Target

How to apply classifier on a new dataset?

This functionality surpasses the TANAGRA framework, which intends only to evaluate and compare data mining algorithms. But, users ask it often; in this tutorial we show how to proceed.

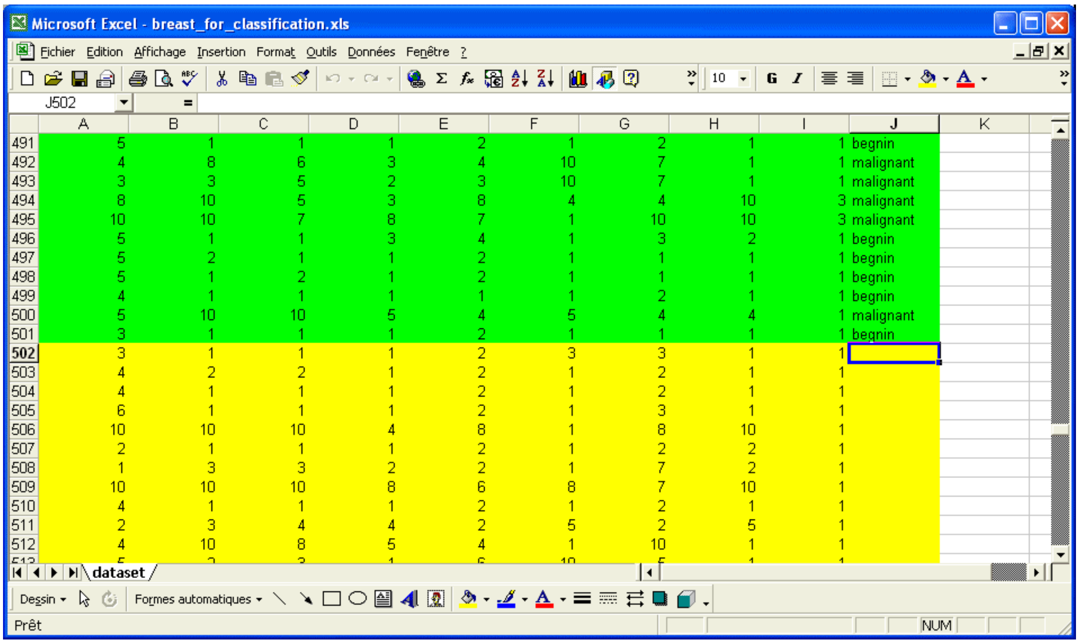
Dataset

Data preparation is a primordial step. Indeed, TANAGRA can handle only one data source. It is not theoretically possible to manipulate two dataset, and therefore apply a classifier on a new dataset. The trick is in the dataset preparation.

We use the BREAST CANCER WINSCONCIN dataset in this tutorial (detect a tumor from cells properties). We subdivide the dataset into 500 examples for the learning phase, and 199 examples for the classification phase.

The dataset is built in several steps:

(1) Join the two dataset in a file, using a spreadsheet for instance, you can set the learning set before the examples to classify.



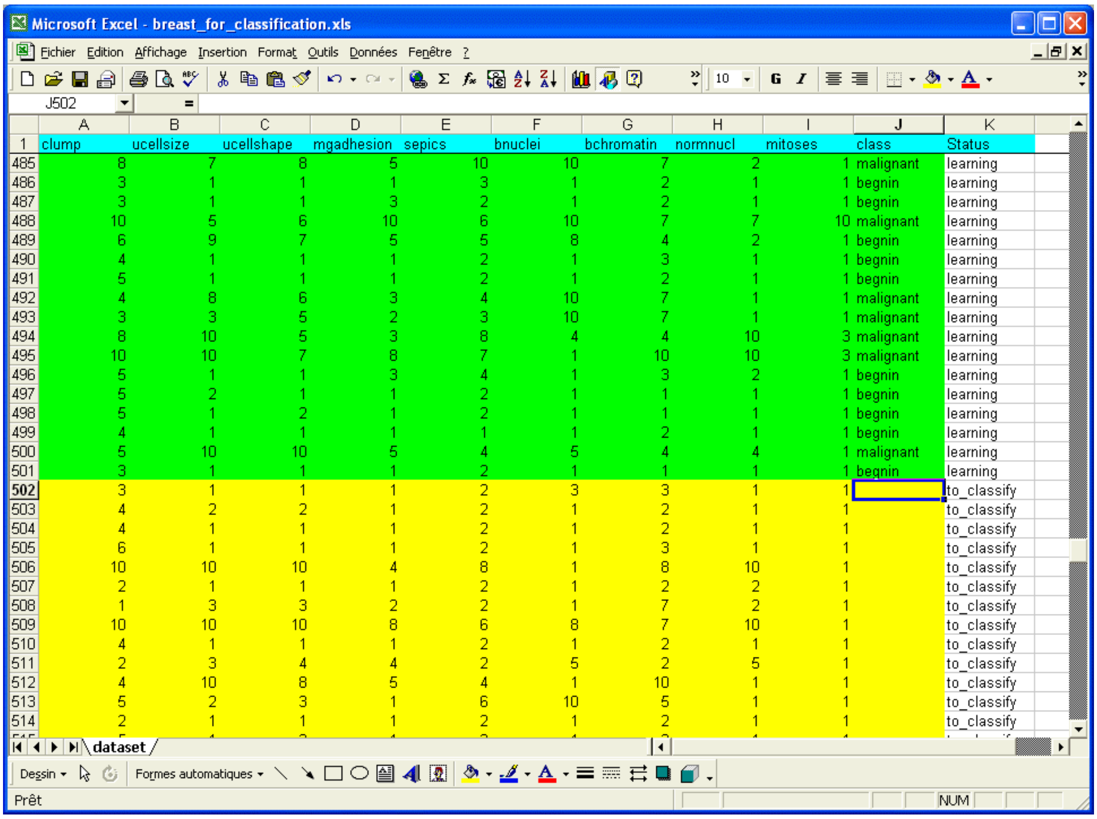
Microsoft Excel - breast_for_classification.xls

