Tutorial overview

This tutorial shows basic characteristics of TANAGRA user interface, through the analysis of the « Breast.txt » dataset.

This well-known dataset come from the medical domain, consists of the characteristics of cells sampled on women presenting (or not) a malignant tumor.

In this tutorial, you'll learn to use the following components:

Tab	Operator (Component)	Function					
Data visualization	View dataset	View the contents of the data file in a grid					
Feature selection Define status		Specify the attributes to use					
Descriptive stats	Univariate continuous stat	Descriptive statistics for continuous attributes					
Descriptive stats	Univariate discrete stat	Descriptive statistics for discrete attributes					
Descriptive stats	Group characterization	Statistics for sub-populations					

Importing and viewing data in TANAGRA

- Creating a new data mining diagram
 - 1 Choose File/New... in the main menu of TANAGRA.
 - 2 Enter a title for the diagram : « TANAGRA Basics ».

3 – Enter the name of the associated file in which you will save your work (« TANAGRA_Basics.bdm »). Before click on place in the vourself directory

« ...\TANAGRA\Tutorials ».

Enregistr	rer sous			? ×
Da <u>n</u> s:	🕞 Tutorials	• E		
J				
No <u>m</u> :	Tanagra_Basics		Ē	nregistrer
<u>Type</u> :	Binary data mining diagram	n	•	Annuler
	C 10 00 00			

4 – Select the text file containing the data you want to explore by clicking on this icon: For this tutorial, choose the file "breast.txt", located in TANAGRA subdirectory « Dataset ».

F	anagra Basics
	ata mining diagram file name :
k	:\Program Files\Tanagra\Tutorials\Tanagra_Basics.bdm 🛛 🛛 🔚
D	ataset (text file) :
IE.	

- 5 Validate by clicking the OK button to start loading.
 - > Adding an operator to the diagram in order to visualize data

1 – Add a **View dataset** component to the diagram. To do this, click on the DATA VISUALIZATION tab of the components palette.

Drag and drop **View Dataset** from components palette to stream diagram, under the "Dataset" node (the node must appear to be selected).

2 – Then click on the "View dataset" node to select it (if not yet), and right-click on it to activate the popup menu: choose the *View* command. Data are displayed in the right frame.

Eile Diagram Component	<u>W</u> indow <u>H</u> elp						_ 8
Tanagra I	Basics		clump	ucellsize	ucellshap	emgadhesic	nsepio
Dataset (breast.txt)		1	4	2	2	1	2
View dataset 1		2	1	1	1	1	2
	Parameters	3	2	1	1	1	2
	Fuseute	4	10	6	6	2	4
		5	4	1	1	1	2
_	<u></u>	6	1	1	1	1	2
		7	1	1	1	1	2
		8	5	1	1	1	2
		9	3	1	1	1	2
		10	1	1	1	1	2
		11	5	3	3	2	3
		12	4	2	2	1	2
		13	1	1	1	1	2
		14	2	1	1	1	2
		15	4	1	1	1	2
		16	1	1	1	1	2
							•
		Co	mponents				
Data visualization	Descriptive stats	Inst	ance selection	Feature c	onstruction		
Feature selection	Regression	Fac	torial analysis	Clus	tering		
Sov learning	Meta-spy learning	Sov le	arning assesment	Asso	ciation		
	t sta	1	g assessment			18	
Correlation scatterplot	Scatterplot	<u> </u>	/iew multiple scatt	erplot			
Export dataset	🚟 View dataset						

Getting some descriptive statistics

Use and importance of the Define status operator

In TANAGRA you can build sequences of operations.

However, almost all operators require, before executing it, that you have defined the attributes to use, and how to use them (View dataset, that we've just used before, is an exception to this rule). To avoid repetition of status definition for each operator, TANAGRA centralizes this declaration in the **Define status** component.

> Basic statistics on each attribute (min, max, average, standard deviation)

1 – Add a **Define Status** component (FEATURE SELECTION tab) to the diagram, under the « Dataset » node. (if, by error, you put it under the « View Dataset » node, you can delete it via the *Diagram / Delete component* menu)

2 – Then click on the "Define status" node to select it, and activate its popup menu by clicking the mouse right button: choose the *Parameters...* Command.

3 - In the dialog box that appears, choose some continuous variables (these marked with a C blue letter): click on them (in the left list), then press the arrow button. They will act as input, since it is the active tab.

Validate with OK, the window closes.

