# Subject

Copy paste feature into the diagram.

When we define the data analysis process into Tanagra, it is possible to copy components (or entire branches of components) towards another location into the diagram. This feature is very helpful when we have to repeat sequences of treatments in different parts of the diagram. The settings are also duplicated.

In this tutorial, we show how to copy a component or a branch. We will see that this feature is helpful when, for instance, we deal with the performance comparisons of supervised learning algorithms on the same dataset. In this context, the processing sequence is always the same, only the method that we want to evaluate is different.

We work on the same project here. We cannot copy paste components between two opened projects. But, in another tutorial, we show how to save a part of the diagram in an external file. Thus, the same processing sequence can be applied on multiple datasets<sup>1</sup>.

# Dataset

We use the SONAR dataset (<u>http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/sonar.xls</u>). We want to compare the accuracy of various supervised learning algorithms: binary logistic regression, linear discriminant analysis, k-nearest neighbor, support vector machine and PLS regression for classification.

# Copy paste feature of the diagram

#### Importing the dataset

We launch Tanagra. We create a new diagram by clicking on the FILE / NEW menu. We select the SONAR.XLS data file. Tanagra can import the Excel file format, even if the spreadsheet is not installed on the computer<sup>2</sup>.

<sup>&</sup>lt;sup>1</sup> <u>http://data-mining-tutorials.blogspot.com/search/label/Diagram%20management</u>

<sup>&</sup>lt;sup>2</sup> About the handling of the Excel file format, see <u>http://data-mining-tutorials.blogspot.com/2008/10/excel-file-format-direct-importation.html</u>; we can also send the data from Excel using an add-in <u>http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html</u>

gra				
👮 TANAGRA 1.4.7				
File Diagram Window Help				
Changes your dataget and start down				
Choose your dataset and start download				
Diagram title :				
Default title				
Data mining diagram file name :				
D:\Temp\Exe\default.bdm		j		
Determined the set of the state				
Dataset (".txt,".am,".xis) :				
UK	Cancel Hei			
	Tanagra			2 X
	Hegarder tabs :	sonar		
		sonar.xls	1	
	Mes documents			
	récents		N	
	TA I		1	
	Bureau		1 I I I I I I I I I I I I I I I I I I I	
	Duicau			
	Mes documents		1	
	mes documents		1	
Data viewalization Statistics	No			
Feature selection Regression	Peste de trauni			
Sov learning Meta-sov learning	So So		V	
	Nr Nr	m du fielde	~	Ouvrir
Correlation scatterplot			4 00001	
	Favoris reseau Fic	Excel File (97	& 2000J 💙	Annuler

# Specifying the type of the variables

We insert the DEFINE STATUS component in order to define the target attribute (CLASS) and the input attributes (ATTRIBUTE\_01 ... ATTRIBUTE\_60).

TANAGRA 1.4.7 - [Dataset (sonar.	xls)]	
Tile Diagram Component Window H	elp	- 8 ×
Default title		<b>^</b>
🖃 🏢 Dataset (sonar 🕼)	Define attribute statuses	
Define status 1	Parameters	ppar.xls
	Attributes : Target Input Illustrative	
	C attribute_50 Class C attribute_51 C attribute_52 C attribute_53 C attribute_55	Г
	C attribute_55 C attribute_56 C attribute_57 C attribute_58	
	C attribute_59	
Data visualization Sta		
Feature construction Feature		
PLS Clu	L Clear selection	
Spv learning assessment		
	OK Cancel Help	- 10 ° - 1

29/06/2009

### Implementation and assessment of a supervised learning algorithm

We add the BINARY LOGISTIC REGRESSION component (SPV LEARNING tab) into the diagram<sup>3</sup>.

💯 TANAGRA 1.4.7 - [De	fine status 1]					
🏆 File Diagram Componer	nt Window Help					_ 8 ×
	Default title				Define status 1	<u>~</u>
🖃 🏢 Dataset (sonar.xls)	)				Parameters	
🖃 🚼 Define status 1	1		Target : 1		r ai ailictor s	
->-> Supervised	Learning 1 (Binary logistic rea	gression)	Input: 60			
			Illustrative : 0			
					Results	
			Attribute	Target	Input Illustrative	
	Ш	>	attribute_1	-	yes -	<b>~</b>
		Com	ponents			
Data visurlization	Statistics	Nonpara	ametric statistics		Instance selection	ר
Feature construction	Feature selection	F	Regression		Factorial analysis	
PLS	Clustering 🤇	S	pv learning	2	Meta-spv learning	
Spv learning assessment	Scoring	A	Association			
Binary logistic regressi	on A.C-RT		🖧 ID3			1 1+E**Log-Reg TRIRLS
-€4.5 €45	t <u>i</u> ≪c-svc		📴 K-NN			Multilayer perceptron
🔀 C-PLS	Ra Decision List		🕍 Linear	discrim	inant analysis	<sup>1</sup> ₩ultinomial Logistic Re
<						>

The resubstitution error rate is 5.29%.