J502 =

	A	B	C	D	E	F	G	H	I	J
491	5	1	1	1	2	1	2	1	1	begnin
492	4	8	6	3	4	10	7	1	1	malignant
493	3	3	5	2	3	10	7	1	1	malignant
494	8	10	5	3	8	4	4	10	3	malignant
495	10	10	7	8	7	1	10	10	3	malignant
496	5	1	1	3	4	1	3	2	1	begnin
497	5	2	1	1	2	1	1	1	1	begnin
498	5	1	2	1	2	1	1	1	1	begnin
499	4	1	1	1	1	1	2	1	1	begnin
500	5	10	10	5	4	5	4	4	1	malignant
501	3	1	1	1	2	1	1	1	1	begnin
502	3	1	1	1	2	3	3	1	1	
503	4	2	2	1	2	1	2	1	1	
504	4	1	1	1	2	1	2	1	1	
505	6	1	1	1	2	1	3	1	1	
506	10	10	10	4	8	1	8	10	1	
507	2	1	1	1	2	1	2	2	1	
508	1	3	3	2	2	1	7	2	1	
509	10	10	10	8	6	8	7	10	1	
510	4	1	1	1	2	1	2	1	1	
511	2	3	4	4	2	5	2	5	1	
512	4	10	8	5	4	1	10	1	1	

Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?
10
dataset
Dessin
Formes automatiques
Prêt
NUM

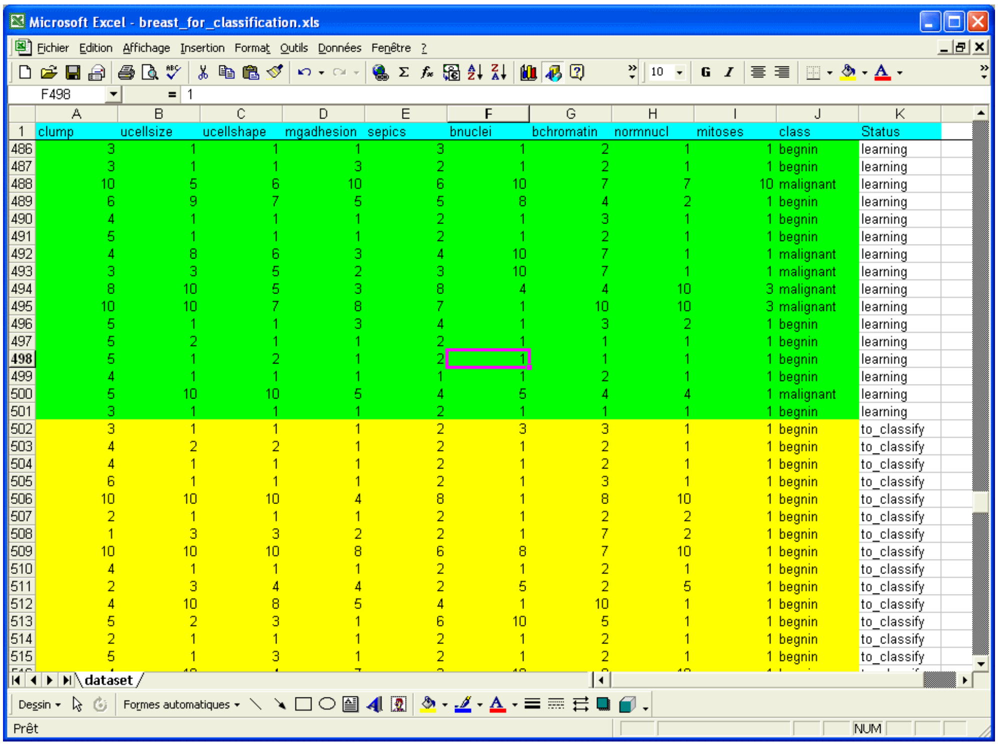
(2) Add a new attribute "STATUS" with two values "learning" and "To_classifiy" which allows us to specify the status of each example.


