Subject

Computing association rule with TANAGRA, ORANGE and WEKA.

We must respect the following steps if we want to compute association rules from a dataset:

- Import the dataset;
- Select the descriptors;
- Set the parameters of the association rule algorithm i.e. the minimal support and the minimal confidence;
- Execute the algorithm and visualize the rules.

Our three packages use attribute-based dataset. Each attribute-value couple becomes an item which be used for generating rules.

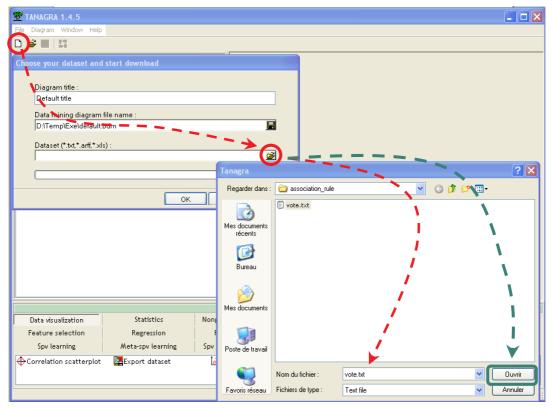
Dataset

We use the VOTE.TXT dataset from the UCI IRVINE repository.

Association rules with TANAGRA

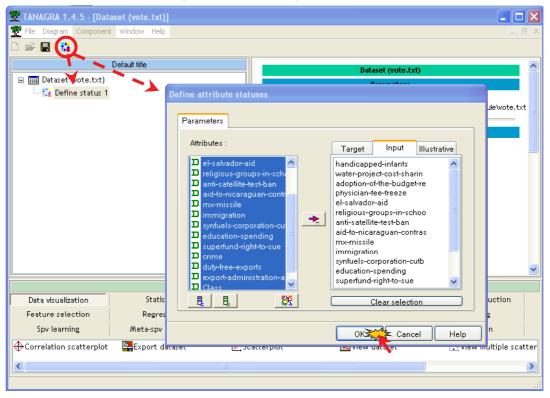
Import the dataset

First, we must create a new diagram and import the dataset with the FILE/NEW menu. We select the VOTE.TXT dataset.



Defining the attributes for the analysis

We add a DEFINE STATUS component in the diagram; we set all attributes as INPUT.



A PRIORI algorithm

There are various algorithms in TANAGRA; some of them come from external libraries. In this tutorial, we use the standard A PRIORI algorithm.

💯 TANAGRA 1.4.5 - [Dat	aset (vote.txt)]				
n File Diagram Component	t Window Help				_ 8 ×
🗅 📽 🖪 🗱					
	Default title			Dataset (vote.txt)	<u> </u>
🖃 🏢 Dataset (vote.txt)				Parameters	
🖻 🎎 Define status 1		Databa	se :		
🔤 🏦 🗛 priori 1		D:\Data	Mining\Databa	ases_for_mining\compariso	on_TOW\associat —
				D	
				Results	
		Dow	nload ir	nformation	
		Datasr	urce proces	sing	
			tation time 16	-	
1			ad mamory - 29		~
				IIII	
		Components			
Data visualization	Statistics	Nonparametric st	atistics	Instance selection	
Feature construction	Feature selection	Regression		Factorial analysis	
PLS	Clustering	Spv learning	3	Meta-spv learning	
Spv learning assessment	Scoring	Association			
🔛 🕮 A priori 🛛 🕄 🕄 A pl	riori MR 💦 🕄 A priori PT	😫 Spv Assoc Ri	ile 👍 Spv A	Assoc Tree	
Produce accoriation rule	, discrete input attribute.				
	, also de input attribute.				

Then we click on the PARAMETERS contextual menu of the component.

Association rule param	eter
Parameters	
Support :	0.50
Confidence :	0.75
Max card itemsets :	4
Lift :	1
ОК	Cancel Help

The minimal support is set to 0.5; the minimal confidence to 0.75; we use only frequent itemsets of cardinal lower or equal to 4; the rules with a LIFT lower than 1 are removed.

