### 1 Introduction

#### Implementing the APRIORI MR component.

Association rule learning is a popular method for discovering interesting relations between variables in large databases (http://en.wikipedia.org/wiki/Association\_rule\_learning). It was often used in market basket analysis domain e.g. *if a customer buys onions and potatoes then he buys also beef.* But, in fact, it can be implemented in various application areas where we want to discover the association between variables.

We were already described the association rule mining tools of Tanagra in several tutorials (<u>http://data-mining-tutorials.blogspot.com/search/label/Association%20rules</u>). The A PRIORI approach is certainly the most popular. But, despite its good properties, this method has a drawback: the number of obtained rules can be very high. The ability to underline the most interesting rules, those which are relevant, becomes a major challenge.

These last years, numerous interestingness measures for association rules are studied. The goal is to associate a numerical indicator to rules. The A PRIORI MR component (ASSOCIATION tab) is an experimental tool which supplies several interestingness measures for evaluating rules. There are widespread measures such as confidence, support, etc; there are also less known measures such as those based on the test value principle. The theoretical foundations of these last measures are described in various tutorials available online http://data-miningtutorials.blogspot.com/2009/02/interestingness-measures-for.html and http://data-miningtutorials.blogspot.com/2009/05/understanding-test-value-criterion.html.

In this tutorial, we show to implement the A PRIORI MR component, how to set the parameters in order to obtain more or less rules, and how to read the results.

## 2 Dataset

We use a modified version of the GERMAN CREDIT dataset in this tutorial. It depicts the characteristics of customers [http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)]. We do not try to explain in particular a variable in our framework. We try mainly to underline the interdependence between the variables (http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/credit\_assoc.xls).

## 3 The A PRIORI MR Component

### 3.1 Creating a diagram and importing the dataset

Tanagra can load directly an Excel file format (XLS) even if the Excel software is not available on our computer. There are two restrictions for handling the data file: it must not be currently opened in other tool; the dataset must be in the first worksheet (see <u>http://data-mining-tutorials.blogspot.com/search/label/Data%20file%20handling</u>).

Into Tanagra, we click on the FILE / NEW menu; we select the CREDIT\_ASSOC.XLS data file.

anagra
<u> </u>

File Diagram Window New	se your dataset and start do	wnload				
Save Save Save as Close Exit	Diagram title : Default title Data mining diagram file n [D:\DataMining\Databases Dataset (*.bt.*.arff,*.xls) : [for_soft_dev_and_compa	ame : _for_mining)dataset_fo	v assoc\credit	nd_compa 🖬 assoc.xls ጅ		
		ОК	Can Tanagra Regarder dans :	cel Help	·	G 🕽 🐂 🗔 -
Data visualization Feature selection Spv learning	Statistics Regression Meta-spv learning	Component Nonparametric stat Factorial analys Spv learning assess	Emplacements récents	Nom Date de m.	Type Tai	lle
Correlation scatterplot Export dataset	Catterplot with label Over dataset	plot	Bureau Maison			
			(Mainateur) Ordinateur			
			<u>.</u>	Nom du fichier. credi Types de fichiers Exce	t_assoc.xls I File (97 & 2000)	<ul> <li>■ Ou</li> <li>■ Anr</li> </ul>

17 attributes and 1000 individuals from the DATASET sheet are now available for the analysis.

### 3.2 A PRIORI MR

We insert the DEFINE STATUS component into the diagram, using the shortcut into the tool bar. We set all the variables as INPUT.

