1 Introduction

Extraction des règles d'association prédictives dans Tanagra.

Les algorithmes d'extraction des règles d'association ont été initialement mis au point pour découvrir des liens logiques entre des variables ayant le même statut. Les règles d'association prédictives en revanche cherchent à produire les combinaisons d'items qui caractérisent au mieux une variable qui joue un rôle à part, on cherche à prédire ses valeurs.

Fondamentalement, l'algorithme est peu modifié. L'exploration est simplement restreinte aux itemsets qui comportent la variable à prédire. Le temps de calcul est d'autant réduit. Deux composants de Tanagra sont dédiés à cette tâche, il s'agit de SPV ASSOC RULE et SPV ASSOC TREE. Ils sont accessibles dans l'onglet ASSOCIATION.

Par rapport aux approches classiques, les composants de Tanagra introduisent une spécificité supplémentaire : nous avons la possibilité de préciser la classe (couple « variable à prédire = valeur ») que l'on souhaite prédire. L'intérêt est de pouvoir ainsi paramétrer finement l'algorithme de recherche, en relation directe avec les caractéristiques des données. Cela s'avère décisif par exemple lorsque les prévalences des modalités de la variable à prédire sont très différentes.

Nous avions déjà présentés le composant SPV ASSOC TREE par ailleurs (<u>http://tutoriels-data-mining.blogspot.com/2008/04/rgles-dassociation-supervises.html</u>). Mais c'était dans le contexte de la caractérisation multivariée de groupes d'individus. Nous l'opposions alors au composant GROUP CHARACTERIZATION. Dans ce didacticiel, nous comparerons le comportement des composants SPV ASSOC TREE et SPV ASSOC RULE sur un problème de prédiction. Nous mettrons en avant leurs points communs, les problèmes qu'ils savent traiter ; et leurs différences, SPV ASSOC RULE, en plus de proposer des mesures d'intérêt des règles originales, a la capacité de simplifier la base de règles.

2 Données

Nous utilisons une version modifiée du fichier GERMAN CREDIT¹. Il décrit les caractéristiques de demandeurs de crédit. Nous avons discrétisé les variables quantitatives. Le fichier est accessible en ligne (<u>http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/credit_assoc.xls</u>).

La variable CLASS joue un rôle particulier dans ce didacticiel. Nous cherchons à mettre à jour les caractéristiques des « bons » clients (CLASS = GOOD). Nous avons donc un double paramétrage à faire avant de pouvoir exécuter les calculs : indiquer que CLASS est la variable TARGET ; parmi les valeurs de CLASS, choisir la modalité GOOD.

¹ <u>http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)</u>

3 Création d'un diagramme

Importer les données. Première étape toujours dans tout projet de Data Mining, nous devons définir un nouveau projet et importer les données. Après avoir lancé Tanagra, nous actionnons le menu FILE / NEW. Tanagra sait lire directement les fichiers au format Excel (XLS)². Nous sélectionnons le fichier CREDIT_ASSOC.XLS.

TANAGRA 1.4.30						
File Diagram Window Help						
New						
😅 Open Ig dia Choose your dataset and start do	ownload					
Save						
Diagram title :				_		
Close Default title						
Exit Data mining diagram file	name :			_		
D:\DataMining\Databases	s_for_mining\da	taset_for_soft_c	lev_and_tompa			
Dataset (*.txt,*.arff,*.xls) :			<u> </u>			
D:\DataMining\Databases	s_for_mining\da	taset_for_soft_c	lev_and_compa	2		
		<i>e</i> 1 51 -				X
		Mer Tanagra				
		Regarder dans :	assoc mr and spv	v assoc	- G 🕫 🖻 🖽	•
		9	Nom Date	e de m Tyre	Taille	
		Emplacements	Credit_assoc.xis			
	Compo	recents			Type : Feuille de calcul Mi Taille : 269 Ko	crosoft Excel
Data visualization Statistics	Nonparamet				Date de modification : 04/	02/2009 19:05
Feature construction Feature selection	Regre	Bureau				
PLS Clustering	Spv le	in.				
Spy learning assessment Scoring	Assoc	Maison				
Correlation scatterplot	🔣 View o	maison				
Export dataset	l <u>t</u> view r					
		Ordinateur		V		
			Nom du fichier :	credit_assoc.xls	•	Ouvrir
~		2	Types de fichiers :	Excel File (97 & 20	00) 💌	Annuler

Tanagra nous indique que 17 variables et 1000 observations ont été chargées.