Define attributes status Parameters Report					
Attributes :		Target	Input	Illustrat	ive:
C clump C ucellsize C ucellshape C mgadhesion C sepics C bnuclei C bchromatin C normnucl C mitoses D class	*	ucellsize ucellshape sepics bchromatin mitoses			
E E		Cli	ear selec	tion	
		ОК	Can	icel	Help

4 – Add a **Univariate continuous stats** component (DESCRIPTIVE STATS tab) to the diagram, under the «Define status 1» node. In the popup menu, choose the *View* command. Descriptive statistics for the selected attributes are displayed in the right frame.

s								
<u>s</u>		_						
2:								
				Univa	riate con	tinuous stat 1		
		Parameters						
View dataset 1								
lous stat 1	Examples : 6	599						
Parameters					Res	utts		
Execute	ttribute	Min	Max	Average	Std-dev	Std-dev/avg		
⊻iew	ellsize	1	10	3.1345	3.0493	1.0279		
	ucellshape	1	10	3.2074	2.9698	1.0800		
	sepics	1	10	3.2160	2.2127	1.4534		
	bchromatin	1	10	3.4378	2.4366	1.4109		
	mitoses	1	10	1.5894	1.7139	0.9274		
	Execution tir Created at 0	me : (15/01	0 ms. 1/04 1	1:59:43				
J,	Comp	oner	nts					
Descriptive stats	Instanc	e se	lectio	n	Featur	e construction		
Regression	Factor	ial a	nalysi	s	C	lustering		
Meta-spv learning	Spv learni	ing a	issesn	nent	A:	ssociation		
Linear correlation	lati,	One	-way	anova		Univariate discret	e stat	
📶 Non-parametrical T	est (KW) 📠	Univ	ariate	e continu	ous stat			
	Execute <u>View</u> Descriptive stats Regression Meta-spv learning Linear correlation Meta-spr learning	Execute View ellsize ucellshape sepics bchromatin mitoses Execution til Created at C Comp Descriptive stats Regression Regression Regression Meta-spv learning Spv learn Muta-spv learning Linear correlation Muta-spx learning Spv learn Muta-spx learning Spv learn Muta-spx learning Spv learn Muta-spx learning Spv learn Muta-spx learning Spv learn Muta-spx learning Spv learn Muta-spx learning Spv learning	Execute View elisize 1 ucelishape 1 sepics 1 bchromatin 1 mitoses 1 Execution time : Created at 05/01 Execution time : Component Descriptive stats Regression Factorial a Meta-spv learning Spv learning Spv learning at 0ne I univ	Execute Hin Max View elisize 1 10 ucellshape 1 10 10 sepics 1 10 10 bchromatin 1 10 10 bchromatin 1 10 10 Execution time : 0 ms. Created at 05/01/04 1 10 Descriptive stats Instance selection Regression Factorial analysi Meta-spv learning Spv learning assess Image: Linear correlation Image: One-way Non-parametrical Test (KW) Image: Univariate	Execute Hin Hax Average View elisize 1 10 3.1345 ucelishape 1 10 3.2074 sepics 1 10 3.2160 bchromatin 1 10 3.4378 mitoses 1 10 3.4378 mitoses 1 10 1.5894 Execution time : 0 ms. Created at 05/01/04 11:59:43 Components Descriptive stats Instance selection Regression Factorial analysis Meta-spv learning Spv learning assesment If Linear correlation If One-way anova Mon-parametrical Test (KW) If Univariate continu	Execute Itribute Hin Hax Average Std-dev Elsize 1 10 3.1345 3.0493 ucellshape 1 10 3.2174 2.9698 sepics 1 10 3.2160 2.2127 behromatin 1 10 3.4378 2.4366 mitoses 1 10 3.4378 2.4366 mitoses 1 10 1.5894 1.7139 Execution time : 0 ms. Created at 05/01/04 11:59:43 Components Execution time : 0 ms. Created at 05/01/04 11:59:43 Components Descriptive stats Instance selection Feature Feature Regression Factorial analysis Com Meta-spv learning Spv learning assesment Astrono If Univariate contributous stat Instance contributous stat	Execute Itribute Hin Max Average Std-dev Std-dev/avg View elisize 1 10 3.1345 3.0493 1.0279 ucellshape 1 10 3.2074 2.9698 1.0800 sepics 1 10 3.2160 2.2127 1.4534 bchromatin 1 10 3.4378 2.4366 1.4109 mitoses 1 10 1.5894 1.7139 0.9274 Execution time : 0 ms. Created at 05/01/04 11:59:43 0.9274 Execution time : 0 ms. Created at 05/01/04 11:59:43 Descriptive stats Instance selection Feature construction Regression Factorial analysis Clustering Meta-spv learning Spv learning assesment Association Mil Univariate discret Mon-parametrical Test Image: One-way anova Image: Univariate continuous stat State	

5 – Add another **Define status** component to the « Dataset » node, and select the discrete attribute « class ».

