Subject

With the 1.4.8 version, we can save a part of the stream diagram. The goal is to perform some succession of analysis on several files.

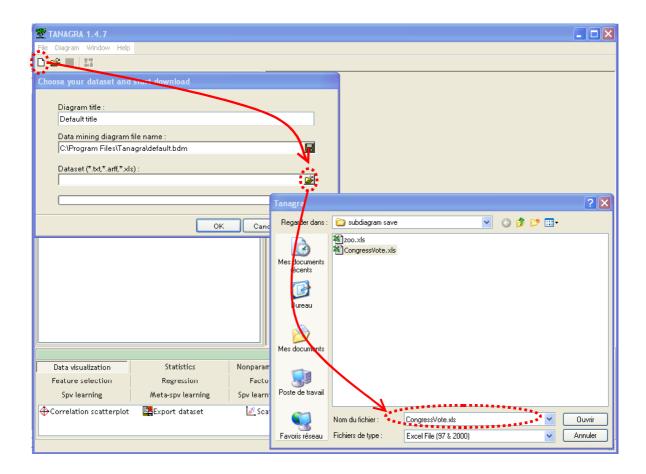
Dataset

We use CONGRESSVOTE.XLS and ZOO.XLS. We want to predict a class attribute from discrete descriptors with or without feature selection. We use the cross-validation in order to compare the error rate.

Saving a sub-diagram

Dataset importation

We click on FILE/NEW in order to create a diagram and import the CONGRESSVOTE.XLS dataset.



Defining class and predictive attributes

We add the DEFINE STATUS component in the diagram: CLASS is the TARGET attribute; the others are INPUT attributes.

Define attribute statuses	
Parameters Attributes : Preligious-groups-in-sch anti-satellite-test-ban aid-to-nicaraguan-contr mx-missile immigration synfuels-corporation-cu education-spending superfund-right-to-sue crime duty-free-exports export-administration-a Class	Target Input Illustrative handicapped-infants Imput Imput water-project-cost-sharin adoption-of-the-budget-re physician-fee-freeze el-salvador-aid religious-groups-in-schoo anti-satellite-test-ban aid-to-nicaraguan-contras mx-missile immigration synfuels-corporation-cutb education-spending superfund-right-to-sue
	Clear selection
	OK Cancel Help

Learning algorithm and performance evaluation

From the 1.4.7 version, we can add directly a supervised learning algorithm in the diagram. TANAGRA inserts automatically the META SPV LEARNING component with implements one instance of this algorithm.

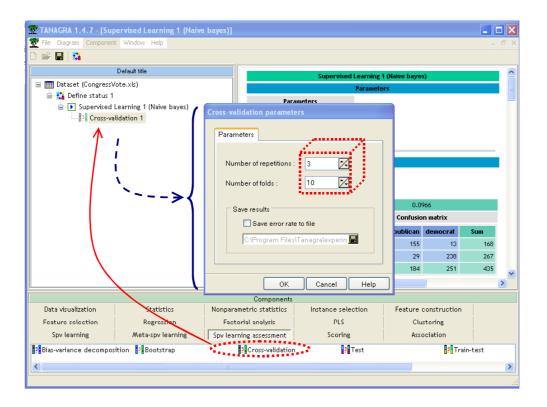
Caution: If you want use aggregation strategy such as BAGGING or BOOSTING, you must follow the old procedure: add in the first step the META SPV component (e.g. BAGGING) and embed in this component the supervised learning algorithm (e.g. NAÏVE BAYES CLASSIFIER).

We insert the NAÏVE BAYES algorithm in the diagram.