Définir le rôle des attributs. Nous insérons le composant DEFINE STATUS via le raccourci dans la barre d'outils pour indiquer le rôle des variables : nous plaçons CLASS en TARGET, toutes les autres en INPUT.

² Sur les différentes manières d'importer des données au format XLS (Excel), voir <u>http://tutoriels-data-</u> <u>mining.blogspot.com/2008/03/importation-fichier-xls-excel-mode.html</u> et <u>http://tutoriels-data-</u> <u>mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html</u>



Nous validons et nous cliquons sur VIEW. Tanagra affiche un récapitulatif.

TANAGRA 1.4.30 - [Define status 1]							
💇 File Diagram Component Wind	ow Help						_ 8 ×
🗅 📽 🖪 🎇							
Default title	A+	tribute	Target	Input	Illustrativa		A
Dataset (credit assoc.xls)			Target	mput	musciacive		
Define status 1	checking	_status	-	yes			
Parameters	credit h	istory		ves			
Evente	Durpose			vec			
View	amount		_	ves			
View				yes			
	savings_	scatus		yes			
	employm	ent	•	yes	•		
	personal	_status	•	yes	•		
	other_pa	arties	-	yes	•		=
	property	_magnitude	•	yes	•		-
	other_pa	ayment_plans	-	yes	-		
	housing		-	yes	•		
	existing	credits	-	yes	•		
	job		-	yes	•		
	own_tele	phone	-	yes	•		
	foreign_	worker	-	yes	-		
	class		yes	-	-		-
		Con	nnonen	to			
Data visualization St	atistics	Nonparam	etric st	tatistic	s In	stance selection	Feature construction
Feature selection Re	ression	Factor	rial ana	ysis		PLS	Clustering
Spv learning Meta-	pv learning	Spv learni	ing asse	ssmen	t	Scoring	Association
Correlation scatterplot	erplot with labe	ι					
Export dataset 🔣 View	dataset						
Scatterplot	multiple scatter	plot					

4 Le composant SPV ASSOC TREE

Le composant SPV ASSOC TREE extrait des règles d'association. L'implémentation utilise en interne un arbre de recherche, d'où son nom, mais au final nous obtenons bien des règles.

4.1 Choix de la modalité à prédire

Nous l'insérons dans le diagramme. Nous actionnons le menu PARAMETERS pour introduire un premier paramétrage obligatoire : nous devons indiquer au logiciel la modalité de la variable à prédire que l'on souhaite expliquer. Dans l'option CLASS VALUE, nous désignons la modalité GOOD.

TANAGRA 1.4.30 - [Define	e status 1]		- D X
💇 File Diagram Compo	onent Window He	łp	_ 8 ×
🗅 📽 🔲 🎎			
Default title			
🖃 🏢 Dataset (credit_ass	oc.xls)	Association tree spv	
🖻 🚰 Define status 1		d Parameters Sorting	
Spv Assoc Ti	ree 1	c	
Para	meters	P	
Exec	ute	a Support: j0.35	
View		s Confidence : 0.75	
		e Max card itemsets : 4	
		P	-
	,		
Data visualization	Statistics	Learning set ratio : 1	
Feature construction	Feature select	di Class value : good	
PLS	Clustering	good	
Spv learning assessment	Scoring		
A priori 8 A p	riori PT 🚦 Spv	A OK Cancel He	lp
A priori MR	oc Outlier 🛛 🏣 Spv	4	

Nous validons et nous lançons les calculs.