Computing the rules

We select the contextual VIEW menu in order to see the rules. We obtain 14 rules.

			Results			
TEMS						
ransactions	435					
Counting iter	ms					
II items	50					
filtered items	12					
Counting item	sets					
ard(itemset) = :	2 4					
ard(itemset) = (3 1					
Rules						
lumber of rules						
	5 14					
ULES	5 14					
	5 14		Number of rules : 14			
	5 14		Number of rules : 14 Consequent	Lift	Support	C
RULES				Lift 1.670	Support 0.503	
RULES	e-freeze=		Consequent			
RULES	e-freeze= prat" - "ad	n" option-of-the-budget-re=y"	Consequent "Class=democrat" - "adoption-of-the-budget-re=y"	1.670 1.670	0.503	
Antecedent Physician-fe Class=demod Class=demod	e-freeze= crat" - "ad crat"	n" option-of-the-budget-re=y"	Consequent "Class=democrat" - "adoption-of-the-budget-re=y" "physician-fee-freeze=n"	1.670 1.670	0.503	
Antecedent Physician-fe Class=demod Class=demod	e-freeze= crat" - "ad crat" the-budge	n" option-of-the-budget-re=y"	Consequent "Class=democrat" - "adoption-of-the-budget-re=y" "physician-fee-freeze=n" "adoption-of-the-budget-re=y" - "physician-fee-freeze=n"	1.670 1.670 1.629	0.503 0.503 0.503	
Antecedent Physician-fe Class=democ Class=democ Class=democ	e-freeze= prat" - "ad prat" the-budge prat"	n" option-of-the-budget-re=y" ət-re=y" - "physician-fee-freeze=n"	Consequent "Class=democrat" - "adoption-of-the-budget-re=y" "physician-fee-freeze=n" "adoption-of-the-budget-re=y" - "physician-fee-freeze=n" "Class=democrat"	1.670 1.670 1.629 1.629	0.503 0.503 0.503 0.503	
Antecedent Physician-fe Class=democ Class=democ Class=democ Class=democ	e-freeze= crat" - "ad crat" the-budge crat" e-freeze=	n" option-of-the-budget-re=y" et-re=y" - "physician-fee-freeze=n" n"	Consequent "Class=democrat" - "adoption-of-the-budget-re=y" "physician-fee-freeze=n" "adoption-of-the-budget-re=y" - "physician-fee-freeze=n" "Class=democrat" "physician-fee-freeze=n"	1.670 1.670 1.629 1.629 1.616	0.503 0.503 0.503 0.503 0.503 0.563	

N°	Antecedent	Consequent	Lift	Support	Confidenc
1	"physician-fee-freeze=n"	"Class=democrat" - "adoption-of-the-budget-re=y"	1.670	0.503	0.88
2	"Class=democrat" - "adoption-of-the-budget-re=y"	"physician-fee-freeze=n"	1.670	0.503	0.94
3	"Class=democrat"	"adoption-of-the-budget-re=y" - "physician-fee-freeze=n"	1.629	0.503	0.820
4	"adoption-of-the-budget-re=y" - "physician-fee-freeze=n"	"Class=democrat"	1.629	0.503	1.000
5	"Class=democrat"	"physician-fee-freeze=n"	1.616	0.563	0.918
6	"physician-fee-freeze=n"	"Class=democrat"	1.616	0.563	0.993
7	"adoption-of-the-budget-re=y"	"Class=democrat" - "physician-fee-freeze=n"	1.537	0.503	0.860
8	"Class=democrat" - "physician-fee-freeze=n"	"adoption-of-the-budget-re=y"	1.537	0.503	0.894
9	"adoption-of-the-budget-re=y"	"physician-fee-freeze=n"	1.524	0.503	0.866
10	"physician-fee-freeze=n"	"adoption-of-the-budget-re=y"	1.524	0.503	0.887
11	"Class=democrat"	"adoption-of-the-budget-re=y"	1.488	0.531	0.865
12	"adoption-of-the-budget-re=y"	"Class=democrat"	1.488	0.531	0.913
13	"Class=democrat"	"aid-to-nicaraguan-contras=y"	1.468	0.501	0.816
14	"aid-to-nicaraguan-contras=y"	"Class=democrat"	1.468	0.501	0.901

Association rules with WEKA

A dialog box appears when we execute WEKA; we choose the **KNOWLEDGE FLOW** paradigm. We have used the **3.5.1** version.