Microsoft Excel - breast_for_classification.xls
Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?
10
G I
dataset
Dessin
Formes automatiques
Prêt
NUM

J502 =

	A	B	C	D	E	F	G	H	I	J	K
1	clump	ucellsize	ucellshape	mgadhesion	sepics	bnuclei	bchromatin	normnucl	mitoses	class	Status
485	8	7	8	5	10	10	7	2	1	malignant	learning
486	3	1	1	1	3	1	2	1	1	begnin	learning
487	3	1	1	3	2	1	2	1	1	begnin	learning
488	10	5	6	10	6	10	7	7	10	malignant	learning
489	6	9	7	5	5	8	4	2	1	begnin	learning
490	4	1	1	1	2	1	3	1	1	begnin	learning
491	5	1	1	1	2	1	2	1	1	begnin	learning
492	4	8	6	3	4	10	7	1	1	malignant	learning
493	3	3	5	2	3	10	7	1	1	malignant	learning
494	8	10	5	3	8	4	4	10	3	malignant	learning
495	10	10	7	8	7	1	10	10	3	malignant	learning
496	5	1	1	3	4	1	3	2	1	begnin	learning
497	5	2	1	1	2	1	1	1	1	begnin	learning
498	5	1	2	1	2	1	1	1	1	begnin	learning
499	4	1	1	1	1	1	2	1	1	begnin	learning
500	5	10	10	5	4	5	4	4	1	malignant	learning
501	3	1	1	1	2	1	1	1	1	begnin	learning
502	3	1	1	1	2	3	3	1	1		to_classify
503	4	2	2	1	2	1	2	1	1		to_classify
504	4	1	1	1	2	1	2	1	1		to_classify
505	6	1	1	1	2	1	3	1	1		to_classify
506	10	10	10	4	8	1	8	10	1		to_classify
507	2	1	1	1	2	1	2	2	1		to_classify
508	1	3	3	2	2	1	7	2	1		to_classify
509	10	10	10	8	6	8	7	10	1		to_classify
510	4	1	1	1	2	1	2	1	1		to_classify
511	2	3	4	4	2	5	2	5	1		to_classify
512	4	10	8	5	4	1	10	1	1		to_classify
513	5	2	3	1	6	10	5	1	1		to_classify
514	2	1	1	1	2	1	2	1	1		to_classify

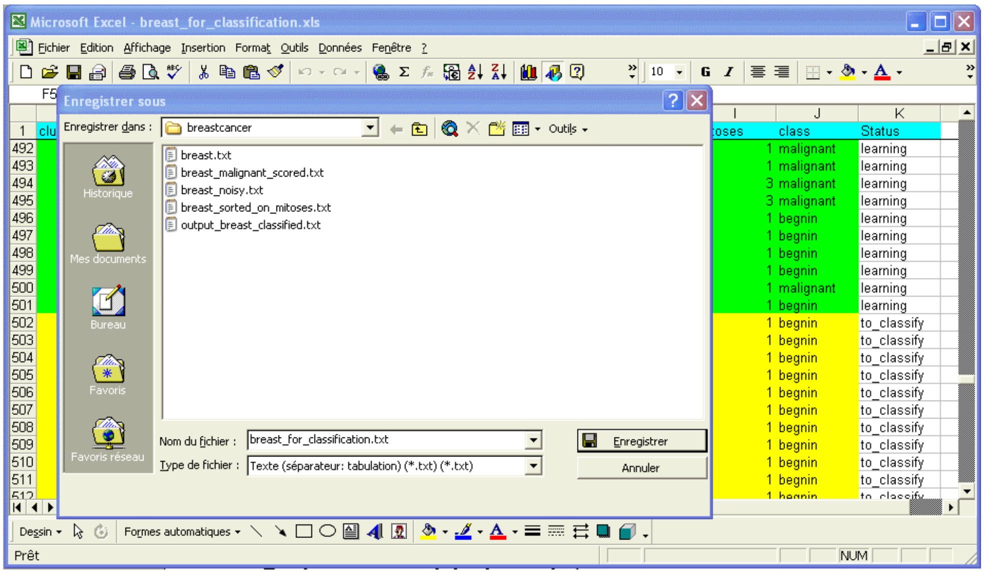
(3) At last, even if it is not very intuitive, you must assign a class value for each example to classify. The reason is that TANAGRA does not handle missing data. You must use one of the existing values. This information will not be used in the following. In our dataset, we set the 199 examples to classify to "begnin".


Microsoft Excel - breast_for_classification.xls
Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?
10
G I
dataset
Dessin
Formes automatiques
Prêt
NUM