<sup>&</sup>lt;sup>3</sup> Before the 1.4.7 version, this insertion had to be made in two steps: first, inserting the meta supervised learning algorithm (single learning, bagging, boosting, etc.); second, inserting the learning method. Since the 1.4.7 version, we can insert directly the learning algorithm for the standard approach. For the non standard approach (bagging, boosting, cost sensitive learning, etc.), we must still follow the two steps.

We know that the error rate computed on the learning sample is optimistic. In order to obtain a more reliable evaluation, we use a resampling method, here a cross-validation algorithm. The component can be found in the SPV LEARNING ASSESSMENT tab. We click on the PARAMETERS menu. We set the following settings (TRIALS = 3; FOLDS = 10). We must use the same scheme for the other learning algorithms that we want to evaluate.

Image: Tanagra 1.4.7       File     Diagram       Component       Image: Tanagram	Window Help		
	Default title	Dataset (sonar.xls)	^
Im Dataset (sonar.xls)     Define status 1     Define status	Learning 1 (Binary logistic regression) lidation 1 Parameters Execute View	Cross-validation parameters Parameters Number of repetitions : 3 Number of folds : 10 Save results Save error rate to file DitTemplExelexperiments_results	stic regression) ary logistic regression) ters ts TCES V
Data visualization Feature selection Sov learning 21 Bios-variance decompos 21 Cross-validation	Statistics Nonp Regression Fa Meta-spv learning Spv le ition ?? Train-test	OK Cancel Help	Feature construction Clustering Association

The cross validation error rate estimate is 30.5%.

				Cross-vali	dation 1		
				Parame	eters		
Cross-v	/alidatio	n parameters		-			
Folds		10	}	•			
Trials		3	J				
				Resu	itts		
С¥ егга	or rate						
Rar	nge						
MIN	0.3000						
MAX	0.3100						
Trial E	rr rate						
1	0.3000						
2	0.3050						
3	0.3100						
Overall	CLO22-A	alidation erro	r rate				
	Error	rate			0.3050		
۷a	lues pro	ediction		Confu	ision matrix		
Value	Recall	1-Precision		Rock	Mine	Sum	
Rock	0.6403	0.3180	Rock	178	100	278	
Mine	0.7422	0.2950	Mine	83	239	322	
			Sum	261	339	600	

# Implementing and assessing another learning algorithm

We want to apply the same experimentation framework on the PLS Regression (C-PLS from the SPV LEARNING tab). We add the component into the diagram.

File Diagram Component	window help				
) 📽 🖪   🎎					
.[	Default title			Cross-validation 1	
🖃 🏢 Dataset (sonar.xls)				Parameters	
🖃 🔛 Define status 1			ross-validatior	n parameters	
Supervised Le	earning 1 (Binary logistic re	gression) F	olds	10	
Supervised La	dation 1 explor 2 (C-DLS)	1	rials	з	
	saming 2 (CPL3)		r		
<b>\</b>				Results	
$\mathbf{X}$		c	V error rate		
$\mathbf{n}$		c	V error rate Range		
		c A	V error rate Range MN 0.3000		1
			Verror rate Range MN 0.3000		
Cata visualization	Statistics	Components Nonparamet	V error rate Range NIN 0.3000	Instance selection	
Data visualization	Statistics	Components Nonparametr	Verror rate Range MN 0.3000 ie statistics	Instance selection	n
Data visualization Feature construction	Statistics Feature selection Clustering	Components Nonparamet Regre	Verror rate Range MIN 0.3000 tic statistics ssion	Instance selection Factorial analysis Meta-sny learning	) )
Data visualization Feature construction PLS	Statistics Feature selection Clustering Scoring	Components Nonparametri Regree Spv le	Verror rate Range NIN 0.3000 tie statistics ssion arning	Instance selection Factorial analysis Meta-spy learning	
Data visualization Feature construction PLS Spy learning assessment	Statistics Feature selection Clustering Scoring	Components Nonparametr Spv le Assoc	Verror rate Range NIN 0.3000 ric statistics ssion arning iation	Instance selection Factorial analysis Meta-spy learning	A Im
Data visualization Feature construction PLS Spy learning assessment	Statistics Feature selection Clustering Scoring MacPPLS	Components Nonparametr Regre Spv le Assoc	Verror rate Range MIN 0.3000 tic statistics ssion arning ation	Instance selection Factorial analysis Meta-spy learning	

The resubstitution error rate estimate is 14.42%.