File Diagram Component	vised Learning 1 (Naive	e bayes)]						
	Hildoff Holp							
De	efault title		6					
🗉 🎹 Dataset (CongressVot	e.xls)		Supervis	ed Learning 1	-	s)		
🛓 🎦 📴 Define status 1			D	Paramete	rs			
Supervised Lea	arning 1 (Naive bayes)	Use laplaciar	Parameters	4				
	٨	Lambda for l		000				
			ional probabilities	1				
		Show condici	onal probabilities					
				Results				
		Classi	fier perfo	rmance	25			
						()		
			Ennon noto		0.00	744		
			Error rate es prediction	L	0.04 Confusio			l
		Valu	es prediction		Confusio	n matrix	Sum	
		Valu Value	es prediction Recall 1-Precision		Confusio republican	n matrix democrat	Sum	
		Valu Value republican	es prediction Recall 1-Precision 0.9226 0.1576	republican	Confusio republican 155	n matrix democrat 13	168	
		Valu Value	es prediction Recall 1-Precision 0.9226 0.1576		Confusio republican	n matrix democrat		
		Valu Value republican	es prediction Recall 1-Precision 0.9226 0.1576	republican democrat	Confusio republican 155 29	democrat 13 238	168 267	>
		Value republican democrat	Recall 1-Precision 0.9226 0.1576 0.8914 0.0518	republican democrat	Confusio republican 155 29	democrat 13 238	168 267	>
Data visualization	Statistics	Value republican democrat	Recall 1-Precision 0.9226 0.1576 0.8914 0.0518	republican democrat Sum	Confusio republican 155 29 184	democrat 13 238	168 267 435	>
Data visualization Feature selection	Statistics Regression	Value republican democrat	Recall 1-Precision 0.9226 0.1576 0.8914 0.0518	republican democrat Sum	Confusio republican 155 29 184 Feature	democrat 13 238 251	168 267 435	>
		Value republican democrat Compone Nonparametric statistic	es prediction Recall 1-Precision 0.9226 0.1576 0.8914 0.0518 units Instance s PLS	election	Confusio republican 155 29 184 Feature Clu	democrat 13 238 251	168 267 435	>
Feature selection	Regression Meta-spv learning	Value republican democrat Compone Nonparametric statistic Factorial analysis Spv learning assessmen	es prediction Recall 1-Precision 0.9226 0.1576 0.8914 0.0518 ints statements Instance s PLS t Score	republican democrat Sum	Confusio republican 155 29 184 Feature Clu Asso	democrat 13 238 251 construction stering	168 267 435	>
Feature selection Spv learning	Regression	Value republican democrat Compone Nonparametric statistic Factorial analysis Spv learning assessmen	es prediction Recall 1-Precision 0.9226 0.1576 0.8914 0.0518 units Instance s PLS	election	Confusio republican 155 29 184 Feature Clu Asso	democrat 13 238 251 construction stering	168 267 435	>

Resubstitution error rate is 9.66%.

In order to obtain an honest estimation of the "true" error rate, we add a cross-validation component in the diagram. We use a repeated (3 times) 10-cross validation.



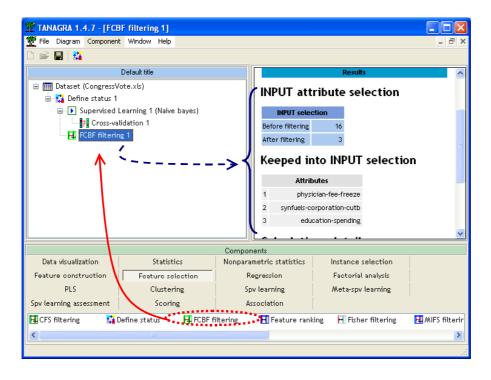
R.R.

The cross-validation error rate is 10.08%.

	Cross-validation 1					
	Parameters					
Cross	s-validatio	on parameters				
Folds		10				
Trials		3				
			Results			
CV er	ror rate					
R	ange					
MIN	0.0977					
MAX	0.1047					
Trial	Err rate					
1	0.1047					
2	0.1000					
3	0.0977					
Overa	Overall cross-validation error rate					
	Err	or rate	0.1008			
	Values	prediction	Confusion matrix			

Feature selection

We want to perform a feature selection before the learning phase. We expect that selecting the relevant attributes improves the classifier performance. We insert in the diagram the FCBF (Liu et al.) feature selection component.



We note that only 3 descriptors among the 16 ones are selected.

In order to evaluate the efficiency of this feature selection, we add again in the diagram the naïve bayes classifier and the cross-validation error rate evaluation. Instead of adding the components manually, we can copy the corresponding sub-diagram (1.4.7 version and higher).

To do that, we select the "SUPERVISED LEARNING 1 (NAÏVE BAYES)" node in the diagram and drag this one on the "FCBF filtering 1" node.

💇 TANAGRA 1.4.7 - [FCE	3F filtering 1]			
Tile Diagram Componen	t Window Help			- 8 ×
🗅 📽 🔚 🔤				
	Default title		Results	^
□ <mark>P1</mark> Cross-va □ P1 FCBF filterir □ ► Supervis	Learning 1 (Naive bayes) 🖊 lidation 1	NPUT select Before filtering After filtering Keeped in Attrit 1 physi	16 3 nto INPUT selectio	n
			cation-spending	v
		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		
🔣 CFS filtering 🛛 👪 I	Define status 🛛 🛃 FCBF f	iltering 🛛 🔣 Feature rank	king 🛛 🕂 Fisher filtering	🔣 MIFS filterir
<				>
				.::

We click on the VIEW menu of the "CROSS VALIDATION 2"; the error rate is 5.5%.