F498 = 1

	A	B	C	D	E	F	G	H	I	J	K
1	clump	ucellsize	ucellshape	mgadhesion	sepics	bnuclei	bchromatin	normnucl	mitoses	class	Status
486	3	1	1	1	3	1	2	1	1	begnin	learning
487	3	1	1	3	2	1	2	1	1	begnin	learning
488	10	5	6	10	6	10	7	7	10	malignant	learning
489	6	9	7	5	5	8	4	2	1	begnin	learning
490	4	1	1	1	2	1	3	1	1	begnin	learning
491	5	1	1	1	2	1	2	1	1	begnin	learning
492	4	8	6	3	4	10	7	1	1	malignant	learning
493	3	3	5	2	3	10	7	1	1	malignant	learning
494	8	10	5	3	8	4	4	10	3	malignant	learning
495	10	10	7	8	7	1	10	10	3	malignant	learning
496	5	1	1	3	4	1	3	2	1	begnin	learning
497	5	2	1	1	2	1	1	1	1	begnin	learning
498	5	1	2	1	2	1	1	1	1	begnin	learning
499	4	1	1	1	1	1	2	1	1	begnin	learning
500	5	10	10	5	4	5	4	4	1	malignant	learning
501	3	1	1	1	2	1	1	1	1	begnin	learning
502	3	1	1	1	2	3	3	1	1	begnin	to_classify
503	4	2	2	1	2	1	2	1	1	begnin	to_classify
504	4	1	1	1	2	1	2	1	1	begnin	to_classify
505	6	1	1	1	2	1	3	1	1	begnin	to_classify
506	10	10	10	4	8	1	8	10	1	begnin	to_classify
507	2	1	1	1	2	1	2	2	1	begnin	to_classify
508	1	3	3	2	2	1	7	2	1	begnin	to_classify
509	10	10	10	8	6	8	7	10	1	begnin	to_classify
510	4	1	1	1	2	1	2	1	1	begnin	to_classify
511	2	3	4	4	2	5	2	5	1	begnin	to_classify
512	4	10	8	5	4	1	10	1	1	begnin	to_classify
513	5	2	3	1	6	10	5	1	1	begnin	to_classify
514	2	1	1	1	2	1	2	1	1	begnin	to_classify
515	5	1	3	1	2	1	2	1	1	begnin	to_classify

(4) Export the dataset in a tab separator text file for TANAGRA.



Enregistrer dans : breastcancer

breast.txt
breast_malignant_scored.txt
breast_noisy.txt
breast_sorted_on_mitoses.txt
output_breast_classified.txt

Nom du fichier : breast_for_classification.txt

Type de fichier : Texte (séparateur: tabulation) (*.txt) (*.txt)

	I	J	K
1	oses	class	Status
492	1	malignant	learning
493	1	malignant	learning
494	3	malignant	learning
495	3	malignant	learning
496	1	begnin	learning
497	1	begnin	learning
498	1	begnin	learning
499	1	begnin	learning
500	1	malignant	learning
501	1	begnin	learning
502	1	begnin	to_classify
503	1	begnin	to_classify
504	1	begnin	to_classify
505	1	begnin	to_classify
506	1	begnin	to_classify
507	1	begnin	to_classify
508	1	begnin	to_classify
509	1	begnin	to_classify
510	1	begnin	to_classify
511	1	begnin	to_classify
512	1	begnin	to_classify

Classifier deployment on a new dataset

Dataset importation in TANAGRA

Import the dataset in TANAGRA and define a stream diagram.