		Superv	riseu L	carning z (v-		
			Para	meters		
1	C-PLS pa	rameters				
Att. tra	ansforma	tion Standard	ize			
# axis			5			
			Re	esutts		
<b>C</b>					_	
Clas	Sifi(	er perf	orr	nance	<b>S</b> 0.1442	>
Clas	SSIFI( Error ( Nues pre	er perf rate ediction	orr		S D. 1442 Ision matrix	>
Clas <sub>Va</sub>	Error I Lues pre Recall	er perf rate ediction 1-Precision	orr	nance: Confu	S D. 1442 Ision matrix Mine	Sum
Clas Va Value Rock	Error o Lues pro Recall 0.9072	er perf rate ediction 1-Precision 0.1927	FOR F	Confu Rock	S D. 1442 sion matrix Mine 9	<b>Sum</b> 97
Value Rock Mine	Error of lues pro Recall 0.9072 0.8108	er perf rate ediction 1-Precision 0.1927 0.0909	FOR F Rock Mine	Confu Rock 88	5 0.1442 sion matrix Mine 9 90	<b>Sum</b> 97 111

#### Tanagra

Again, we want to implement the cross-validation with the same settings. If the copy paste feature does not exist, we must insert the component and specify the appropriate settings. It is rather tedious. With the copy paste functionality, we can duplicate the treatment(s) by copying the component from the diagram.

In order to duplicate the existing CROSS-VALIDATION 1 component, we must select it. Then, by using the drag and drop principle, we copy the component on the SPV LEARNING 1 (C-PLS) treatment. The operation must be achieved with the mouse. None shortcut or menu allows to make this.



The new component is automatically numbered (CROSS-VALIDATION 2). By launching the cross-validation, we obtain a reliable estimation of the error rate: 25.67%.

We note that the settings are well transmitted to the duplicated component. We have 3 TRIALS of 10 FOLDS cross validation.

### Tanagra

				Cross-val	idation 2		
				Param	eters 👘		
Cross	-validatio	n parameters	רו				
Folds		10	<u>۲</u>				
Trials		3					
			-				
				Res	utts		
Verı	ror rate						
Ri	ange						
MIN	0.2300						
MAX	0.2900						
Trial	Err rate						
1	0.2900						
2	0.2500						
3	0.2300						
)vera	ll cross-v	alidation erro	r rate				
	Error	rate		0	.2567		
١	alues pro	ediction		Confus	ion matrix		
∀alu	e Recall	1-Precision		Rock	Mine	Sum	
Roci	<b>k</b> 0.7806	0.3000	Rock	217	61	278	
Mine	e 0.7112	0.2103	Mine	93	229	322	
			Sum	310	290	600	

# Assessing many supervised learning algorithm

By repeating these operations, we fill out the diagram in the following way.

🖃 🏢 Dataset (sonar.xls)
🖮 🏰 Define status 1
🖃 🕩 Supervised Learning 1 (Binary logistic regression)
Cross-validation 1
🚍 🕩 Supervised Learning 2 (C-PLS)
Cross-validation 2
🚍 🕩 Supervised Learning 3 (C-SVC)
Cross-validation 3
🖃 🕩 Supervised Learning 4 (Linear discriminant analysis)
Cross-validation 4
🖮 🕨 Supervised Learning 5 (K-NN)
Cross-validation 5

We can then obtain a table which displays the estimated error rate according to the assessed supervised learning method.

Method	Resubstitution error rate estimate (%)	Cross validation error rate estimate (%)
Logistic regression	5.29	30.5
PLS Regression (C-PLS)	14.42	25.67
Linear SVM (C-SVC)	12.02	25.33
Linear Discriminant Analysis	10.10	23.50
K-Nearest Neighbor	9.62	14.17

Table	1	-	Error	rate	estimate	according	the	learning	algorithm
-------	---	---	-------	------	----------	-----------	-----	----------	-----------

In this framework, the K-NN algorithm seems the most accurate despite the unfavorable ratio between the number of predictive variables (60) and the number of instances (208).

