each transaction. Since the transactions file counts only the presence (or absence) of the item, the possible values are o or 1.

We check this with the INTERACTIVE TABLE. We have indeed the appropriate table with the o/1 coding scheme.

Therefore, as in the previous section, we have to add a BIT VECTOR GENERATOR after the PIVOTING and extract component the association rules by following the same approach.

The main advantage of the procedure described in this section is that the data transformation is into the included automated process. Thus, it remains applicable if the source file is modified i.e. if we want to refresh the transactions file and perform a new extraction of the rules.

Out of curiosity, we show in this subsection the reverse transformation, from the binary table to the transactions table. As we can see, the VBA code is very simple also. The generalization to other databases, with more items and transactions is easy.



6 Conclusion

The ability to handle transactions file is an asset for the extraction of association rules. The ease of operations with SPAD and SAS shows this. Curiously, this feature is lacking in some generalists (academic) tools.

In this tutorial, we show that the data transformation in order that free tools such as Tanagra can handle this kind of dataset is rather simple. We can use a little program written in VBA (under Excel). The limitation is not too restrictive. We remind that Excel (version 2007 and later) can treat 1,048,575 "transaction id - item id" pairs (we do not count the first row) and 16,381 items (if the start the binary table form the D column into the worksheet). We can do many things with that.