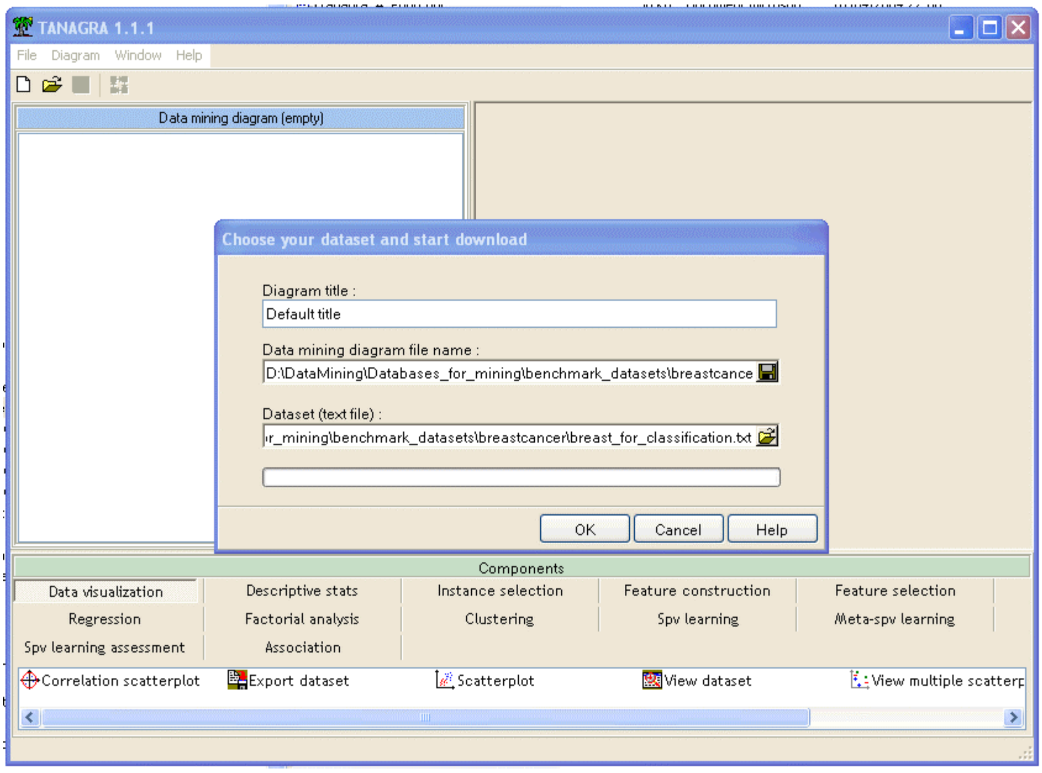


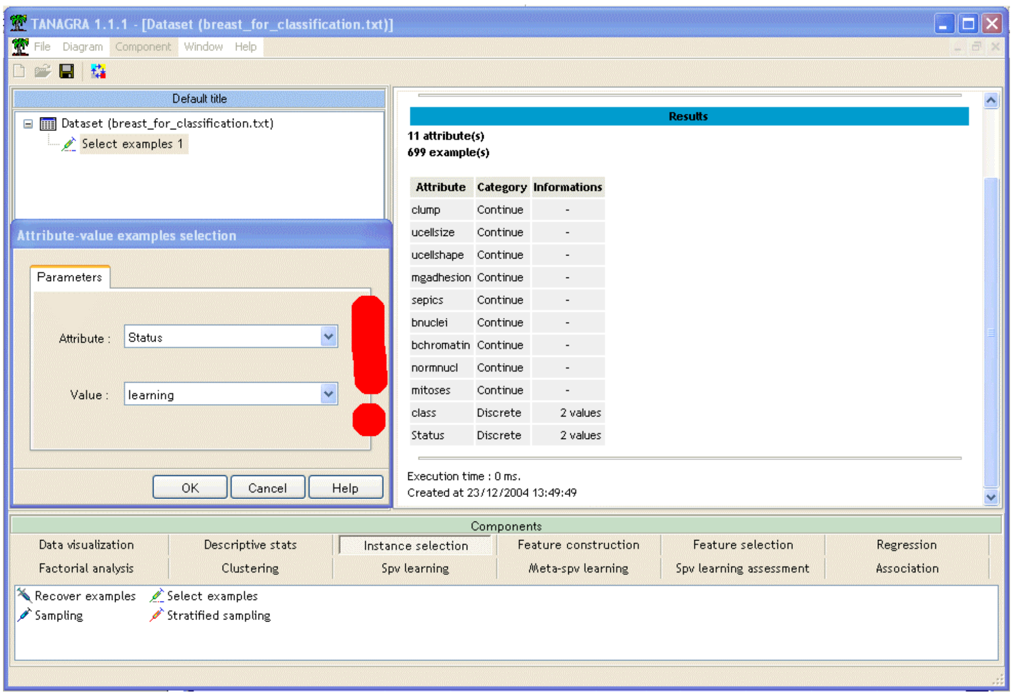
Diagram title : Default title

Data mining diagram file name : D:\DataMining\Databases_for_mining\benchmark_datasets\breastcance

Dataset (text file) : r_mining\benchmark_datasets\breastcancer\breast_for_classification.txt

Select the learning set

Above all, you must select the learning set. Use SELECT EXAMPLES component using the STATUS attribute: we have 500 examples for the learning phase (active examples), and 199 examples for the classification phase (idle examples).



Attribute : Status

Value : learning

Results

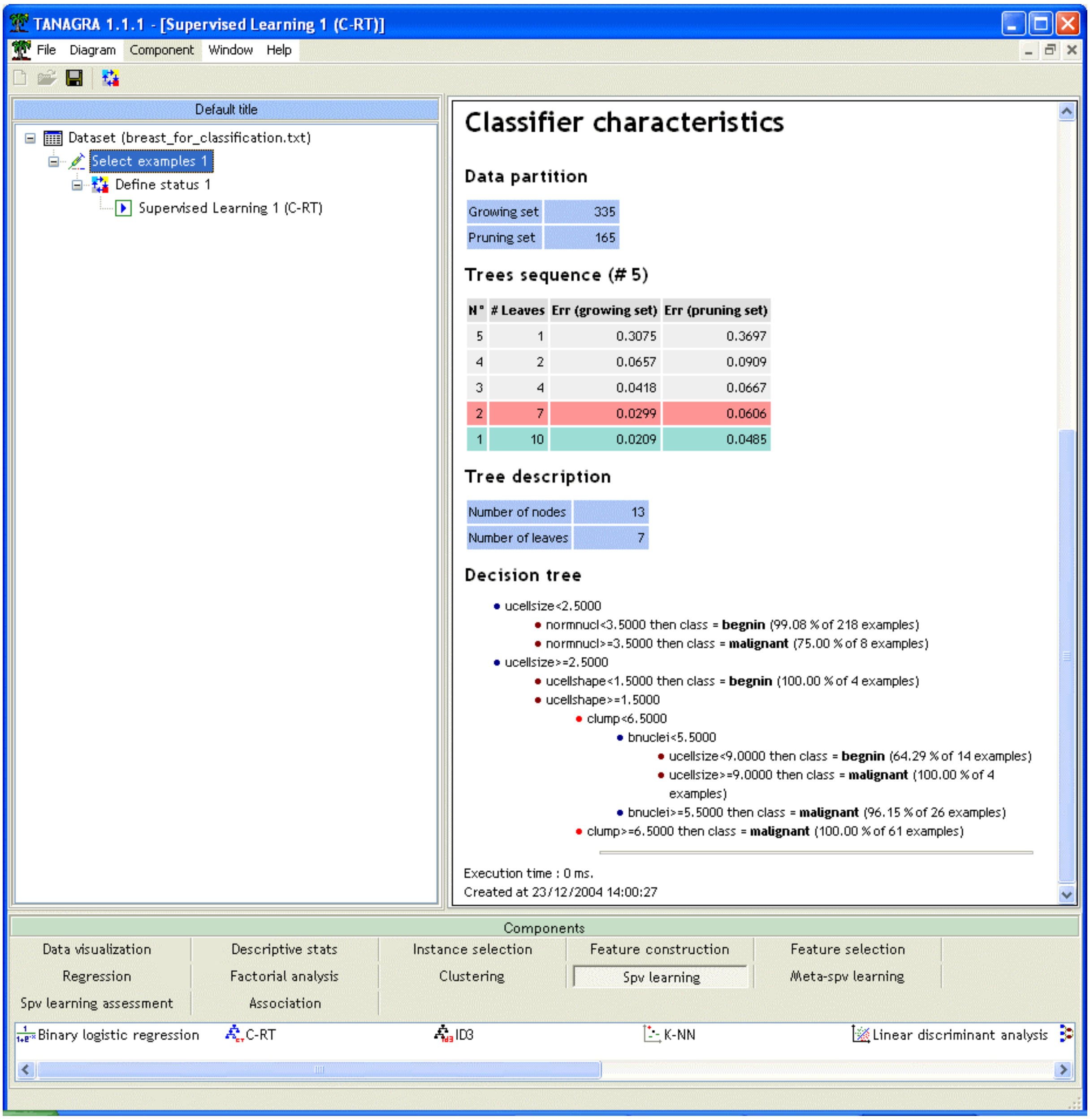
11 attribute(s)
699 example(s)