Tile Diagram Component Window Help	TANAGRA 1.4.30 - [Dataset (credit_assoc.xls)]	_ 0 <b>_ X</b>
Default title       Secretor since       Unissee         Sheet size       1001 x 17         Define status 1       Define attribute statuses         Parameters       Attributes:         Attributes:       Target         Image: status 1       Personal, status         Image: status 1       Personal, status 1         Image: status 2       Personal, status 2         Image: status 2       Personal, worker         Image: status 2       Pe	🏆 File Diagram Component Window Help	_ & ×
Default ble     Default ble       Sheet size     1001 x 17       Sheet size     1001 x 17       Sheet size     1001 x 17         Parameters       Attributes :     Target       Image: Status     Image: Status       Image: Stat		
Dataset (credit_assoc.xls)       Define attribute statuses         Parameters       Attributes :         Attributes :       Target Input Illustrative         Oher_parties       Oher_parties         personal_status       Oher_parties         property_magnitude       other_parties         potentiation       Statis         Image: Spy learning       Meta-spy         Image: Spy learning       Meta-spy         Image: Spy learning       Scatterplot		
Define status 1 Parameters Attributes : Parameters Attributes : Property_magnitude other_payment_plans housing Property_magnitude other_payment_plans housing existing_credits Spv learning Meta-spv Clear selection Statis Export dataset Sy View dat Scatterplot Scatterplot Cite Scatterplot Cite Scat	Sheet size 1001 x 17	
Data visualization       Statis         Feature selection       Regres         Spv learning       Meta-spv         Correlation scatterplot       Scatterplot         Export dataset       Wiew dat         Wiew mut       OK	Define attribute statuses	
	Define status 1       Define status 1         Parameters         Attributes :         Image: status 1         Personal_status 1         Image: status	uction 3

Then, we add the A PRIORI MR component (ASSOCIATION tab). We click on the VIEW menu. We obtain a first result with the default settings.



32 rules are generated. They are sorted in a decreasing order according the LIFT measure.

### 3.3 A PRIORI MR settings

Several essential settings are displayed in the PARAMETERS part of the report.

TANAGRA 1.4.30 - [A prior	ri MR 1]				
💇 File Diagram Compo	nent Window	Help			_ 8 ×
D 📽 🖪 👫					
Default title					
Dataset (credit asso	ac v(r)		A prio	ri MR 1	
Define status 1	,		Param	eters	
E A priori MR 1		A-Priori p	arameters		
		Support min	0.33		
		Confidence	min 0.75		
		Max rule len	igth 4		
		Lift filtering	g 1.10		
		Learning se	t ratio 1.00		
		Repetition	1		
			Res	ults	-
		•			4
			Components		
Data visualization	Statisti	ics	Nonparametric statistics	Instance selection	
Feature construction	Feature se	lection	Regression	Factorial analysis	
PLS Clus		ing	Spv learning	Meta-spv learning	
Spv learning assessment	Scorin	ng	Association		
🕄 A priori 🛛 🕄 A pr	iori PT 🛛 😩	Spv Assoc Ru	ule		
🕄 A priori MR 🔹 Asso	oc Outlier 🛛 📛	Spv Assoc Tr	ree		
				141 141 1441 1	

- SUPPORT MIN states the minimum support of the mined rules;
- CONFIDENCE MIN states the minimum confidence;

- MAX RULE LENGTH states the maximum number of items (attribute = value) of rules ;
- LIFT FILTERING states the minimum of LIFT.

These settings allow to restrict the number of mined rules. The first 3 parameters determines the computation time and the memory occupation; the last one (LIFT) filters only the displayed rules.

- LEARNING SET RATIO states the proportion of the dataset used for the learning phase. Indeed, Tanagra can subdivide the dataset into two parts: the first is used to extract the rules; the second is used to asses them. This framework is often used in the supervised learning task. Here, it allows to evaluate the reliability of the numerical indicator associated to the rules. The default value is 1 i.e. all the observations are used in the learning phase.
- REPETITION is the number of the replication of the Monte-Carlo procedure for the computation of the VT-100.

#### 3.4 The results supplied by the A PRIORI MR Component

The ITEMS part describes the number of mined frequent itemsets (those of which the support is higher than SUPPORT MIN), gathered by cardinality. The total number of items is 66, 19 of them are frequent. For the itemsets of cardinality equal to two, we have 68; etc.

