Objectif

Montrer comment importer directement dans TANAGRA un fichier au format EXCEL (version 97 & 2000).

Le principal avantage de cette procédure est qu'il est possible de modifier à la volée le fichier source sans avoir à reconstruire le diagramme de traitements (!).

Fichier

Nous allons utiliser le fichier ADULT, il contient 48842 individus et 15 variables. L'objectif est de caractériser les individus ayant un revenu élevé (variable « class » : supérieur à 50K\$ ou non)

Manipuler un fichier EXCEL

Importer le fichier EXCEL

La première étape consiste à importer les données. Il faut donc construire un nouveau diagramme de traitements selon la procédure habituelle (File / New), puis choisir le fichier de données.

TANAGRA 1.1.2										
File Diagram Window Help										
Data mining o	diagram (empty)									
Choose your dataset and start download										
	Diagram title									
	Default title	• •								
	Data mining	diagram file name :								
	D:\DataMin	ing\Databases_for_mining\t	oenchmark_datasets\adult\adu	lt.tc 🔚						
	Dataset (tex	tfile)								
				2						
			OK Cancel	Help						
		Components								
Data visualization	Descriptive stats	Instance selection	Feature construction	Feature selection						
Regression	Factorial analysis	Clustering	Spv learning	Meta-spv learning						
Spv learning assessment	Association									
+Correlation scatterplot	Export dataset	🖉 Scatterplot	🔣 View dataset	👫 View multiple scat	terplot					

Dans la boîte de dialogue permettant de sélectionner le fichier, plusieurs formats sont maintenant disponibles, notamment le format EXCEL. Sélectionnez-le, puis chargez le fichier ADULT.XLS.



Plusieurs précautions sont nécessaires pour importer correctement un fichier au format EXCEL :

(1) les seuls formats acceptés correspondent aux versions 97 & 2000 de EXCEL ;

(2) si le classeur contient plusieurs feuilles, les données doivent être situées dans la première feuille ;

(3) les données doivent être alignées en haut et à gauche dans la feuille de calcul càd elles doivent commencer aux coordonnées A1 ;

(4) la première ligne des données correspond au nom des variables ;

(5) il n'y a pas d'identifiant des lignes ;

(6) la plage de données ne doit pas contenir des cellules / lignes / colonnes vides.

Il est à noter que le nombre maximum de variables et d'individus que l'on peut traiter est limité par les capacités d'EXCEL et non de TANAGRA : 65534 observations et 256 variables.

Voici les résultats de l'importation, il est relativement rapide (7 secondes sur un P4 à 3 Ghz).

Didacticiel - Etudes de cas Charger un fichier EXCEL

TANAGRA 1.1.2 - [Dat	aset (adult.xls)]							
Tile Diagram Componen	t Window Help					_ @ >		
D 📽 🖬 🗱								
	Default title				Dataset ((adult.xls)		
Dataset (adult.xls)				Paran	meters			
	Database : D:\DataMining\Databases_for_mining\benchmark_datasets\adult\adult.xls							
				Res	sutts			
			15 attribute(s) 48842 example(s	5)				
			Attribute	Category	Informations			
			age	Continue	-			
			workclass	Discrete	8 values			
			fnlwgt	Continue	-			
			education	Discrete	16 values			
			education_num	Continue	-	~		
			Component	s				
Data visualization	Descriptive stats	Instar	nce selection	Featur	e construction	Feature selection		
Regression	Factorial analysis	с	lustering	Sp	ov learning	Meta-spv learning		
Spv learning assessment	Association							
↔ Correlation scatterplot	📴 Export dataset	Sc:	atterplot		/iew dataset	🖾 View multiple scatterplot		

Caractériser la variable « CLASS »

Nous allons essayer de prédire la variable classe à partir des autres variables. Nous utilisons pour cela le composant GROUP CHARACTERIZATION après avoir placé CLASS en TARGET et les autres attributs en INPUT. **Enregistrer le diagramme de traitements correspondant**.