Attribute	Category	Informations
clump	Continue	-
ucellsize	Continue	-
ucellshape	Continue	-
mgadhesion	Continue	-
sepics	Continue	-
bnuclei	Continue	-
bchromatin	Continue	-
normnucl	Continue	-
mitoses	Continue	-
class	Discrete	2 values
Status	Discrete	2 values

Execution time : 0 ms.
Created at 23/12/2004 13:49:49

Supervised learning

In the next step, we must select examples and define the learning algorithm. We use the Breiman's et al. (1984) famous classification tree algorithm.



TANAGRA 1.1.1 - [Supervised Learning 1 (C-RT)]
File
Diagram
Component
Window
Help
Default title
Dataset (breast_for_classification.txt)
Select examples 1
Define status 1
Supervised Learning 1 (C-RT)

Classifier characteristics

Data partition

Growing set	335
Pruning set	165

Trees sequence (# 5)

N°	# Leaves	Err (growing set)	Err (pruning set)
5	1	0.3075	0.3697
4	2	0.0657	0.0909
3	4	0.0418	0.0667
2	7	0.0299	0.0606
1	10	0.0209	0.0485

Tree description

Number of nodes	13
Number of leaves	7

Decision tree

- ucellsize<2.5000
    - normnucl<3.5000 then class = **begnin** (99.08 % of 218 examples)
    - normnucl>=3.5000 then class = **malignant** (75.00 % of 8 examples)
- ucellsize>=2.5000
    - ucellshape<1.5000 then class = **begnin** (100.00 % of 4 examples)
    - ucellshape>=1.5000
        - clump<6.5000
            - bnuclei<5.5000
                - ucellsize<9.0000 then class = **begnin** (64.29 % of 14 examples)
                - ucellsize>=9.0000 then class = **malignant** (100.00 % of 4 examples)
            - bnuclei>=5.5000 then class = **malignant** (96.15 % of 26 examples)
        - clump>=6.5000 then class = **malignant** (100.00 % of 61 examples)

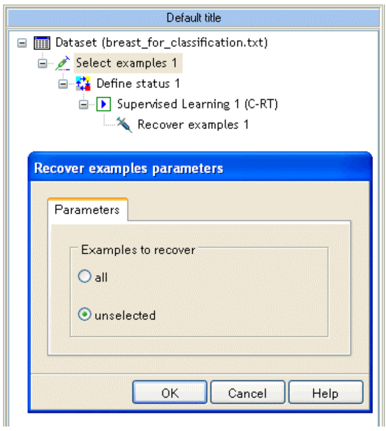
Execution time : 0 ms.
Created at 23/12/2004 14:00:27

TANAGRA adds automatically a new attribute, the predicted values.

Note that if the learning phase is realized on the learning set, the classification is realized on the whole examples, including the idle examples. We exploit this property to classify new dataset or to test the classifier on an external test sample.

View classified examples

We can view the predicted attribute on the examples to classify using a VIEW DATASET component. First, we must set idle examples to active examples with RECOVER EXAMPLES component.