TANAGRA 1.4.30 - [A pric	ori MR 1]				
💇 File Diagram Compo	onent Window	Help			_ & ×
🗅 📽 🔚   🎇					
Default title					
Dataset (credit_ass	oc.xls)	ITEMS			
🚊 🔁 Define status 1		Transaction	ns 1000		_
🔤 🕄 A priori MR	1	Countir	ng items		E
		All items	66		
		Filtered ite	ms 19		
		Counting	; itemsets		
		card(itemse	et) = 2 68		
		card(itemse	et) = 3 91		
		card(itemse	et) = 4 36		
		Ru	iles		
		Number of	rules 32		
		٠ III			
		L	Company		
Data visualization	Statist	ics	Nonparametric statistics	Instance selection	
Feature construction	Feature se	lection	Regression	Factorial analysis	
PIS Clust		ring Sov learning Meta-cov learning		Meta-spy learning	
Spv learning assessment	Scori	ng	Association		
B A priori B A p	riori PT 🛛 🕄	Spy Assoc R	ule		
A priori MR	oc Outlier 🗄	Spv Assoc Ti	ree		
1					

We can decrease the number of extracted itemsets: increasing the SUPPORT MIN parameter, it influences the number of itemsets extracted whatever the cardinality; decreasing the MAX RULE LENGTH, the itemsets with high cardinality are not computed.

In a second phase, Tanagra computes the rules starting from the itemsets with cardinality higher to two. The CONFIDENCE MIN can take effect here. If we increase its value, we obtain fewer rules.

The **RULES section** describes the rules. The rule is subdivided in two parts: the antecedent part corresponds to the premise of the rule; the consequent part corresponds to the conclusion. Then, we have various numerical indicators which allow to assess the reliability of the rule. Some of them are well known (confidence, support, lift ...), the others are supplied only by Tanagra (test value).

TANAGRA 1.4.30 - [A priori	i MR 1]													3
💇 File Diagram Compon	nent Window	Help											- 8	×
Default title														*
Dataset (credit_asso	c.xls)	ROLLS												
Define status 1		Rules	evaluation											
		N°	Antécédent	с	onséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Levera	ш
		"other 1 "credi paid"	_parties=none" - :_history=existing	"other_pa - "existing	ayment_plans=none" g_credits=one"	1000	472	522	370	0.3700	0.7839	1.5017	0.1236	
		2 <sup>"credi</sup> paid"	2 "credit_history=existing paid"		"other_payment_plans=none" - "existing_credits=one"		530	522	415	0.4150	0.7830	1.5000	0.1383	
		3 <sup>"other</sup> - "exis	3 "other_payment_plans=none" - "existing_credits=one"		istory=existing	1000	522	530	415	0.4150	0.7950	1.5000	0,1383	
		"foroir	III										•	Ť
				Compone	ents									
Data visualization	Statis	tics	Nonparametric st	atistics	Instance sele	Instance selection		Feature construction			tion			
Feature selection	Factorial anal	ysis	PLS			Clustering								
Spv learning Meta-spv learning Spv learn				ssment	Scoring				Ass	ociation				
A priori Assoc A priori MR	😫 A priori 📲 Assoc Outlier ﷺ A priori MR 📲 Sov Assoc Rule													
8 A priori PT	Assoc Tree													
L						.41	at	atat	atatatit.		at atati	at ada a	dat atat	H-AA

4 Fine tuning the parameters of A PRIORI MR

### 4.1 How to obtain fewer rules?

Assoc rule MR parameters	
Parameters	
Support : Confidence :	0.33
Max card itemsets : Lift :	4
Repetition :	1
ОК	Cancel Help

We want to obtain fewer rules using the MIN CONFIDENCE parameter. We click on the PARAMETERS contextual menu. We set 0.9 i.e. the itemsets with a support lower than 90% are not extracted. Then, we click on the VIEW menu. We obtain now 8 rules.

#### Tanagra



#### 4.2 Exploring the rules

Filtering the rules is an interesting functionality. But, we must know how to set appropriately the parameters. It is not obvious. Tanagra supplies an option which allows to deeply explore the extracted rules. We can copy the results in a spreadsheet, then we can use the abilities of Excel to organize (sort) the rules in different ways, according various interestingness measures.