💌 Weka GUI Cho 🔳 🗖 🔀	불 Weka KnowledgeFlo	v Environment				
	DataSources DataSi	nks Filters Classifiers Clusterers A	ssociations Evaluation Visual	ization		, R
Waikato Environment for	DataSources					
Knowledge Analysis		bod Gdd ⁽¹⁰⁾	yes, 8 no, 3 CSV		100010 010011 SERIAL	
Version 3.5.1	Arff	C45	CSV	Database	Serialized	
(c) 1999 - 2005	Loader	Loader	Loader	Loader	InstancesLoader	
University of Waikato	<u><</u>					>
New Zealand	Knowledge Flow Layout					
						<u>^</u>
GUI						
Simple CLI <u>Explorer</u>	<					>
Experimenter KnowledgeFlow	Status					
ArffViewer Log	Welcome to the Weka Know	ledge Flow				Log

Import the dataset

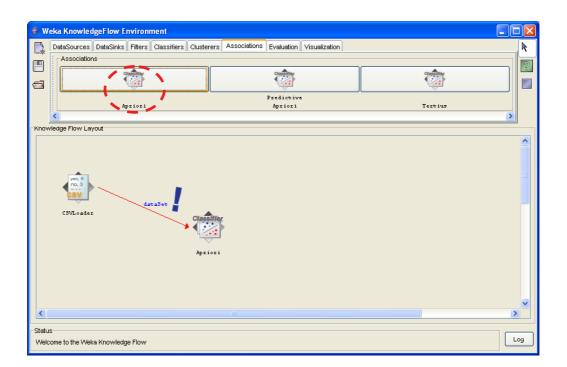
The CSV LOADER enables to handle text file format. We select the VOTE.TXT dataset with the CONFIGURE contextual menu.

👙 Weka KnowledgeFlow Environment							
DataSources DataSinks Filters Classifiers Clusterers Asso	ociations Evaluation Visualization						
DataSources							
Arff C45 Loader Loader	CSV Database Serialized Loader InstancesLoader						
	<u>></u>						
-Knowledge Flow Layout							
	About						
Yes. 8	Reads a source that is in comma separated or tab separated format.						
CSV Edit	Rechercher dans : 📄 association_rule 🔮 🦻 🖽 📰						
CSULoade Configure	vote totm						
Connections	Mes El vote.txt						
dataSet instance	documents récente						
Actions Start loading							
	Bureau						
	Mes dotuments						
	Poste de travail						
<	Poste de travail						
	Nom de fichier : vote txt						
∽Status Welcome to the Weka Knowledge Flow	Favoris réseau Fichiers du type : Tous les fichiers V Annuler						

A PRIORI algorithm

The default selections are all instances and all attributes, so we must add only the A PRIORI component from the ASSOCIATION tab in the diagram.

We use the DATASET connection type.

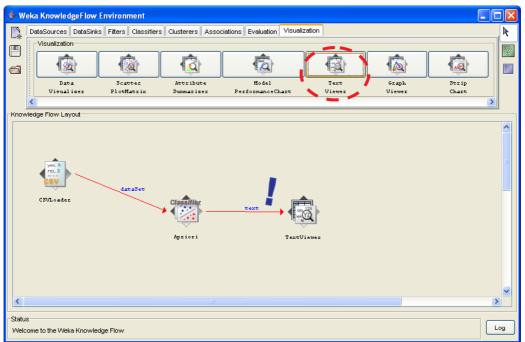


The CONFIGURE contextual menu allows to set the parameters values.