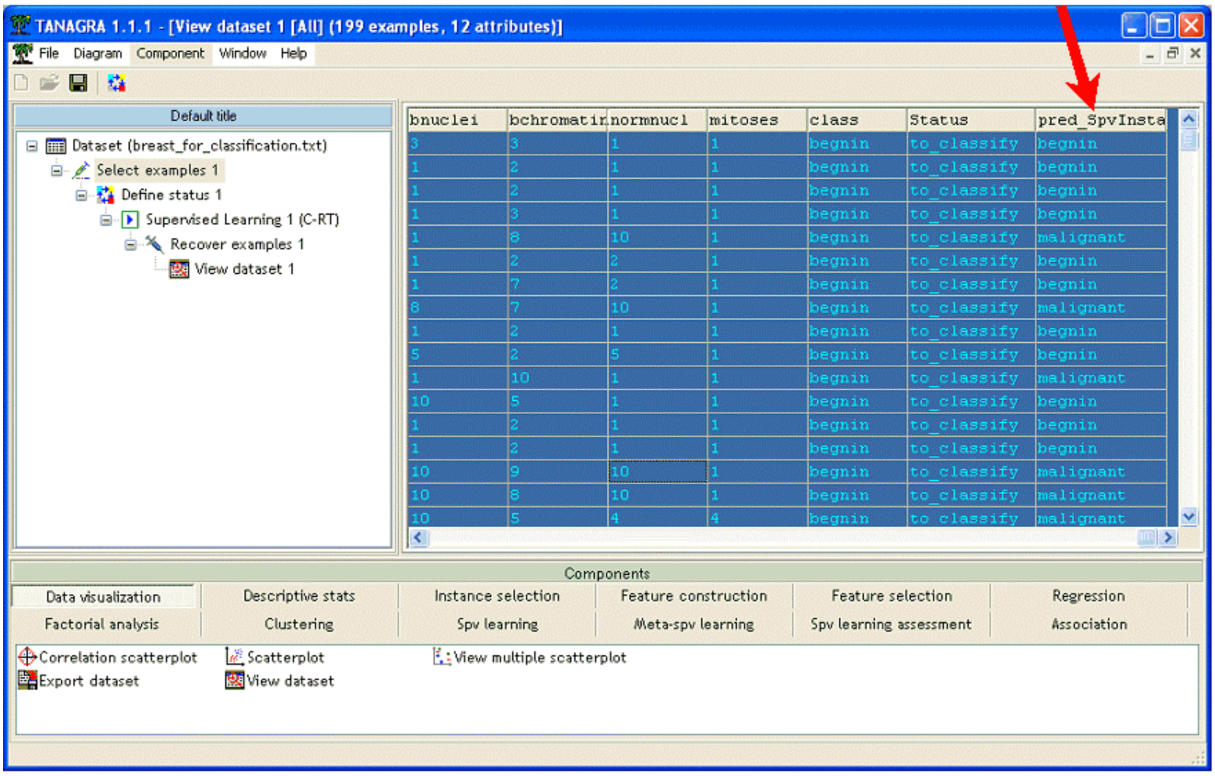
Default title
Dataset (breast_for_classification.txt)
Select examples 1
Define status 1
Supervised Learning 1 (C-RT)
Recover examples 1
Recover examples parameters
Parameters
OK
Cancel
Help