We click on the COMPONENT / UNFORMATTED COPY in the main menu. Then, we launch EXCEL. We paste the results. With the various tools supplied by Excel, we can explore deeply the rule base. In the screenshot below, we have sorted the rules according the LEVERAGE criterion.

<b>X</b> N	licrosoft	Excel - Classeu	ır1						-						
8	<u>F</u> ichier	Edition Af	fichage <u>I</u> nserf	tion Form	na <u>t O</u> util	s <u>D</u> onné	es Fe <u>n</u> êt	re <u>?</u> Sij	pina Tan	agra Ado	be <u>P</u> DF	Tapez une	question	•	_ 8 ×
D	□ ☞ 冒 品 戦 ● 国 ♥ 以 時 商・ダ ロ・マー 急 Σ・弱 対 科 雌 刷 100% - ・ ⑦ 、 気 デ .														
Ari															
-	H3	-	f.	1 0				000 ,00 +,	0   =;- =;-	1	-				
	A	В	C	D	E	F	G	Н		J	K	L	М	N	0 —
1	A-Priori	parameters		_	_		_			-		_			
2	Suppor	0.33													
3	Confide	0.9													
4	Max rul	4													
5	Lift filte	1.1													
6	Learnin	1													
7	Repetit	1													
8	Sample	Samples size	•												
9	Training	1000													
10	Test	0													
11	Numbe	8													
12	Rules e	valuation													
13	N°	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importanc	Conviction	Surprise	VT-Hy
14	6	credit_history	existing_cred	1000	530	633	478	0.478	0.9019	1.4248	0.1425	1.006	3.7406	0.673	5.74
15	8	foreign_worke	existing_cred	1000	510	633	459	0.459	0.9	1.4218	0.1362	0.93	3.67	0.6445	5.51
16	1	other_payme	existing_cred	1000	452	633	415	0.415	0.9181	1.4505	0.1289	0.8364	4.4834	0.5972	5.22
17	5	other_parties	existing_cred	1000	472	633	427	0.427	0.9047	1.4292	0.1282	0.841	3.8494	0.6035	5.13
18	7	foreign_worke	existing_cred	1000	458	633	413	0.413	0.9017	1.4246	0.1231	0.7982	3.7352	0.5814	4.99
19	3	foreign_worke	existing_cred	1000	432	633	396	0.396	0.9167	1.4481	0.1225	0.787	4.404	0.5687	4.97
20	2	other_parties	existing_cred	1000	403	633	370	0.37	0.9181	1.4504	0.1149	0.7343	4.4818	0.5324	4.82
21	4	class=good -	existing_cred	1000	361	633	330	0.33	0.9141	1.4441	<b>V</b> 0.1015	0.6564	4.2738	0.4724	4.36
192	22 I ← → N Feuil1 / Feuil2 / Feuil3 /														
Deg	isin • 🖟	Formes auto	omatiques 🔹 🔨		0 🖾 🔺	1 🗘 🙍	- 🖄 🔝	🔏 • 🗚	• =	≓∎(					
Prêt														NUM	/

Tanagra

#### 4.3 Subdividing the dataset into training and test sets

We can subdivide the dataset into two parts for computing the rules (train set) and assessing them (test set). We obtain thus an honest estimation of the interestingness measure associated to the rule.

We click on the PARAMETERS menu. We set the LEARNING RATIO to 0.66 i.e. 660 observations on 1000 are used in the training phase; 340 in the testing phase.

0.33
1

We validate and we click on the VIEW menu.