### Dimensionality reduction: duplicating a branch of the diagram

Because the number of descriptors is high in relation to the number of observations, a dimensionality reduction seems an appropriate strategy. The goal is to apply a mapping of the instances to a new representation space with fewer dimensions; the loss of information must be as weak as possible.

We proceed in the following way: (1) we perform a PCA (Principal Component Analysis); (2) we use the relevant  $^4$  factors as INPUT variables for the K-NN algorithm.

We insert a DEFINE STATUS component into our diagram. We set all the continuous variables (ATTRIBUTE\_01 to ATTRIBUTE\_60) as INPUT.

<sup>&</sup>lt;sup>4</sup> It is an open problem, especially in the context of the dimensionality reduction for subsequent supervised learning algorithm. By varying the number of factors to retain, we can certainly change the accuracy of the classifiers.

### Tanagra

TANAGRA 1.4.7 - [Cross-validation 5]		
File Diagram Component Window Help	-	E ×
Default title	Concernalidation 5	~
Dataset (sonar.xls)	Cross-vauadon 3	
😑 🎎 Define status 1	Parameters	
🚊 💽 Supervised Learning 1 (Binary logistic regress	on) Define attribute statuses	
Cross-validation 1	Derme altribute statuses	
Supervised Learning 2 (C-PLS)	Parameters	
Cross-validation 2		
Supervised Learning 3 (C-SVC)	Attributes : Target Input Illustrative	1
Cross-validation 3		
Supervised Learning 4 (Linear discriminant an	attribute_51 attribute_2	
Cross-validation 4	Cattribute_52 attribute_3	
Cross-unlidation 5	Cattribute_53 attribute_4	
Perfine status 2	C attribute_55	
Set Dennie Status 2	C attribute_56	
	attribute_57 attribute_9	
	C attribute_58 attribute_10	
	C attribute_60	~
	Class attribute_13	
Data visualization Statistics No		
Feature selection Regression	Pactor Viear selection	
Spy learning Meta-spy learning Sc	v learnin	
	OK Cancel Help	_
Correlation scatterplot	<u>Manager and a second s</u>	

We add the PRINCIPAL COMPONENT ANALYSIS component into the diagram. We use the default settings i.e. the component generates the 10 best factors.

TANAGRA 1.4.7 - [Defi	ne status 2]								
Tile Diagram Component	Window Help								- 8 ×
D 📽 🖬   🎎									
	Default title					Def	ine status 2		<u> </u>
🖃 🏢 Dataset (sonar.xls)						P2	rameters		
🖨 🙀 Define status 1		Target : 0							
🖶 🕩 Supervised Learning 1 (Binary logistic regression)				Input : 60					
Cross-val	Cross-validation 1								
😑 🕩 Supervised L	earning 2 (C-PLS)								
Cross-val	idation 2						Results		
😑 💽 Supervised L	earning 3 (C-SVC).			Attribute	Target	Input	Illustrative		
Cross-val	idation 3		at	ttribute_1	-	yes	-		
😑 💽 Supervised Learning 4 (Linear discriminant analysis)			at	ttribute_2	-	yes	-		
🔤 🖓 Cross-val	idation 4		at	ttribute_3	-	yes	-		
🖨 💽 Supervised Learning 5 (K-NN)			at	ttribute_4	-	yes	-		
2 Cross-validation 5			at	ttribute 5	-	ves	-		
🖻 🏰 Define status 2			at	ttribute 6	-	ves	-		
			a	ttribute 7	-	ves			
<b>∧</b>			at	ttribute 8		ves	_		~
		Compon	onto						
Data visualization	Statistics	Nonparametric	statis	stics	Instan	ce seli	ection		
Feature construction	Feature selection	Regression			Factorial analysis				
PLS	Clustering	Sov learning			Meta-spy learning				
Spv learning assessment	Scoring	Association					0		
Construction Analysis - Northeast Construction									
😰 Canonical Discriminant Analysis 📉 Nirwoo									
ter makepie con espondance makiste tester intelpat component makister									
									.::

We click on the VIEW menu. The report shows that the 10 first factors contain 72% of the available information.