Examples to recover

○ all

● unselected

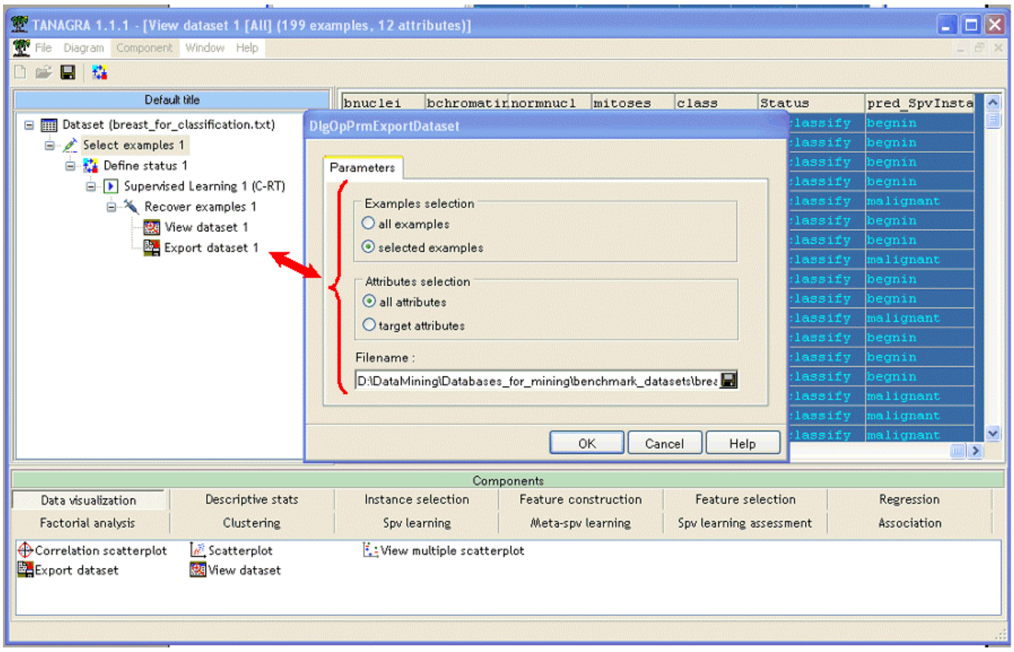
VIEW DATASET



bnuclei	bchromatin	normnucl	mitoses	class	Status	pred_SpvInsta
3	3	1	1	begnin	to_classify	begnin
1	2	1	1	begnin	to_classify	begnin
1	2	1	1	begnin	to_classify	begnin
1	3	1	1	begnin	to_classify	begnin
1	8	10	1	begnin	to_classify	malignant
1	2	2	1	begnin	to_classify	begnin
1	7	2	1	begnin	to_classify	begnin
8	7	10	1	begnin	to_classify	malignant
1	2	1	1	begnin	to_classify	begnin
5	2	5	1	begnin	to_classify	begnin
1	10	1	1	begnin	to_classify	malignant
10	5	1	1	begnin	to_classify	begnin
1	2	1	1	begnin	to_classify	begnin
1	2	1	1	begnin	to_classify	begnin
10	9	10	1	begnin	to_classify	malignant
10	8	10	1	begnin	to_classify	malignant
10	5	4	4	begnin	to_classify	malignant

Export results

The last step is to export the examples with EXPORT DATASET component. We choose to export all attributes on examples to classify.



DlgOpPrmExportDataset

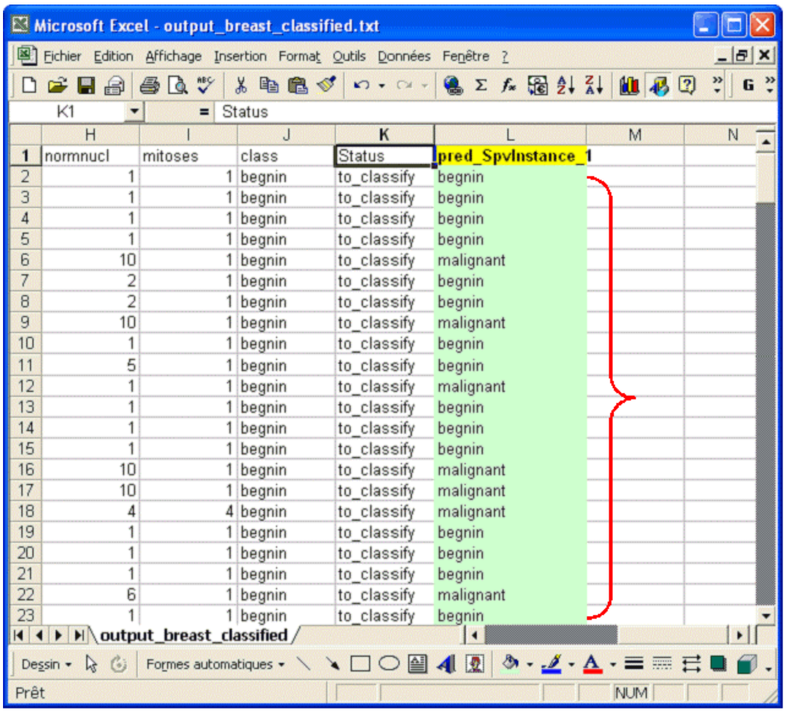
Parameters

Examples selection
○ all examples
● selected examples

Attributes selection
● all attributes
○ target attributes

Filename : D:\DataMining\Databases_for_mining\benchmark_datasets\brea

We can view the dataset in a spreadsheet.



Microsoft Excel - output_breast_classified.txt
Fichier Edition Affichage Insertion Format Outils Données Fenêtre ?
K1 = Status
output_breast_classified
Dessin Formes automatiques
Prêt
NUM

	H	I	J	K	L
1	normnucl	mitoses	class	Status	pred_SpvInstance_1
2	1	1	begnin	to_classify	begnin
3	1	1	begnin	to_classify	begnin
4	1	1	begnin	to_classify	begnin
5	1	1	begnin	to_classify	begnin
6	10	1	begnin	to_classify	malignant
7	2	1	begnin	to_classify	begnin
8	2	1	begnin	to_classify	begnin
9	10	1	begnin	to_classify	malignant
10	1	1	begnin	to_classify	begnin
11	5	1	begnin	to_classify	begnin
12	1	1	begnin	to_classify	malignant
13	1	1	begnin	to_classify	begnin
14	1	1	begnin	to_classify	begnin
15	1	1	begnin	to_classify	begnin
16	10	1	begnin	to_classify	malignant
17	10	1	begnin	to_classify	malignant
18	4	4	begnin	to_classify	malignant
19	1	1	begnin	to_classify	begnin
20	1	1	begnin	to_classify	begnin
21	1	1	begnin	to_classify	begnin
22	6	1	begnin	to_classify	malignant
23	1	1	begnin	to_classify	begnin

Conclusion

We can follow the same way for classifier validation on an external test set. To obtain classification matrix, you can build a contingency table between class attribute and predicted attribute on the idle examples.

I know this method is very complex but it is the only way to apply a classifier on a new dataset.

For classifier validation with subset of examples which not used for the learning phase, you can also use SAMPLING or ASSESSMENT components.