TANAGRA 1.1.2 - [Group caracterization 1]												
File Diagram Component Window Help									- é			
📽 🔲 👫												
Default title	Default title Group caracterization 1											
Dataset (adult.xls)	Parameters Results											
😑 🏠 Define status 1												
Group caracterization 1												
	Description of "class"											
	class	=less			class=n	nore						
	Examples			37155	Examples			11687				
	Att - Desc	Test value	Group	Overral	Att - Desc	Test value	Group	Overral				
	Continuous attributes				Continuous attributes							
	fnlwgt	1.4	190039.57	189664.13	education_num	73.5	11.60	10.08				
	capital_loss	-32.6	54.15	87.50	age	50.9	44.28	38.64				
	capital_gain	-49.3	147.01	1079.07	hours_per_week	50.3	45.45	40.42				
	hours_per_week	-50.3	38.84	40.42	capital_gain	49.3	4042.24	1079.07				
	age	-50.9	36.87	38.64	capital_loss	32.6	193.53	87.50				
	education_num	-73.5	9.60	10.08	fnlwgt	-1.4	188470.57	189664.13				
	Discrete attributes				Discrete attributes							
	marital_status=Never-married	70.5	41.40%	33.00%	marital_status=Married-civ-spouse	98.5	85.43%	45.82%				
	relationship=Own-child	49.9	20.10%	15.52%	relationship=Husband	89.2	75.69%	40.37%				
	sex=Female	47.4	38.82%	33.15%	sex=Male	47.4	84.86%	66.85%				
	relationship=Not-in-family	42.1	30.43%	25.76%	occupation=Exec-managerial	46.6	24.88%	12.46%				
	occupation=Other-service	34.3	12.70%	10.08%	education=Bachelors	39.9	28.35%	16.43%				
		С	omponents									
Data visualization Descriptive stats	Instance selection	Feature c	onstruction	n F	eature selection Regr	ression	1	Factorial analys	is			
Clustering Spv learning	Meta-spv learning	Spv learning	g assessme	nt	Association							
Cross Tabulation Group caracterization Group caracterization	on di Non-parametrical cont stat di One-way ANOVA	Test (KW)	g assessmer ≰Univariat ∭Univariat	e continuo e discrete	Association us stat stat							

Nous constatons que les personnes à revenus élevés sont surtout des hommes mariés assez âgés exerçant une activité managériale, ils ont un haut niveau d'éducation et travaillent beaucoup (sic !).

Ce sont également des personnes qui ont eu à la fois des gains et des pertes en capital plus élevés que la moyenne. La question est de savoir comment interpréter cette dernière information ?

En revenant sur le fichier initial (ADULT.XLS), nous constatons que ces deux variables regroupent deux types d'informations : si la personne ne place pas sont argent, le gain (la perte) est égal à zéro ; et s'ils placent leur argent, la variable indique le montant du gain (de la perte). De plus, un individu ayant réalisé un gain ne peut pas réaliser de perte, et inversement.

× 1	licrosoft Exce	el - adult.xls												-		×
	<u>Fichier</u> <u>E</u> dition	<u>A</u> ffichage <u>I</u> nse	ertion Forma <u>t</u> (<u>O</u> utils <u>D</u> onnées	Fe <u>n</u> être	2								J	_ 8	×
	🖻 🖪 🔒	a 🐧 🖤 🛛	X 🖻 🛍 ダ	10 × 01 ×	Ω	f _*	E 2↓	Z↓		4 2	» •	G	* .0	,00 • .0	<u></u>	• *
_	L12 •	= 0							-			1				_
	Н		J	K	L			М		N			0			
1	relationship	race	sex	capital gain	capital	oss	hours	per	wer	ative co	untr	class	-			
2	Not-in-family	White	Male	2174		0			40 L	Jnited-St	ates	less		_		-
3	Husband	White	Male	0		0			13 U	Jnited-St	ates	less				
4	Not-in-family	White	Male	0		0			40 U	Jnited-St	ates	less				
5	Husband	Black	Male	0		0			40 U	Jnited-St	ates	less				
6	Wife	Black	Female	0		0			40 C	Cuba		less				
7	Wife	White	Female	0		0			40 U	Jnited-St	ates	less				
8	Not-in-family	Black	Female	0		0			16 J	amaica		less				
9	Husband	White	Male	0		0			45 U	Jnited-St	ates	more				
10	Not-in-family	White	Female	14084		0		ę	50 U	Jnited-St	ates	more				
11	Husband	White	Male	5178		0			40 JU	Jnited-St	ates	more				
12	Husband	Black	Male	0		0		{	30 U	Jnited-St	ates	more				
13	Husband	Asian-Pac-Isl	Male	0		0			40 li	ndia		more				
14	Own-child	White	Female	0		0			30 U	Jnited-St	ates	less				
15	Not-in-family	Black	Male	0		0			50 L	Jnited-St	ates	less				
16	Husband	Asian-Pac-Isl	Male	0		0			40 L	Jnited-St	ates	more				
17	Husband	Amer-Indian-E	Male	0		0			45 N	/lexico		less				
18	Own-child	White	Male	0		0			35 U	Jnited-St	ates	less				
19	Unmarried	White	Male	0		0			40 L	Jnited-St	ates	less				
20	Husband	White	Male	0		0			50 U	Jnited-St	ates	less				
21	Unmarried	White	Female	0		0			45 U	Jnited-St	ates	more				
22	Husband	White	Male	0		0			50 U	Jnited-St	ates	more				
23	Unmarried	Black	Female	0		0			20 U	Jnited-St	ates	less				
24	Husband	Black	Male	0		0			40 L	Jnited-St	ates	less				-
• •	► ► \adult	/					4								_ ▶	
Deg	isin + 🗟 🌀	Formes automa	tiques 👻 🔪 🍾		4 🗕	<u></u> -	- 🚄 -	· <u>A</u> ·	- =	• = E			•			
Prê	t											NUM				