	Cross-validation 2			
			Parameters	
Cross	-validatio	on parameters		
Folds		10		
Trials		3		
			Results	
CV er	ror rate			
R	ange			
MIN	0.0465			
MAX	0.0628			
Trial	Err rate			
1	0.0465			
2	0.0628			
3	0.0558			
Overa	ll cross-v	alidation error	r rate	
	Err	or rate	0.0550	

The FCBF feature selection improves significantly the naïve bayes performances on the VOTE dataset.

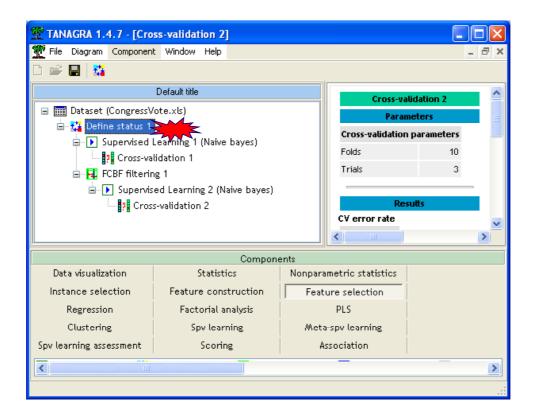
The same analysis on another dataset

We want to evaluate this framework (FCBF feature selection + naïve bayes classifier) on the ZOO dataset. To do that, we must define the same diagram on this dataset, in order to compare the performance of the classifier with or without feature selection.

From the 1.4.8 version, we can save a part of the diagram in a file (SDM file extension) and insert this one in another diagram. We follow three steps: save the sub-diagram from the selected node; open or create a new diagram; insert the saved sub-diagram under the selected node in the new diagram.

Saving the sub-diagram

We select the "DEFINE STATUS 1".



We click on the DIAGRAM / SAVE SUBDIAGRAM item of the main menu.

TANAGRA 1.4.7 - [Cros:				-	. a x
] 💕 💡 Execute					
Copy as image	title		Cross-validation 2		- ^
🕞 🥅 🥙 Create report			Parameters		
🖻 🔠 Delete component		Cross-validation	parameters		
🗃 Load subdiagram	g 1 (Naive bayes)	Folds	10		- and
🔚 Save subdiagram		Trials	3		
E Pag roor nicering	d Learning 2 (Naive bayes)	-			-
	validation 2		Results		
		CV error rate			~
		<	Ξ.		>
		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			
🔣 CFS filtering 🛛 🙀 De	efine status 🛛 🔂 FCBF	filtering 🛛 🖪 Feature rank	ing 🛛 🕂 Fisher filtering	🔛 MIFS filtering	E
					>

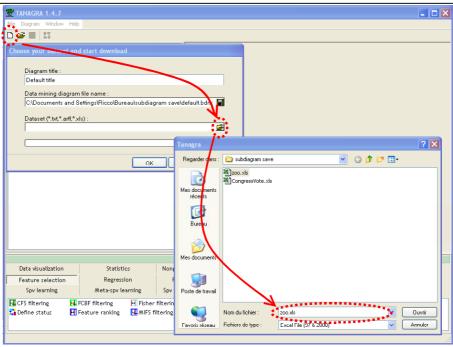
A dialog box appears, we set the sub-diagram file name.

Save a subdiag	ram				? 🛛
Enregistrer dans :	🛅 subdiagram save		G 🖸	🖻 🛄 •	
Mes documents récents Bureau					
) Mes documents					
Poste de travail				7	
	Nom du fichier : 🔶	feature sel for naive bayes class	sifier 📕	~	Enregistrer
Favoris réseau	Туре:	Tanagra subdiagram		*	Annuler

When the sub-diagram is saved, we can close the current diagram (FILE/CLOSE menu).

New diagram and data importation

We select the FILE/NEW menu in order to create a new diagram and import the ZOO.XLS dataset.



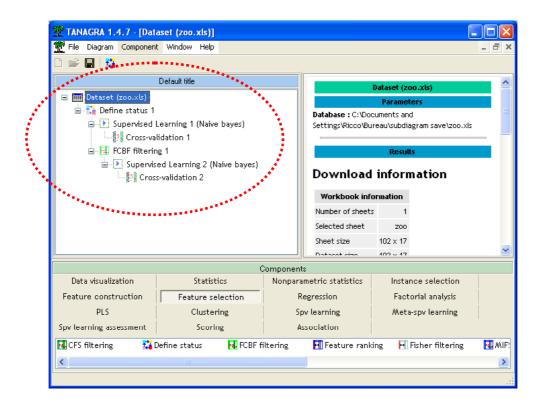
We obtain the following diagram.