💇 TANAGRA 1.4.30 - [Spv /	Assoc Tree 1]								x		
💇 File Diagram Comp	onent Window	Help						-	ъ×		
🗅 📽 🔛 🎎											
Default title	Default title Results										
Dataset (credit_as	Dataset (credit_assoc.xls)										
E Spv Assoc T	"⊨ Spv Assoc Tree 1 "class" is "good" IF										
		N°	N° Antecedent Length Support Confidence Lift								
		1 foreign_worker=yes - checking_status=no checking 2 0.335 (0.00) 0.879 (0.00) 1.256 (0.00)									
		2 other_parties=none - checking_status=no checking 2 0.331 (0.00) 0.887 (0.00) 1.268 (0.00)									
		3 checking_status=no checking 1 0.348 (0.00) 0.883 (0.00) 1.262 (0.00)									
									-		
			Components								
Data visualization	Statist	ics	Nonparametric statistics	Instance	selecti	on Fe	ature constr	uction			
Feature selection	Regress	ion	Factorial analysis	F	LS		Clustering	g			
Spv learning	Meta-spv le	earning	Spv learning assessment	Sci	oring		Associatio	n			
🕄 A priori	soc Outlier										
🕄 A priori MR 🛛 🕄 Sp	v Assoc Rule										
🕄 A priori PT 🛛 🛨 Sp	Assoc Tree										
								на нама н	4.4.2000		

Le composant produit 3 règles. Elles sont énumérées dans la partie basse de la fenêtre d'affichage. Le support, la confiance et le lift sont fournis. Lorsque les chiffres sont entre parenthèses, cela veut dire que l'indicateur a été calculé sur un échantillon test, n'ayant pas participé à l'élaboration des règles. Ici, nous avons systématiquement la valeur zéro. Toutes les observations sont utilisées pour l'élaboration des règles.

4.2 Partition « apprentissage – test » des données

Le danger lors de l'induction est de produire des solutions trop spécifiques aux données traitées, qui ne correspondent à aucune réalité dans la population. Pour éviter cet écueil, il est généralement conseillé de partitionner les observations en deux parties : la première, dite échantillon d'apprentissage, sert à élaborer les règles de prédiction ; la seconde, dite échantillon de test, sert uniquement à leur évaluation. Les indicateurs calculés dans ce second contexte sont nettement plus crédibles, traduisant les performances réelles de la règle. A mon sens, il faut surtout s'inquiéter lorsque le même indicateur calculé sur les deux échantillons prend des valeurs anormalement différentes.

Pour subdiviser les données, nous revenons sur la boîte de paramétrage via le menu PARAMETERS. Nous modifions le paramètre LEARNING SET RATIO, nous le fixons à 0.6 (*attention, le point décimal dépend de la configuration de votre système*) c.-à-d. nous utilisons 60% des données pour la construction des règles, 40% pour leur évaluation.

TANAGRA 1.4.30 - [Define status 1]	
💇 File Diagram Component Window Help	_ <i>B</i> ×
Association tree spy	
Default title Attribute Default title Attribute Dataset (credit_assoc.xls) checking_status duration duration Credit_history parameters Parameters nurpose Execute amount View savings_status employment Lift : Ltearning set ratio :	0.33 0.75 4 1.1 0.6
Data visualization Statistics Nonparam Class value :	good 🔻 Dn
Feature selection Regression Facto	
Spv learning Meta-spv learning Spv learn	
A priori Assoc Outlier A priori MR Spv Assoc Rule A priori PT Spv Assoc Tree	OK Cancel Help
	anna a a a a a a a a a a a a a a a a a

Nous validons puis nous cliquons sur VIEW.

	ssoc Tree 1]								
File Diagram Compo	nent Window Help						- 1		
🖻 🖬 😫									
Default title ⊡∽∰ Dataset (credit_assi	oc.xls) "class" is "g	good" IF							
🗄 🚰 Define status 1	N°	Antecedent		Length	Support	Confidence	Lift		
Spv Assoc Tr	ree 1 1 foreign_wor checking	1 foreign_worker=yes - other_parties=none - checking_status=no 3 0.343 (0.28) 0.900 (0.86) 1.301 (1.20)							
	2 foreign_wor	2 foreign_worker=yes - checking_status=no checking 2 0.353 (0.31) 0.887 (0.87) 1.282 (1.22)							
3 other_parties=none - checking_status=no checking 2 0.352 (0.30) 0.902 (0.86) 1.304 (1.21)									
	4 other_paym	ent_plans=none - housing=own		2	0.428 (0.47)	0.763 (0.78)	1.103 (1.09)		
	5 checking_st	tatus=no checking		1	0.363 (0.33)	0.890 (0.87)	1.286 (1.22)		
		Componen	ts			1			
Data visualization	Statistics	Nonparametric statistics	Instance selection		Feature cons	truction			
Feature selection	Regression	Regression Factorial analysis PLS Clustering							
Spv learning	Meta-spv learning	Spv learning assessment	Scoring		Associat	ion			
	And DT Courterer	Rule							

Première information importante, nous avons davantage de règles, sans que nous ayons modifié les autres paramètres. Cela veut dire simplement que certaines règles (la n°1 pour le critère de support ; la n°4 pour le critère de confiance) étaient toutes proches des seuils de coupures, elles avaient été arbitrairement éliminées précédemment.