		Pr	incipal Component Analysis 1	
			Parameters	
ene	rated axis : 1	0		
			Results	
ia	on valu	26		
- 5	en value	- 3		
\xis	Eigen value	% explained	Histogram	% cumulated
1	11.907436	19.85%		19.85%
2	11.184611	18.64%		38.49%
3	5.068340	8.45%		46.93%
4	3.295124	5.49%		52.43%
5	2.965207	4.94%		57.37%
6	2,409907	4.02%		61.38%
7	2.003368	3.34%		64.72%
8	1.726094	2.88%		67.60%
9	1.497870	2.50%		70.10%
10	1.476919	2.46%	<b>.</b>	72.56%
11	1.251381	2.09%		74.64%
12	1.199677	2.00%		76.64%

We want to launch the learning process on the factors generated by the PCA.

We can then specify the processing sequence above i.e. "learning algorithm" + "cross validation error rate estimate" for each learning method to evaluate. The settings must be the same one. Of course, defining the whole sequence manually is very tedious. Here, by using the mouse drag and drop copy paste principle, we can duplicate all the treatments.

So, we click on the DEFINE STATUS 1 component into the diagram. We drag it with the mouse on the PRINCIPAL COMPONENT ANALYSIS 1. The sub tree below DEFINE STATUS 1 is duplicated. The components are automatically numbered.

We note that Tanagra performs a "real" cross validation i.e. at each learning step process during the cross validation, all the sequence is launched (PCA + SUPERVISED LEARNING METHOD).



Into the DEFINE STATUS 3 component, the 10 factors as INPUT variables; the TARGET attribute is CLASS.

💯 TANAGRA 1.4.7 - [Def	ine status 2]				
File Diagram Componen	t Window Help		- 🖻 🗙		
D 📽 🖬 👫					
	Default title	Define attribute statuses	~		
Supervised  Supervised Super	Learning 4 (Linear dis lidation 4 Learning 5 (K-NN) lidation 5 omponent Anatysis 1 tatus 3 arvised Learning 6 (Bin Cross-validation 6 arvised Learning 7 (C- Cross-validation 7 ervised Learning 8 (C- Cross-validation 8	Parameters         Attribute_s:         C attribute_60         C class         C PCA_1_Axis_1         C PCA_1_Axis_2         C PCA_1_Axis_3         C PCA_1_Axis_4         PCA_1_Axis_6         PCA_1_Axis_7         PCA_1_Axis_8         C PCA_1_Axis_7         PCA_1_Axis_8         PCA_1_Axis_9         PCA_1_Axis_10			
Data visualization	Statistics	E E Clear selection			
PLS Spv learning assessment	Clustering Scoring	OK Cancel Help			
Canonical Discriminant Analysis 🖾 Multiple Correspondance Analysis 🖄 NIPALS 🖄 Principal Component					
<			>		

Now, we must launch the whole diagram. The easiest way is to click on the DIAGRAM / EXECUTE menu.

🖞 TANAGRA 1.4.7 - [Define status 2]					
🝸 File Diagram Component Window Help 🛛 🛛 🗖					- 8 ×
Copy as image	Default title ng 4 (Linear discriminar t n 4	Defi Pa Target : 0	ne status 2 rameters		
Supervised I	Learning 5 (K-NN) lidation 5		Input : 60   Illustrative : (	0	
🖃 🔁 Define status 2					_
🛓 🔀 Principal Co	mponent Analysis 1			Results	
🚊 💱 Define s	tatus 3		Attribute	Target Input	Illustr.
🖨 🕑 Supe	rvised Learning 6 (Binary logi	stic regression)	attribute_1	- yes	-
		attribute_2	- yes	-	
🖨 🕑 Supervised Learning 7 (C-PLS)					<u> </u>
	C	omponents	1		
Data visualization	Statistics	Nonparametric statistics	Instance	e selection	
Feature construction	Feature selection	Regression	Factori	ial analysis	
PLS	Clustering	Spv learning	Meta-sp	ov learning	
Spv learning assessment	Scoring	Association			
🗹 Canonical Discriminant Analysis 🛛 🕅 Multiple Correspondance Analysis 🔀 NIPALS 🕅 🔀 P					Ì∑ P
					>
					:

The Table 2 outlines the performances. We note that all the linear classifiers have benefit from the form of regularization incorporated by the PCA. The result associated to the K-NN seems state that the number of retained factors is important. We can fine tuning this parameter.

Table 2 - Error rate accordi	ng the learning method
------------------------------	------------------------

Method	Resubstitution error rate	Cross validation error rate	
	(%)	(%)	
Logistic Regression	16.35	20.17	
PLS Regression(C-PLS)	16.83	21.67	
Linear SVM (C-SVC)	18.27	20.83	
Discriminant Analysis	15.87	21.50	
K-Nearest Neighbor	7.69	16.33	

We note above all that the copy paste feature is very useful in our context.