File Diagram Component Window Help	🏆 TANAGRA 1.4.7 - [Data	set (zoo.xls)]			
Default title Dataset (zoo.xkb) Parameters Dataset (zoo.xkb) Parameters Dataset (zoo.xkb) Dataset (zoo.xkb) Parameters Dataset (zoo.xkb) Parameters Dataset (zoo.xkb) Results Download information Workbook information Number of sheets 1 Selected sheet 200 Components Data visualization Statistics Instance selection Feature construction Feature selection Regression Factorial analysis PLS Clustering Nonparametric statistics Instance selection Feature construction Feature selection Regression Factorial analysis PLS Clustering Spv learning Meta-spv learning Spv learning assessment <th< td=""><td>Tile Diagram Component</td><td>Window Help</td><td></td><td></td><td>_ @ ×</td></th<>	Tile Diagram Component	Window Help			_ @ ×
Dataset (zoo.xls) Parameters Download information Number of sheets 1 Selected sheet zoo Sheat virue (102 v 47) Vorkbook information Number of sheets 1 Selected sheet zoo Statistics Nonparametric statistics Instance selection Regression Feature construction Feature selection PLS Clustering Spv learning assessment Scoring Association PI Feature ranking PLS Define status PCBF filtering PLS Patient Selection Patient Patient Regression Patient Patient Regression Patient Regression Public	0 📽 🖪 🙀				
Dataset (zoo.xb) Parameters Database : C:Documents and Settings/Ricco/Bureau/subdiagram. save/zoo.xds Results Download information Workbook information Workbook information Number of sheets 1 Selected sheet 200 Charterize 102 v 17 Components Instance selection Feature construction Feature selection PLS Clustering Spv learning assessment Scoring Association Instance filtering CFS filtering Peature ranking Parameterize Instance filtering	1	Default title			
Part attractors and Settings/Ricco/Bureau/subdiagram save/zoo.xls Results Download information Workbook information Workbook information Number of sheets 1 Selected sheet 200 Chain of statistics Nonparametric statistics Instance selection Feature construction Feature selection Regression Factorial analysis PLS Clustering Spv learning Meta-spv learning Spv learning assessment Scoring Association Fisher filtering MIPS filtering	Dataset (zoo.xls)				
savelzoo.xls Results Download information Workbook information Workbook information Number of sheets 1 Selected sheet 200 Chast rize 107 v 17 Data visualization Statistics Pata visualization Statistics Peature construction Feature selection Regression Factorial analysis PLS Clustering Spv learning assessment Scoring Association H Fisher filtering CFDS filtering FCBF filtering			Database • C:3Docur	Concerning and the second second	reaulsubdiagram
Download information Workbook information Number of sheets 1 Selected sheet 200 Chast cira 107 v 17 Data visualization Statistics Pata visualization Statistics PLS Clustering Spv learning assessment Scoring Association H Feature ranking CFS filtering FCBF filtering					
Download information Workbook information Number of sheets 1 Selected sheet 200 Chast cira 107 v 17 Data visualization Statistics Pata visualization Statistics PLS Clustering Spv learning assessment Scoring Association H Feature ranking CFS filtering FCBF filtering			÷		
Workbook information Number of sheets 1 Selected sheet 200 Cheat riva 107 v 17 Data visualization Statistics Feature construction Feature selection Regression Factorial analysis PLS Clustering Spv learning assessment Scoring Association If Seature ranking If CFS filtering If Seature ranking If Sheer filtering If Seature ranking				Resutts	
Number of sheets 1 Selected sheet 200 Chase rize 107 v 17 Data visualization Statistics Peature construction Feature selection Regression Factorial analysis PLS Clustering Spv learning assessment Scoring Association H Feature ranking CFS filtering FCBF filtering			Download i	nformation	
Selected sheet zoo Selected sheet zoo Components Intrance selection Pata visualization Statistics Nonparametric statistics Instance selection Peature construction Feature selection Regression Factorial analysis PLS Clustering Spv learning Meta-spv learning Spv learning assessment Scoring Association Fisher filtering MIFS filtering			Workbook infor	mation	
Cheat rize In7 v 17 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 Fisher filtering MIFS filtering			Number of sheets	1	
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 Image: CFS filtering Fisher filtering Image: CFS filter			Selected sheet	200	
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 Image: CFS filtering Feature ranking			Chaot siza	107 v 17	*
Feature construction Feature selection Regression Factorial analysis PLS Clustering Spv learning Meta-spv learning Spv learning assessment Scoring Association Image: CFS filtering Image: Fisher filtering Im			Components		
PLS Clustering Spv learning Meta-spv learning Spv learning assessment Scoring Association ICFS filtering Item Provide the status Item Filtering	Data visualization	Statistics	Nonparametric statistics	Instance selection	
Spv learning assessment Scoring Association Image: CFS filtering Image: CF	Feature construction	Feature selection	Regression	Factorial analysis	
🖫 CFS filtering 🗱 Define status 🛃 FCBF filtering 🔄 Feature ranking 📑 Fisher filtering 🛃 MIFS filtering	PLS	Clustering	Spv learning	Meta-spv learning	
	Spv learning assessment	Scoring	Association		
	🖪 CFS filtering 🛛 🙀 De	efine status 🛛 🙀 FCBF 🕯	filtering 🛛 🖪 Feature ranking	g 🕂 Fisher filtering	MIFS filtering
	<				>