Deuxième information importante, nous avons une meilleure visibilité sur la stabilité de la règle en comparant le même indicateur calculé sur l'échantillon d'apprentissage et de test. Nous noterons entre autres que les chiffres en test sont moins optimistes.

4.3 Produire plus de règles et les organiser

Comme nous avons pu le constater, les paramètres de filtrage jouent un rôle important lors de l'extraction. SPV ASSOC TREE propose les éléments suivants :

- SUPPORT définit le support minimum des règles extraites ;
- CONFIANCE définit la confiance minimum des règles extraites ;
- MAX CARD ITEMSETS restreint la longueur maximale de la règle ;
- LIFT indique le LIFT minimum des règles.

Les trois premiers paramètres restreignent l'espace de recherche, ils pèsent sur le temps de calcul. Le dernier agit a posteriori, uniquement pour limiter le nombre de règles à afficher.

Un fois les règles produites, il faut pouvoir les organiser de manière à ce que les plus intéressantes apparaissent en premier. Tanagra sait classer les règles selon un des critères numériques ci-dessus.

Dans ce qui suit, nous allons diminuer le support minimum pour obtenir plus de règles (SUPPORT = 0.10). Puis nous les classerons selon un lift décroissant.

Nous cliquons sur le menu PARAMETERS.

TANAGRA 1.4.30 - [Define status 1] File Diagram Component Window Help Default tile Attribute Checking_status duration Execute View employment personal_status other_parties property_magnitus other_parties property_magnitus other_parties Sort rules by no sort Sort rules by no sort construction staring Meta-spv learning Sta priori Sta priori Sta priori Sta size Coutlier Exprise multielingth rule length rule length rule length rule length rule length			Associat	tion tree spv			1
View employment personal_status other_parties other_parties Association tree spv property_magnitude other_payment_plan Data visualization Statistics Spv learning Meta-spv learning Spv learning Meta-spv learning Spv learning Meta-spv Assoc Rule A priori Spv Assoc Outlier Spv Assoc Outlier Spv Assoc Tree Image: A priori MR Spv Assoc Tree	TANAGRA 1.4.30 - [Define status 1] Time Diagram Component Wir Default title Default title Default title Define status 1 Spv Assoc Tree 1 Paran Execu	ndow Help Attribute checking_status duration neters	Para	ameters Sorting Support: Confidence : Max card itemsets : Lift : Learning set ratio : Class value :	, 0.10 0.75 4 2 1.1 0.6		
Data visualization Statistics No Feature selection Regression Spv learning Meta-spv learning Sp Image: Spr Assoc Rule Image: Regression Image: Spr Assoc Rule Image: Regression Image: Spr Assoc Rule Image: Regression Image: Regression Image: Regression Image: Spr Assoc Rule Image: Regression Image: Regression Image: Regresion Image: Regression	View	employment personal_status other_parties property_magnitude other_payment_plan	Associat	tion tree spv ameters Sorting	OK Cancel	Help	
	Data visualization Feature selection R Spv learning Meta A priori A priori PT A priori MR Assoc Outlier	Statistics No legression Sp a-spv learning Sp Spv Assoc Rule E Spv Assoc Tree		Sort rules by Tule length Tule confidence Tule lift			construction ustering sociation

Nous cliquons sur VIEW pour obtenir les résultats.