File Diagram Component Window Help         Image: State of the st	💇 TANAGRA 1.4.30 - [A pric	ori MR 1]				
Default üle         Default üle         Defaustet (credit_assoc.xls)         Befine status 1         Bata visualization         Statistics         Nonparametric statistics         Ins	💇 File Diagram Compo	onent Window	Help		- 8	×
Default title         Dataset (credit_assoc.xls)         Define status 1         Bamples size         Training 660         Test         Test         340         ITEMS         Counting items         All items         660         Counting items         All items         66         Fittered items         Counting items         Card(itemset) = 2         66         card(itemset) = 3         66         Components         Number of rules         7         Image: Components         Nata visualization         Feature construction         Feature selection         Regression         PLS         Clustering         Spv learning assessment         Scoring         Association         Stappiori PT         Spv Assoc Rule         Stappiori MR         Stappiori MR	🗅 📾 🖪 🛛 👪					
Sample characteristics         Sample size         Training 660         Test         340         ITEMS         Transactions         660         Counting items         Al items         66         Counting itemsets         card(itemset) = 2         66         card(itemset) = 3         66         card(itemset) = 4         84         Rules         Number of rules         7         Construction         Feature construction         Feature selection         Regression         PLS         Clustering         Spv learning         Spv learning assessment         Scoring         Association         Stassoc Outlier         Spv Assoc Rule	Default title					
Components         Data visualization       Statistics       Nonparametric statistics       Instance selection         Feature construction       Feature selection       Regression       Factorial analysis         PLS       Clustering       Spv learning       Meta-spv learning         Spv learning assessment       Scoring       Association         St A priori       Spv Assoc Rule       Spv Assoc Tree	Defar tue Dataset (credit_ass	oc.xls) 1	Samples siz Training 66 Test 34 ITEMS Transaction Counting All items Filtered item card(itemse card(itemse card(itemse Ruk	characteristics		
Components         Data visualization       Statistics       Nonparametric statistics       Instance selection         Feature construction       Feature selection       Regression       Factorial analysis         PLS       Clustering       Spv learning       Meta-spv learning         Spv learning assessment       Scoring       Association       Image: A priori PT         A priori       A priori PT       Spv Assoc Rule       Image: A priori MR       Spv Assoc Tree			<b>۲</b>		4	-
Data visualization       Statistics       Nonparametric statistics       Instance selection         Feature construction       Feature selection       Regression       Factorial analysis         PLS       Clustering       Spv learning       Meta-spv learning         Spv learning assessment       Scoring       Association         Image: A priori       A priori       Spv Assoc Rule         A priori       Spv Assoc Outlier       Spv Assoc Tree			Cor	nponents		
Feature construction     Feature selection     Regression     Factorial analysis       PLS     Clustering     Spv learning     Meta-spv learning       Spv learning assessment     Scoring     Association     Meta-spv learning       St A priori     St A priori PT     Spv Assoc Rule       St A priori MR     Spv Assoc Outlier     Spv Assoc Tree	Data visualization	Statist	ics	Nonparametric statistics	Instance selection	
PLS     Clustering     Spv learning     Meta-spv learning       Spv learning assessment     Scoring     Association     Meta-spv learning       Image: A priori     Image: A priori PT     Image: Spv Assoc Rule     Image: A priori MR     Image: Assoc Outlier       Image: A priori MR     Image: Assoc Outlier     Image: Assoc Tree     Image: Assoc Outlier     Image: Assoc Outlier	Feature construction	Feature construction Feature set		Regression	Factorial analysis	
Spv learning assessment     Scoring     Association       Image: A priori     Image: A priori PT     Image: Spv Assoc Rule       Image: A priori MR     Image: Assoc Outlier     Image: Assoc Tree	PLS	Cluster	ing	Spv learning	Meta-spv learning	
Image: A priori	Spv learning assessment	Scori	ng	Association		
	St A priori St A p A priori MR St Ass	riori PT 🛛 🕄 oc Outlier '⊟	Spv Assoc Ru Spv Assoc Tr	ıle ee		

Tanagra extracts 7 rules. In the right part of the table enumerating the rules, after the TEST column (in red), we have the interestingness measures computed on the test set. The comparison of the values obtained on the train and the test set allows to assess the stability of the rules.

# 5 Conclusion

The A PRIORI MR component of Tanagra extracts rules from data using the A PRIORI algorithm. It differentiates oneself from other by offering additional tools for exploring and assessing the mined rules: original measures based on the "test value" principle allow to evaluate differently the rules; the ability to copy the results into a spreadsheet allows a more detailed exploration of the rule base; by subdividing the dataset into train and test sets, we obtain a more reliable values of the interestingness measures of rules.