Nous allons donc créer une variable indicatrice permettant de savoir si l'individu place son argent ou pas.

	🛚 Microsoft Excel - adult.xls														
	Eichier Edition	Affichage Inse	ertion Forma <u>t</u>	<u>O</u> utils <u>D</u> onnées	Fe <u>n</u> être <u>?</u>									_ 8	×
	📽 🖪 🔒	a 🛯 🖓 🛛	አ 🖻 💼 🚿	N + CI +	🤮 Σ f _n	Ę			1	? »	G	:00	,00 ,0	<u>.</u>	- >>
-	M2 -	= =5	SI(OU(K2>0;L2	2>0):"ves":"no	") 🗲				for	mu 1	a 🗌				
	Н	1	J	K	L		М		1011	Indi	a _	0		(Contract	-
1	relationship	race	sex	capital gain	capital los:	s	capital	inves	thours	per we	native	e cou	untr	class	
2	Not-in-family	White	Male	2174		0	yes	_	7	40	Unite	d-Sta	ates	less	
3	Husband	White	Male	0		0	no			13	Unite	d-Sta	ates	less	
4	Not-in-family	White	Male	0		0	no			40	Unite	d-Sta	ates	less	
5	Husband	Black	Male	0		0	no			40	Unite	d-Sta	ates	less	
6	Wife	Black	Female	0		0	no			40	Cuba			less	
7	Wife	White	Female	0		0	no			40	Unite	d-Sta	ates	less	
8	Not-in-family	Black	Female	0		0	no			16	Jama	ica		less	
9	Husband	White	Male	0		0	no			45	Unite	d-Sta	ates	more	
10	Not-in-family	White	Female	14084		0	yes			50	Unite	d-Sta	ates	more	
11	Husband	White	Male	5178		0	yes			40	Unite	d-Sta	ates	more	
12	Husband	Black	Male	0		0	no			80	Unite	d-Sta	ates	more	
13	Husband	Asian-Pac-Isl	Male	0		0	no			40	India			more	
14	Own-child	White	Female	0		0	no			30	Unite	d-Sta	ates	less	
15	Not-in-family	Black	Male	0		0	no			50	Unite	d-Sta	ates	less	
16	Husband	Asian-Pac-Isl	Male	0		0	no			40	Unite	d-Sta	ates	more	
17	Husband	Amer-Indian-B	Male	0		0	no			45	Mexi	со		less	
18	Own-child	White	Male	0		0	no			35	Unite	d-Sta	ates	less	
19	Unmarried	White	Male	0		0	no			40	Unite	d-Sta	ates	less	
20	Husband	White	Male	0		0	no			50	Unite	d-Sta	ates	less	
21	Unmarried	White	Female	0		0	no			45	Unite	d-Sta	ates	more	
22	Husband	White	Male	0		0	no			60	Unite	d-Sta	ates	more	
23	Unmarried	Black	Female	0		0	no			20	Unite	d-Sta	ates	less	
24	Husband	Black	Male	0		0	no			40	Unite	d-Sta	ates	less	-
K	I I I Adult						4							•	
Deg	sin • 🗟 🕑	Formes automa	itiques 🕶 🔨 🦄		4 🗕 🖄	9 -	<u></u>	A -		≓∎					
Prê	t										NUM				1

Après avoir enregistré le fichier modifié puis fermé EXCEL, il suffit de fermer puis de ré-ouvrir le diagramme de traitements dans TANAGRA.