TANAGRA 1.4.30 - [Spv Assoc Tree 1]											
💇 File Diagram Compon	💇 File Diagram Component Window Help										
D 📽 🖬 🗱 🔚 🙀											
Default title			Results						*		
🖃 🏢 Dataset (credit_asso	c.xls)	Dulas									
🖻 🚰 Define status 1	Rules	Kules									
□ E Spy Assoc Tree 1											
	Class 15 g000 Ir										
	N°	N° Antecedent Length Support Confidence Lift									
other_payment_plans=none - checking_status=no checking - credit_history=critical/other 3 0.135 (0.13) 0.988 (0.96) 1.428 (1.3) 1 existing 3 0.135 (0.13) 0.988 (0.96) 1.428 (1.3)								1.428 (1.35)			
	2 housing=ow	n - checking_status=no checking	 credit_history=critical/other 	existing	3	0.127 (0.11)	0.974 (0.92)	1.409 (1.29)			
	3 other_parties=none - checking_status=no checking - purpose=radio/tv 3 0.118 (0.11) 0.973 (0.91) 1.406 (1.28										
	4 other_paym	ent_plans=none - checking_statu	us=no checking - purpose=radio/	tv	3	0.107 (0.10)	0.970 (0.93)	1.402 (1.31			
	5 existing_cre	edits=one - checking_status=no c	checking - own_telephone=yes		3	0.107 (0.07)	0.970 (0.86)	1.402 (1.20			
	6 checking_st	tatus=no checking - purpose=radi	io/tv		2	0.123 (0.12)	0.961 (0.92)	1.389 (1.29	-		
		1	Components								
Data visualization	Statistics	Nonparametric statistics	Instance selection	Feature construc	tion	Featur	e selection				
Regression	Factorial analysis	orial analysis PLS Clustering Spv learning Meta-spv learning									
Spv learning assessment	Spv learning assessment Scoring Association										
StA priori Sta priori PT Stype Assoc Rule StA priori MR Assoc Outlier Stype Assoc Tree											
									.4.01 .430		

327 règles sont maintenant produites. Les plus intéressantes au sens du LIFT apparaissent en premier. On remarquera encore la différence entre les valeurs des indicateurs selon que nous les calculons sur l'échantillon d'apprentissage ou de test. C'est le chiffre entre parenthèses (échantillon test) qui est réellement crédible.

5 Le composant SPV ASSOC RULE

Le composant SPV ASSOC RULE produit aussi des règles prédictives. Il répond au même cahier de charges que le SPV ASSOC TREE, à la différence que : (a) l'algorithme interne est un peu différent ; (b) il propose une vaste panoplie de mesures ; (c) il intègre la possibilité de réduire la base de règle par simplification logique. Voyons comment tout cela fonctionne.

5.1 Paramétrage et lecture des résultats

Nous insérons le composant dans le diagramme. Nous actionnons le menu PARAMETERS. Nous indiquons la modalité de la variable à prédire que nous souhaitons caractériser c.-à-d. CLASS VALUE = GOOD.

TANAGRA 1.4.30 - [Spv Assoc Tree 1]									
💇 File Diagram Compone	nt Window Help		- 8 ×						
🗅 📽 🔚 🏭		Spv Assoc Rule Parameters							
Default title		Parameters							
□	.xls)								
🖃 🚰 Define status 1	Rules								
⊨ Spv Assoc Tree	1	Support : 0.33							
Spv Assoc Rule	class is good	Confidence : 0.75							
Parar	meters	Max card itemsets 4	idence Lift						
Execu	ute er_payment_plans=		B (0.96 1.428 (1.35)						
View	sing=own - checking	μπ: 1.1	4 (0.92 1.409 (
	2 credit_history=critica	Learning set ratio:	1.29)						
	3 other_parties=none - o		B (0.91 1.406 (
		Repetition : 1	*						
		Class value : good 🗸							
Data visualization	Statistics Nonpa		ction						
Feature selection	Regression Fa	VT Boundary Filtering : 2							
Spv learning	Meta-spv learning Spv le								
A priori A prior	ri PT Spv Assoc Rule								
-g- A priori MK - Assoc	Outlier E Spy Assoc Tree	OK Cancel Help							
			-44.4 -48.4 -48.4 -48.4 -48.4 -48.4 -48.4 -48.4 -48.4 -48.4 -48.4 -48.4 -49.4						

D'autres paramètres sont proposés. Ils sont en relation avec les nouvelles mesures d'évaluation de règles intégrées dans l'outil. Il y a en particulier la valeur test que nous décrivons longuement par ailleurs (http://tutoriels-data-mining.blogspot.com/2009/02/mesures-dinteret-des-regles-dans-priori.html).

Citons rapidement ces paramètres :

- REPETITION définit le nombre de réplications lors du calcul à l'aide de la procédure de Monte Carlo ;
- VT Boundary Filtering définit la valeur test seuil qui permet d'accepter une règle. Il se réfère à l'indicateur **Z (HYP)**. Il agit a posteriori pour limiter le nombre de règles à afficher.

