## **Tutorial overview**

This tutorial shows how to apply a supervised learning method on your data; for this example we're going to analyze the « breast.txt » dataset, using the ID3 method (decision tree).

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

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

Tab	Operator (Component)	Comment
Feature selection	Define status	Specify the attributes to use
Meta-spv learning	Supervised learning	A container for machine learning operators
Spv learning	ID3	A machine learning operator

## Loading the data in TANAGRA

- Opening an existing diagram
- 1 Choose *File/Open...* in the main menu of TANAGRA.
- 2 In the "Dataset" subdirectory of TANAGRA, select the file named « breast.bdm ».

## Using a supervised learning operator

> Specify the status of the attributes for the analysis

1 – Add a **Define Status** operator under the "Dataset" node, by clicking on its icon in the shortcuts toolbar (as shown below). A dialog box appears automatically, allowing the definition of the status of the attributes.

🎌 Tanagra - [Dataset (breas	t.txt)]					
m Eile Diagram Component	<u>W</u> indow <u>H</u> elp					_ 8 ×
Tanagra Decis	sion Tree			Datacot	t (broad tyt)	-
🖃 🥅 Dataset (breast.txt)			_	Datase		
🔂 Define status 1		Database : C	:)Program]	Files\Tapagra\[	)ataset\breast.txt	
				, nos tranagia e		
				F	Results	
		10 attribute 699 example	(s) e(s)			
		Attribute	Category	Informations		
		clump	Continue	-		
		ucellsize	Continue	-		
		ucellshape	Continue	-		
		mgadhesion	Continue	-		
		sepics	Continue	-		
		bnuclei	Continue	-		
		bchromatin	Continue	-		_
	]	<u> </u>	• ··			
Data visualization	Descriptive stats	Instance	onencs e selection	n Feat	ture construction	
Feature selection	Regression	Factor	ial analysis		Clustering	
Spv learning	Meta-spv learning	Spv learni	ng assesmi	ent	Association	
Binary logistic regression	 Linear discriminant	analysis 🗽 P	roptotye-	NN		
A. ID3	Multilayer percept	ron 🔅 R	adial basis	function		
tiri K-NN	Reso Naive bayes					

2 – Before all, be sure that the active tab in the dialog is the "Input" one. Then select the continuous attributes in the left list by clicking the corresponding button below the list (as shown in the following screenshot), and hit the arrow button to bring them in the Input list.

Define attributes status Parameters Report			
Attributes : C clump C ucellsize C ucellshape C mgadhesion C sepics C bnuclei C bchromatin C normnucl C mitoses D class	Target	Input    ust	rative
	(C)A	ar selection	
Select continuous attributes	ОК	Cancel	Help

3- In the same dialog box, activate the Target tab. Select the « class » attribute in the list and click the arrow button.

Define attributes status Parameters Report	
Attributes :	Target Input Illustrative
C clump C ucellsize C ucellshape C mgadhesion C sepics C bnuclei C bchromatin C normnucl C mitoses	Class Add selected attributes
E. E.	Clear selection
	OK Cancel Help

4 – Now you have defined the class attribute (« class » = Target), and the descriptors to do this (the others = Input). Click OK to validate and close this dialog box.

> Choosing the meta-operator

To obtain a better prediction model, you may execute a machine learning method many times on a weighted examples, and aggregate classifiers. This method, known as "boosting - arcing" can be applied to any learning algorithm.

TANAGRA implements this concept by encapsulating the machine learning operator in a metaoperator. It is the meta-operator that launches the method many times, taking some specific parameters into account.

In this quick start tutorial we won't describe these advanced experimentations, we'll simply launch the ID3 method once. But as TANAGRA always needs a meta-operator, a basic one exists that suits our case, launching only one time the encapsulated method: the **Supervised Learning** operator.

1 – Add a **Supervised learning** component (META-SPV LEARNING tab) to the diagram, under the «Define status 1» node.

🎌 Tanagra - [Dataset (brea	ast.txt)]					
💇 <u>F</u> ile <u>D</u> iagram Componen	t <u>W</u> indow <u>H</u> elp					
D 📽 🖬 🗠						
Tanagra De	cision Tree			Dataco	t (broad tot)	-
🕞 🛅 Dataset (breast.txt)			_	Datase		
Define status 1		Database : (	NProgram I	ra Files\Tenegre\T	Pameters	
	ervised earning			rilostranagraia		
					Results	
		10 attribute	(5)			
		699 example	e(s)			
		Attribute	Category	Informations		
		clump	Continue	2		
		ucellsize	Continue	-		
		ucellshape	Continue	-		
		mgadhesion	Continue	-		
		sepics	Continue	-		
		bnuclei	Continue	-		
		bchromatin	Continue	-		
<u>  </u>		<u> </u>	-			
Data da alla da	Der de la constante	Compo	onents	-		2
Data visualization	Descriptive stats	Instanc	e selection	n Fea	ture construction	
Feature selection	Regression	Factori	ial analysis		Clustering	
Spv learning	Meta-spv learning	Spv learni	ng assesme	ent	Association	
Arcing [Arc-x4]	Supervised Learning					
Bagging						
Boosting						
<u> </u>						
						1.