1.1.2 TANAGRA							
File Diagram Windov	w Help						
	Tanagra					? 🔀	
Data mir	Regarder dans :	🚞 adult		🕑 G 🏚	≥		
	Mes documents récents Bureau Mes documents	adult, tdm					
Data visualizati Feature select Spv learning EC Cross Tabulatio	Poste de travail	Nom du fichier : Fichiers de type :	adult.tdm Text data mining diagram ((*.tdm)		Ouvrir Annuler	nt stat
<							.:

R.R.

Nous constatons alors :

(1) que le diagramme est bien conservé ;

(2) que les données maintenant sont enrichies de la nouvelle colonne que nous venons de créer.

TANAGRA 1.1.2 - [Dataset (adult.xls)]								
💇 File Diagram Component Window Help					- 8 ×			
D 📽 🖪 🔛								
Default title			Results		~			
Dataset (adult.xls) = 100 Dataset (adult.xls)	16 attribute(s) 48842 example(s))						
Group caracterization 1	Attribute	Category	Informations					
	age	Continue	-	-				
	workclass	Discrete	8 values					
	fnlwgt	Continue	-					
	education	Discrete	16 values					
	education_num	Continue	-					
	marital_status	Discrete	7 values					
	occupation	Discrete	14 values					
	relationship	Discrete	6 values					
	race	Discrete	5 values					
	sex	Discrete	2 values					
	capital_gain	Continue	-					
	capital_loss	Continue	-					
	capital_invest	Discrete	2 values	\leftarrow				
	hours_per_week	Continue	-	•				
	native_country	Discrete	41 values					
	class	Discrete	2 values					
	L.				<u> </u>			
	Compone	nts						
Data visualization Descriptive stats	: Instan	ice selecti	ion Fe	ature construction				
Feature selection Regression	Facto	orial analys	sis	Clustering				
Spv learning Meta-spv learning	g Spv learr	ning assess	ment	Association				
E Cross Tabulation II Group caract	erization 🚺	Linear c	🕂 More Univaria	te cont stat				
K					>			

Dans DEFINE STATUS, ajouter la nouvelle variable parmi les INPUT. Nous constatons dès lors que la variable CAPITAL indique surtout un comportement vis à vis de son épargne : les personnes ayant un revenu élevé ont tendance à placer leur capital (12.9% des personnes placent leur argent, ils sont 31% à le faire parmi les personnes à revenus élevés).

TANAGRA 1.1.2 - IGroup caracterization	41											
File Diagram Component Window Help										_ 2	×	
										Jam Jam		
Defaultible										3		
		class=les	5			-	class=mor	e				
Dataset (adutt.xis)					37155	Examples				11687		
Group caracterization 1	c			Group	Overral	Att - Desc		Test value	Group	Overral	- III	
	us attri	butes				Continuo	us attributes	-			-	
	1		1.4	190039.57	189664.13	education	า_กนก	73.5	11.60	10.08		
	055		-32.6	54.15	87.50	age		50.9	44.28	38.64		
	ain		-49.3	147.01	1079.07	hours_pe	r_week	50.3	45.45	40.42		
	r_week	r_week		38.84	40.42	capital_gain		49.3	4042.24	1079.07		
	n_num		-50.9	36.87	38.64	capital_loss		32.6	193.53	87.50		
			-73.5	9.60	10.08	fnlwgt			188470.57	189664.13		
	attribu	tes			Discrete attributes							
	tatus=N	lever-married	70.5	41.40%	33.00%	marital_status=Married-civ-spouse		98.5	85.43%	45.82%		
	nvest=n	D	67.1	92.78%	87.07%	relationship=Husband		89.2	75.69%	40.37%		
I I I I I I I I I I I I I I I I I I I	hip Our	r ohitd	10.0	20.40%	15.52	capital_in	west=yes	67.1	31,10%	12.93%		
•	ale		47.4	38.82%	33.15%	sex=Male		47.4	84.86%	66.85%		
	hip=Not	-in-family	42.1	30.43%	25.76%	occupatio	on=Exec-managerial	46.6	24.88%	12.46%		
	on=Othe	er-service	34.3	12.70%	10.08%	education	n=Bachelors	39.9	28.35%	16.43%		
	hip=Unm	arried	31.7	12.96%	10.49%	education	n=Masters	38.5	12.48%	5.44%	~	
	<					1	I.			>		
			Cor	nponents								
Data visualization Descriptive sta	its	Instance selec	tion	Feat	ure constru	uction	Feature selection		Regre	ssion		
Factorial analysis Clustering		Spv learning	8	M	eta-spv lear	ning	ng Spv learning assessment			Association		
<											>	