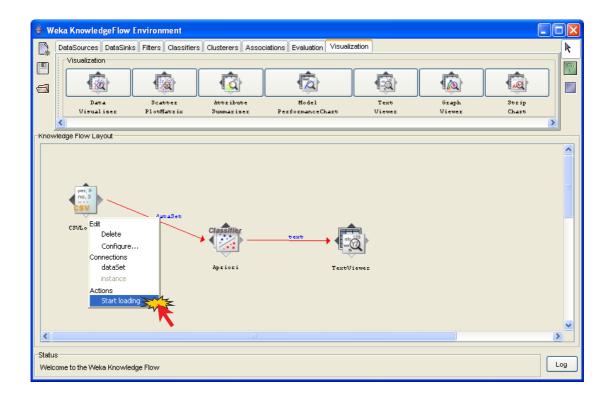
4	
About	
Finds association ru	les. More
car	False
classindex	-1
delta	0.05
lowerBoundMinSupport	0.5
metricType	Confidence
minMetric	0.75
numRules	100
removeAllMissingCols	False
significanceLevel	-1.0
upperBoundMinSupport	1.0

LOWERBOUNDMINSUPPORT set the minimal support of rules; MINMETRIC is the minimal confidence, if we set CONFIDENCE as METRIC TYPE; NUMRULES set the maximal number of rules that we can generate.

In order to visualize the rules, we add the TEXT VIEWER component in the diagram; we use the TEXT connection.



To execute the computation, we click on the START LOADING of the first component (CSV LOADER).



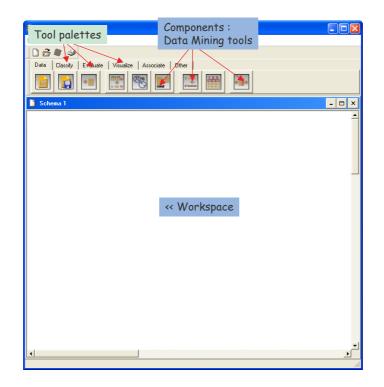
We can see the rule by clicking the SHOW RESULTS menu of the TEXT VIEWER component.

esult list	ir Text		
sour not			
:58:34 - Model: Aprior			
	Minimum support: 0.5		
	Minimum metric <confidence>: 0.75</confidence>		
	Number of cycles performed: 10		
	Number of cycles performed. To		
	Generated sets of large itemsets:		
	Size of set of large itemsets L(1): 12		
	Size of set of large itemsets L(2): 4		
	Size of set of large itemsets L(3): 1		
	Best rules found:		
	 adoption-of-the-budget-re=v physician-fee-freeze=n 219 ==> Class=democrat 219 	conf:(1)	
		conr:(1)	
	2. physician-fee-freeze=n 247 ==> Class=democrat 245 conf:(0.99)	a 10 051	
		conf:(0.95)	
	 Class=democrat 267 ==> physician-fee-freeze=n 245 conf: (0.92) 		
	5. adoption-of-the-budget-re=y 253 ==> Class=democrat 231 conf: (0.91)		
	 aid-to-nicaraguan-contras=y 242 ==> Class=democrat 218 conf: (0.9) 		
		conf:(0.89)	
	 physician-fee-freeze=n 247 ==> adoption-of-the-budget-re=y 219 conf: (0.89) 		
	9. physician-fee-freeze=n 247 ==> adoption-of-the-budget-re=y Class=democrat 219	conf:(0.89)	
	10. adoption-of-the-budget-re=y 253 ==> physician-fee-freeze=n 219 conf:(0.87)		
	11. adoption-of-the-budget-re=y 253 ==> physician-fee-freeze=n Class=democrat 219	conf:(0.87)	
	12. Class=democrat 267 ==> adoption-of-the-budget-re=y 231 conf:(0.87)		
	13. Class=democrat 267 ==> adoption-of-the-budget-re=y physician-fee-freeze=n 219	conf:(0.82)	
	14. Class=democrat 267 ==> aid-to-nicaraguan-contras=y 218 conf:(0.82)		

We obtain the same 14 rules as TANAGRA.