> Adding a supervised learning component

1 – In the components palette, SPV LEARNING tab, drag with the mouse a **ID3** component on the « Supervised Learning » node you've added just before.

The operator is included in the meta one, so we don't see another node under the meta node, but we notice that the name of the method appears in the meta node label.

🎢 Tanagra - [Dataset (breast	.txt)]					
🕎 Eile Diagram Component	<u>W</u> indow <u>H</u> elp					_ 8
D 📽 🖬   📬						
Tanagra Decisi	on Tree			Dataco	t (broad tut)	
🖃 🛄 Dataset (breast.txt)			_	Datase		
🗄 🤹 Define status 1	$\frown$	Database : C	:)Program]	Files\Tanagra\[	Dataset\breast.txt	
Supervised Lear	rning 1 (ID3)					
	N N			1	Results	
		10 attribute	(\$)			
		699 example	e(s)			
		Attribute	Category	Informations		
		clump	Continue	2		
		ucellsize	Continue	-		
		ucellshape	Continue	-		
		mgadhesion	Continue	-		
		sepics	Continue	2		
		bnuclei	Continue	-		
		bchromatin	Continue	-		1
ļ		<u>`</u>				
Data visualization	Descriptive stats	lostano	onents e selection	- Fea	ture construction	
Easture calestion	Pegression	Eactor	al analysis	1 100	Clustering	
	Regression	Caulaana	iat amatysis		Clustering	
Spv learning	Weta-spv learning	spv learni	ng assesmi	ent	Association	
Binary logistic regression	Linear discriminant	: analysis 🛛 🚺 P	roptotye-	NN		
n as 103	Ser Multilayer percept	ron 🕬 R	adial basis	runction		
	Inaniaque payes					
1						

> Defining the method parameters (for ID3 operator)

1 – Activate the popup menu of the « Supervised learning (ID3) » node by right-clicking on it. There's a new command under the well-known *Parameters*: *Supervised parameters*...

TANAGRA tutorials- January 2004 – Quick start : Using a supervised learning method:

<sup>®</sup> <b>Tanagra - [Dataset (breas</b> ■ <u>F</u> ile <u>D</u> iagram Component	t. <b>txt)]</b> ∭undow <u>H</u> elp					
) 🚅 🔚   🎲						
Tanagra Decis	sion Tree			Datase	et (breast.txt)	
∃ 🛅 Dataset (breast.txt)				Pa	rameters	
🖻 🚰 Define status 1		Database : C	∷\Program I	Files\Tanagra\	Dataset\breast.txt	
Supervised Lea	Parameters					
	<u>S</u> upervised param	neters	(*)		Results	
	Execute	nple	(S) 2(S)			
	_ ⊻iew					
		Attribute	Category	Informations		
		clump	Continue	-		
		ucellsize	Continue	-		
		ucellshape	Continue	-		
		mgadhesion	Continue	-		
		sepics	Continue	-		
		bnuclei	Continue	-		
		bchromatin	Continue	-		
		Compo	onents			
Data visualization	Descriptive stats	Instance	e selection	n Fea	ture construction	
Feature selection	Regression	Factori	al analysis	l l	Clustering	
Spv learning	Meta-spv learning	Spv learni	ng assesme	ent	Association	
≂Binary logistic regression aID3	Linear discriminant	analysis 🗽 P ron 🔅 R	roptotye-l adial basis	NN function		
<u>-</u> K-NN	Naive bayes					

The first (*Parameters...*) applies to the meta-operator, the second (*Supervised parameters...*) applies to the supervised learning operator. Choose this last one.

2 - In the opening dialog box, to take into account the size of the data file (699 records), we modify the ID3 parameters as follows:

IT 🗌	ee stopping	rule		
	Min size f	fro split : 50	-/-	
	Min size of	leaves : 10	-⁄-	
1	Max depth of	the tree 10	-/-	
1	Min entropy g	gain for splittin	<u></u> ۵.0300	

3 - Click OK to validate.

## **Executing the learning process**

1 – In the contextual menu of the meta node, choose the *View* command. Results are displayed in the right frame.

anagra - [5upervised Learning   [103]] file Diagram Component Window Help						-	_ 2
Tanagra Decision Tree	Classi	fior n	erfo	rman	- AC		
🔟 Dataset (breast.txt)	Classi	ner p	erro		LES		
🖻 🙀 Define status 1	E	rror rate			0.0	0472	
Supervised Learning 1 (ID3)	Value	es predictio	on		Confusi	on matrix	
	Value	Sencihility	Pred.		begnin	malignant	Sum
	* aluc	Sensibility	error	begnin	442	16	458
	begnin	0.9651	0.0370	malignant	17	224	241
	malignant	0.9295	0.0667	Sum	459	240	699
	Classi	fier c	hara	cteris	itics		
	Classi	fier c	hara	cteris	itics		
	Classi Tree descri	fier c	hara	octeris	itics		
	Classi Tree descri Number of n Number of le	fier c	hara	icteris	itics		

The resubstitution error rate seems to be good (4,72 %). We see in the confusion matrix that false positive and false negative examples are equivalent.

Finally we notice, looking at the results of our analyses, the importance of the "ucellsize" attribute in this problem.