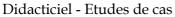
We want to implement the same analysis than the CONGRESSVOTE dataset. We click on the DIAGRAM / LOAD SUBDIAGRAM menu. We select the previous sub-diagram file.

💯 TANAGRA 1.4.7 - [Datas	set (zoo.xls)]			
Tile Diagram Component	Window Help		-	Ξ×
🗋 🚅 😵 Execute				
Copy as image	title		Dataset (200.xls)	~
🔚 🥙 Create report			Parameters	
👘 Delete component		Concernance and the second		=
🗃 Load subdiagram		Load and inser	ert a subdiagram ? 🔀 🖤	
Save subdiagram		Regarder dans :	s: 🔁 subdiagram save 🛛 🕜 🧷 🗗 🛄 🗸 🔤	
		A	feature sel for naive bayes classifier.sdm	
		Mes documents		
		-		
		Bureau		
		Mes documents		~
		-		
Data visualization	Statistics	Poste de travail		
Feature construction	Feature selection	i oste de travali		
PLS	Clustering		Nom du fichier : feature sel for naive bayes classifier.sdm 💌 Ouvrir	
	-	Favoris réseau	Fichiers de type : Tanagra subdiagram 🗸 Annuler	
Spv learning assessment	Scoring			
🖪 CFS filtering 🛛 👪 De	efine status 🛛 🔂 FCI	BF filtering	🔣 Feature ranking 🛛 Fisher filtering 🛛 🔣 MIFS filtering	
<				>

The diagram is supplemented in the following way.



So, we must set the right predictive and class attributes. Of course, we must clear the previous selection before (use the CLEAR SELECTION button).



2 TANAGRA 1.4.7 - [Dat	aset (zoo.xis)]				
 File Diagram Component Tele Dataset (zoo.xk) Define status 1 Supervised FCBF filterin Supervised 	t Window Help Default title Parameters yet Execute	Define attribute statuses Parameters Attributes : I hair feathers eggs milk aquatic predator toothed backbone breathes venomous fins lens E		Input Illustrative	
			ок	Cancel Help	
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	
	FCBF filtering 🛛 Fisher Feature ranking 🔂 MIFS f		:ering 📥 Runs filtering stant		

Then, we can execute the cross-validation when we use all the attributes (VIEW menu on CROSS-VALIDATION 1) and when we use only the selected attributes (VIEW menu of CROSS-VALIDATION 2).

		Cross-validation 1			(Cross-validation 2
		Parameters				Parameters
ross-validation para	meters			Cross-validation p	arameters	
olds	10			Folds	10	
rials	3			Trials	3	\sim
		Results				Results
error rate				CV error rate		
Range			`	Range		
N 0.0600			\mathbf{N}	MIN 0.1000		
X 0.0600				MAX 0.1100		
ial Err rate			1	Trial Err rate		
0.0600				1 0.1100		
0.0600				2 0.1100		
0.0600				3 0.1000		
erall cross-validatio	on error rate		V	Overall cross-valio	ation error rate	
Error rate	в		0.0600	Error	rate	0.

The error rate without feature selection is 6%. When we insert the FCBF feature selection before the learning algorithm, the error rate becomes 10.67%. We note that this feature selection is not efficient on the ZOO dataset.

We note especially that this new functionality makes it possible to transpose very easily a succession of analysis on another dataset.