Association rules with ORANGE

When we execute ORANGE, we have the following interface.



Importing the dataset

ORANGE can handle text file format (tabulation separator). When we select the tool, a new component is inserted in the diagram. We can select the file with the OPEN contextual menu.

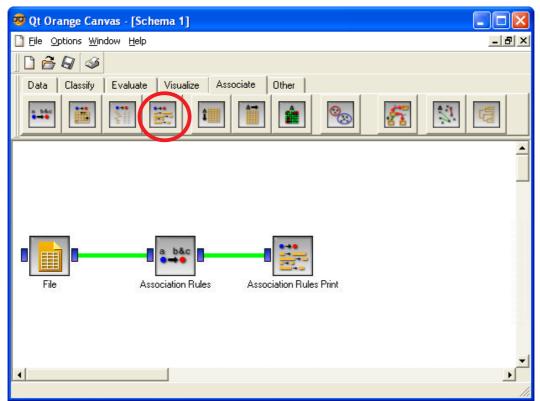
🤓 Qt Orange Canvas	
Elle Options Window Help	
D 2 8 3	
Classify Evaluate Visualize Associate Other	
📑 Schema 1	- 🗆 ×
	-
Classificat Reload	I Reset domain at next reload

A PRIORI algorithm

In order to compute the rules, we add ASSOCIATION RULES component. All examples and attributes are used. We click on the OPEN menu for parameters setting.

Ele Options Window Help Data Classify Evaluate Visualize Associate Other Image: Classify Evaluate Visualize Associate Image: Classify Evaluate Visualize Associate Image: Classify Evaluate Visualize Association Image: Classify Evaluate Visualize Association Image: Classify Evaluate Visualize Association Image: Classify Evaluate Imag	🤓 Qt Orange Canvas - [Schema 1]	
Data Classify Evaluate Visualize Associate Other Image: Im	📔 Eile Options Window Help	_ B ×
Image:		
File Association Ru Open Minimal support [%] Induce classification rules Minimal confidence [%] Induce classification rules Minimal support [%] Induce classification rules Minimal confidence [%] Induce classification rules Induce classification rules Minimal confidence [%] Induce classification rules	Data Classify Evaluate Visualize Associate Other	
Build algorithm Use algorithm for sparse data Induce classification rules Minimal support [%] Sename F2 Remove Del		
	File Association Ru	Build algorithm Use algorithm for sparse data Induce classification rules Minimal support [%] Minimal confidence [%] 75 Maximal number of rules 10000
		>

The rules are automatically computed when we connect the FILE component to ASSOCIATION RULE. We add the ASSOCIATION RULE PRINT component for rules visualization.



We click on the OPEN menu in order to view the rules.

🔣 Qt Association Ru	les Pri	nt			
Measures Support Confidence Lift Leverage Strength Coverage	supp 0.531 0.531 0.563 0.563	conf 0.865 0.913 0.918 0.992	lift 1.488 1.488 1.616 1.616	rule Class=democrat -> adoption-of-the-budget-re=y adoption-of-the-budget-re=y -> Class=democrat Class=democrat -> physician-fee-freeze=n physician-fee-freeze=n -> Class=democrat	

We obtain only 4 rules in ORANGE; we had found 14 with TANAGRA and WEKA. ORANGE seems to have a preference for "shorter" rules. I did not find the reasons of this difference.

Conclusion

Our three packages are very simple to use for association rules induction.

These packages are largely sufficient for the majority of the analyses. The situation is a little more difficult if we wish to treat big databases with thousands of items. The number of generated rules can become very high and, the performances, the possibility of even carrying out calculations, very strongly depend on the RAM memory available of the machine used.