Nous cliquons sur VIEW.

Pile Diagram Component Window Help Pile Default Me Default Me Filtered = 1 rules Image: Second Se	🃅 TANAGRA 1.4.30 - [Spv As	ssoc Rule 1]											
Image: Second status Filtered = 1 rules Image: Second status Filtered = 1 rules Image: Second status Second status Image: Second status Image: Second status Image: Second status <th>💇 File Diagram Compo</th> <th>nent Window Helj</th> <th>р</th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th></th> <th>- 8 ×</th>	💇 File Diagram Compo	nent Window Helj	р										- 8 ×
Default tile Filtered = 1 rules Spv Assoc Rule 1 Filtered = 1 rules Spv Assoc Rule 1 N define status 1 Spv Assoc Rule 1 N* Antécédent Conséquen n n[A] n[C] n[A*C] Support Confiance Lift Leverage Important All rules Rules evaluation N* Antécédent Conséquent n n[A] n[C] n[A*C] Support Confiance Lift Leverage Important All rules Rules evaluation Nonsequent n n[A] n[C] n[A*C] Support Confiance Lift Leverage Important * Checking * tatus=no class=good* 1000 344 700 348 0.3480 0.8832 1.2618 0.0722 0.4191 * Checking * tatus=no class=good* 1000 373 700 348 0.3480 0.8832 1.2618 0.0722 0.4191 * Checking * tatus=no class=good* 1000 373 700 348 0.3480 0.8832 1.2618 0.0722 0.4191 * Checking * tatus=no class=good* 1000 373 700 311 0.310 0.8874 1.2677 0.0699 0.4107 0.4197 0.4197 0.4197 <th>D 📽 🔚 👪</th> <th></th>	D 📽 🔚 👪												
N* Antécédent Conséquent n n[A] n[C] n[A*C] Support Confiance Lift Leverage Important 1 "checking." "class=good" 1000 394 700 348 0.3480 0.8832 1.2618 0.0722 0.4191 Antécédent Conséquent n n[A] n[C] n[A*C] Support Confiance Lift Leverage Important N° Antécédent Conséquent n n[A] n[C] n[A*C] Support Confiance Lift Leverage Important 1 "checking.status=no rclass=good* 1000 394 700 348 0.3480 0.8832 1.2618 0.0722 0.4191 2 "checking.status=no rclass=good* 1000 394 700 348 0.3480 0.8832 1.2618 0.0722 0.4191 2 "checking.status=no rclass=good* 1000 373 700 331 0.3870 0.8874 1.267 0.0699 0.4107 2 "checking.status=no	Default title Default title Filtered = 1 rules Spv Assoc Tree 1 Spv Assoc Rule 1												
1 *checking: *class=good 1000 394 700 348 0.3480 0.8832 1.2618 0.0722 0.4191 All rules Rules evaluation 1 *checking: *class=good 1000 394 700 348 0.3480 0.8832 1.2618 0.0722 0.4191 All rules ***********************************		N	Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importar
All rules Nules evaluation N° Antécédent conséquent n n(A) n(C) n(A ^C) Support Confiance Lift Leverage Important 1 Checking_status=no class=good 1000 394 700 348 0.3480 0.8832 1.2618 0.0722 0.4191 2 Checking_* class=good 1000 373 700 331 0.3310 0.8874 1.2677 0.6699 0.4107 2 checking_* class=good* 1000 381 700 335 0.3350 0.8793 1.2561 0.6683 0.3995 2 Checking_* class=good* 1000 381 700 335 0.3350 0.8793 1.2561 0.6683 0.3995 2 Components Components Clustering Clustering Clustering Especiation Feature construction Feature construction Feature construction Feature selection Regression Factorial analysis PLS Clustering Spv learning assessment Scoring <td< td=""><td></td><th>1</th><td>"checking_status=no checking"</td><td>"class=good"</td><td>1000</td><td>394</td><td>700</td><td>348</td><td>0.3480</td><td>0.8832</td><td>1.2618</td><td>0.0722</td><td>0.4191</td></td<>		1	"checking_status=no checking"	"class=good"	1000	394	700	348	0.3480	0.8832	1.2618	0.0722	0.4191