Add a **Univariate discrete stats** operator under the node « Define status 2», and choose the *View* command as previously. Below is the result you should obtain.

Tanagra B	lasics						
⊐			_	Univariate discre	ete stat 1		
View dataset 1		Parameters Attributes : 1 Examples : 699					
🖃 🙀 Define status 1							
🛄 📶 Univariate con	tinuous stat 1						
🖻 🙀 Define status 2				Results			
III Univariate disc	rete stat 1	Attribute	Gini	Dis	tribution		
				Values	Count	Percent	
		class	0.4518	begnin	458	65.52 %	
				malignant	241	34.48 %	
		Execution f Created at	time : 0 r : 05/01/0	ns.)4 15:59:31			
		Execution † Created at	time : 0 r : 05/01/(ns. 14 15:59:31			
		Execution t Created at	time : 0 r : 05/01/0 ponent:	ns. 14 15:59:31 8			
Data visualization	Descriptive stats	Execution of Created at Com Instan	time : 0 r : 05/01/0 ponent: ce selec	ns. 04 15:59:31 s stion Feature of	Instruction		
Data visualization	Descriptive stats Regression	Execution 1 Created at Com Instan Facto	time : 0 r : 05/01/0 ponent: ce seleo prial ana	ns. 04 15:59:31 s stion Feature of lysis Clust	Instruction		
Data visualization 5 Feature selection Spv learning	Descriptive stats Regression Meta-spv learning	Execution of Created at Com Instan Facto Spv learn	time : 0 r : 05/01/0 ponent: ce selec orial ana ning assi	ns. 14 15:59:31 s stion Feature of lysis Clust esment Assoc	Instruction ering iation		
Data visualization	Descriptive stats Regression Meta-spv learning M. Non-parametrical	Execution f Created at Com Instan Facto Spv learn Fest (KW)	time : 0 r : 05/01/0 ponent: ce selec orial ana ning asso I Univari	ns. 14 15:59:31 s stion Feature or lysis Clust esment Assoc late discrete stat	ering iation		

Statistics for each sub-population (comparing the characteristics of the women presenting -- or not -- a malignant tumor)

1 -- Add another **Define status** operator to the « Dataset » node. Choose the *Parameters...* command in its popup menu. In the dialog window, select some continuous attributes as Input, and the discrete attribute as Target.

2 – Under this node add a **Group caracterization** operator, and choose *View* in its popup menu.

Tanagra - [Group caracte	rization 1] Window Help								-	
									_	222
Tanagra I	Basics			Gri		cterization 1				
⊡-∰ Dataset (breast.txt)		Parameters								
🔤 View dataset 1		r ar amotor s								
🖻 🄹 Define status 1				Res	autts				61	
Univariate cor	itinuous stat 1			D	escriptio	n of "class"				11
III Univariate dis	crete stat 1	class=begnin class=malignant								
🖻 🙀 Define status 3	Examples	10		458	Examples					
Group caract	Group caracterization 1		Test value	Group	Overral	Att - Desc	Test value	Group	Overral	
		Continuous	attributes		Continuous	attributes				
		mitoses	-11.2	1.06	1.59	ucellshape	21.6	6.56	3.21	
		sepics	-18.0	2.12	3.22	ucellsize	21.6	6.57	3.13	
		mgadhesion	-18.4	1.36	2.81	bnuclei	21.5	7.60	3.56	
		normnucl	-18.8	1.29	2.87	bchromatin	20.0	5.98	3.44	
		clump	-18.9	2.96	4.42	clump	18.9	7.20	4.42	-
		Compor	nents							
Data visualization	Descriptive stats	Instance	selection		Featur	e constructi	on			
Feature selection	Regression	Factorial	l analysis		C	lustering				
Spv learning	Meta-spv learning	Spv learning	g assesmer	nt	A	ssociation				
Cross Tabulation	📶 Non-parametrical Tr	est (KW) 📶 Ur	nivariate d	iscrete	e stat					_
🚺 Group caracterization	📶 One-way anova									
Linear correlation	📶 Univariate continuo	us stat								

Inspecting these results, we notice that, on average, women with benign tumor present smaller values of "mitoses" (1.06, versus 1.59 for the complete population). On the other hand, the values of "ucellshape" attribute are, still on average, higher for women with malignant tumor (6.56 versus 3.21).