N* Antécédent Conséquent n n[A] n[C] n[A*C] Support Confiance Lift Leverage Important 1 "checking" "class=good" 1000 394 700 348 0.3480 0.8832 1.2618 0.0722 0.4191 1 "checking" "class=good" 1000 373 700 331 0.3310 0.8874 1.2677 0.0699 0.4107 2 "checking"- "class=good" 1000 381 700 331 0.3350 0.8793 1.2677 0.0699 0.4107 3 "checking"- "class=good" 1000 381 700 335 0.3350 0.8793 1.2671 0.0683 0.3995 * "incering"-worker=yes" "class=good" 1000 381 700 335 0.3350 0.8793 1.2561 0.0683 0.3995 * "incering"-worker=yes" "incering"-worker=yes" Instance selection Feature construction Feature selection Regression Factorial analysis PLS Cluster	All rules Rules evaluation												
1 "checking_status=no" relass=good" 1000 394 700 348 0.3480 0.8832 1.2618 0.0722 0.4191 2 "checking" "class=good" 1000 373 700 311 0.3310 0.8874 1.2677 0.0699 0.4107 2 "checking_status=no" "class=good" 1000 373 700 311 0.3310 0.8874 1.2677 0.0699 0.4107 3 checking_status=no" "class=good" 1000 381 700 335 0.3350 0.8793 1.2561 0.0683 0.3995 Components Components Pactorial analysis PLS Clustering Spv learning Spv learning assessment Scoring Association Weta-spv learning Scatterplot Scatterplot with label Wiew dataset Sc		N	• Antécédent	Conséquent	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importan
************************************		1	"checking_status=no checking"	"class=good"	1000	394	700	348	0.3480	0.8832	1.2618	0.0722	0.4191
checking_status=no *class=good*1000 381 700 335 0.3350 0.8793 1.2561 0.0683 0.3995 *foreign_worker=yes** *class=good**** 1000 381 700 335 0.3350 0.8793 1.2561 0.0683 0.3995 *foreign_worker=yes** * ************************************		2	"checking_status=no checking" - "other_parties=none"	"class=good"	1000	373	700	331	0.3310	0.8874	1.2677	0.0699	0.4107
Image: Components 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 Correlation scatterplot Export dataset Scatterplot Scatterplot with label View dataset Image: Correlation scatterplot Image: Correlation scatterplot scatterplot Image: Correlation scatterplot		3	"checking_status=no checking" - "foreign_worker=yes"	"class=good"	1000	381	700	335	0.3350	0.8793	1.2561	0.0683	0.3995
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 Correlation scatterplot Export dataset Scatterplot Scatterplot with label View dataset Im Im Im Im Im			III										+
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 Correlation scatterplot Export dataset Scatterplot Scatterplot with label Wiew dataset			1	Component	s								
Feature selection Regression Factorial analysis PLS Clustering Spv learning Meta-spv learning Spv learning assessment Scoring Association Correlation scatterplot Export dataset Correlation Scatterplot with label Wiew dataset	Data visualization	Statistics	Nonparame	tric statistics		Ins	tance	e select	ion	Feature	constr	uction	
Correlation scatterplot Export dataset I Correlation scatterplot Export dataset I Correlation Correlatio Correlation Correlation	Feature selection	Feature selection Regression Factorial analysis PLS Clustering											
4	Correlation scatterplot Export dataset Correlation Scatterplot Correlation												
	•												ł

Les règles sont affichés dans 2 blocs séparés : **ALL RULES** indique toutes les règles extraites, nous en obtenons 3, tout comme avec le composant SPV ASSOC TREE ; **FILTERED RULES** indique la base de règles après simplification c.-à-d. élimination des règles redondantes, dans notre cas il ne reste plus qu'une règle.

En effet, nous nous rendons compte que les règles n°2 et n°3 n'apportent pas d'informations supplémentaires, en termes logiques, par rapport à la première règle n°1. Il suffit que « CHECKING STATUS = NO » pour que la classe soit GOOD. L'adjonction de conditions (items) supplémentaires ne modifie pas la conclusion. Notons un élément important, **le module de simplification est uniquement basé sur des critères logiques. On considère que les règles ont le même poids dès lors qu'elles ont passé le filtrage numérique**.

5.2 Produire plus de règles

Comme précédemment, nous pouvons réduire nos exigences numériques pour obtenir plus de règles. L'intérêt de SPV ASSOC RULE est la possibilité de réduire après coup la base de règles en éliminant les règles redondantes à l'aide du module de simplification.

Nous actionnons le menu PARAMETERS. Nous partitionnons les données en 60% pour l'apprentissage et 40% pour le test (LEARNING SET RATIO = 0.6 – *attention toujours au point décimal*), nous réduisons le support minimal à 10% (SUPPORT = 0.1).

TANAGRA 1.4.30 - [Spv A	ssoc Rule 1]				
File Diagram Compo	onent Window Help				_ 8 ×
🗅 📽 🖬 🔛		Sp	v Assoc Rule Parameters		
Default title 	oc.xls) Filtere	d	Parameters		•
Spv Assoc Tr	ree 1 Rules ev	alı	Support :	0.1	
Par	rameters N° Antéc	éde	Confidence :	0.75	Leverage Importance
Exe Vie	w 1 checking	sta	Max card itemsets :	4	8 0.0722 0.4191
	All rule	es	Lift : Learning set ratio :		
	Kutes e v	'	Repetition :		•
			Class value :	good 🔻	
Data visualization	Statistics N Regression	long	VT Boundary Filtering :	2	truction
Spv learning	Meta-spv learning	Spv (1-	ion
Correlation scatterplot	Export dataset	1 ₁₁			lataset 🔛
•				OK Cancel Help	•

Nous validons et cliquons sur VIEW. Les indicateurs calculés sur l'échantillon test sont affichés dans la deuxième partie du tableau, tout à droite.

TANAGRA 1.4.30 - [Spv A	ssoc Rule 1]												- O X
Tile Diagram Component Window Help													
🗅 📽 🔚 🔤													
Default title			Rules										*
⊡ ⊡ Dataset (credit_assoc.xls)		Number of rules 81											
🖻 👬 Define status 1		RULES											
Spv Assoc Tr	ree 1		.5										=
English Spy Assoc Ru	ule 1	Filte	ered = 4 rule	es 🔶		_							
		N°	Antécédent	Conséquen	n	n[A]	n[C]	n[A^C]	Support	Confiance	Lift	Leverage	Importance
		"cree 1 exist "othe	dit_history=critical/other ting" - er_parties=none"	"class=good"	600	162	415	137	0.2283	0.8457	1.2227	0.0416	0.2870
		"cree 2 exist "othe	dit_history=critical/other ting" - er_payment_plans=none"	"class=good"	600	147	415	129	0.2150	0.8776	1.2687	0.0455	0.3293
		"dura 3 "crea exist	ation=lo_1_year" - dit_history=critical/other ting "	"class=good"	600	69	415	66	0.1100	0.9565	1.3829	0.0305	0.3752
		4 "che chec	cking_status=no :king"	"class=good"	600	245	415	218	0.3633	0.8898	1.2865	0.0809	0.4722
Components													
Data visualization	Statistics		Nonparametric stati	atistics Ir		Instance selection			Feature construction				
Feature selection	Regression		Factorial analysi	ysis		PLS			Clustering				
Spv learning Meta-spv learni		earning	Spv learning assessment		Scoring				Association				
Correlation scatterplot	🧖 Scatterplot	Scatterplot with				t with la	abel 🔛 View dataset				🚉 View multi		
✓ []								•					

La base complète comporte 81 règles. Nous en avons moins qu'avec SPV ASSOC TREE car VT BOUNDARY FILTERING a également agit pour en limiter le nombre : une règle est acceptée si et seulement si [Z (HYP) > VT BOUNDARY FILTERING]. Si nous avions fixé VT BOUNDARY FILTERING à 0, nous aurions obtenu 327 règles, exactement comme avec SPV ASSOC TREE.

Après simplification des règles redondantes, nous n'avons plus que 4 règles. L'interprétation des résultats est grandement simplifiée.

6 Conclusion

Dans ce didacticiel, nous avons présenté deux composant originaux de Tanagra pour l'élaboration de règles d'association prédictives. Elles se distinguent par la stratégie utilisée pour dépasser l'écueil de la profusion des règles inhérente à l'algorithme d'extraction : SPV ASSOC TREE offre la possibilité d'organiser les règles selon un critère numérique choisi par l'utilisateur ; SPV ASSOC RULE introduit un module de simplification logique de la base de règles.