# Subject

TANAGRA, ORANGE and WEKA: Comparison of learning algorithms using a predefined learning and test set.

Very often, we use the accuracy to compare the performances of the algorithms. We then select the method that is the most accurate. So that the comparison is rigorous, it is necessary that we use the same dataset in training and test phase.

We show in this tutorial, how to implement this process in three data mining software: TANAGRA, ORANGE and WEKA. We chose to compare the performances of a SVM (linear kernel), a logistic regression and a decision tree.

## Dataset

We use the BREAST dataset (UCI IRVINE). We have a binary class attribute (benign or malignant tumor), 9 continuous descriptors, and 699 examples.

We have selected 499 examples for the training phase, 200 examples for the test. **We use the same subdivision for our three packages.** 

## Algorithms comparison with ORANGE

When we execute ORANGE, we have the following interface.



#### Data preparation

We divide the whole dataset into two files: BREAST\_TRAIN.TXT for training, BREAST\_TEST.TXT for testing. We set two data access components in the diagram; we parameterize them by activating the OPEN menu.



#### Learning components

We want to compare three learning methods from the CLASSIFY tab, we set them in the diagram.



#### **Evaluation component**

The comparison can be gathered in only one component, the TEST LEARNERS component from the EVALUATE tab. We connect the three learning method to this new component.



We must now specify which are the data to be used for the training. We connect the first data source [FILE] to the TEST LEARNERS. A dialog box appears, it is of primary importance

because it enables us to check that we transmit the training set (DATA). The training phase is automatically started.



In the next step, we connect the second data source [FILE (2)] to the TEST LEARNERS component. ORANGE considers that this second data source is the test set (SEPARATE TEST DATA). We can modify this type of the connection when we double-click on the link; it is not necessary here.



#### Seeing the results

To display the results, we select the OPEN menu of the TEST LEARNERS component.



We check the "Test on test data" option. Various statistics are available; we are interested primarily in the accuracy in our tutorial:

- Classification tree: 93.5% (error rate 6.5%);
- Logistic regression: 95.5%
- Linear SVM<sup>1</sup>: 94.5%.

<sup>&</sup>lt;sup>1</sup> Check that you use really a linear kernel in your diagram (KERNEL – LINEAR).

### Algorithms comparison with WEKA

A dialog box appears when we execute WEKA; we choose the **KNOWLEDGE FLOW** paradigm. We have used the **3.5.1** version.

🔭 Weka GUI Cho 🔚 🗖 🔀	👙 W	eka KnowledgeFlow Enviror	iment				
		DataSources DataSinks Filters	Classifiers Clusterers Ass	ociations Evaluation Visuali	zation		R
Waikato Environment for		DataSources					
Knowledge Analysis		Prosto The second se	D.A.	yes, 8 no, 3	W0010	100010 010011	
Version 3.5.1	9	ARFF	<b>1</b>	csv	ARFF	SERIAL	
		Arff Loader	C45 Loader	Loader	Database Loader	Serialised InstancesLoader	
(c) 1999 - 2005		<					>
University of Walkato	-Know	ledge Flow Layout					
							<b>^</b>
The case of the second s							
All Martin Cas							
A MARCE MARK							
CALLER DE							
and the second							
a lange and the second							
The second second							
THE COMPANY OF THE OWNER							
GUI							
Simple CLI Explorer							~
Experimenter KonsulationElaus	<						>
Experimenter Rhowledgerow	Status	;					
ArtfViewer Log	Welc	ome to the Weka Knowledge Flow					Log

#### **Data preparation**

We have to use two separate dataset with WEKA. We use the ARFF file format; we check carefully that the description of the attributes is the same one.

<pre>@relation breast_train.arff</pre>	<pre>@relation breast_test.arff</pre>
@attribute clump REAL	@attribute clump REAL
@attribute ucellsize REAL	@attribute ucellsize REAL
@attribute ucellshape REAL	@attribute ucellshape REAL
@attribute mgadhesion REAL	@attribute mgadhesion REAL
@attribute sepics REAL	@attribute sepics REAL
@attribute bnuclei REAL	@attribute bnuclei REAL
@attribute bchromatin REAL	@attribute bchromatin REAL
@attribute normnucl REAL	@attribute normnucl REAL
@attribute mitoses REAL	@attribute mitoses REAL
<pre>@attribute class {begnin,malignant}</pre>	<pre>@attribute class {begnin,malignant}</pre>

We set two ARFF LOADER components in the diagram; we select the datasets.



We must now specify the role of these data in the diagram. We use the TRAINING SET MAKER and TEST SET MAKER components (EVALUATE tab). We set the adequate connections.



### Learning methods

We set three learning algorithms components (CLASSIFIERS tab) in the diagram. About SMO, we check that we really use a linear kernel (exponent = 1, not RBF kernel). We connect the three TRAINING SET MAKER, ...



... and the three TEST SET MAKER.



### **Evaluation components**

To compute the accuracy of the classifiers, we set CLASSIFIER PERFORMANCE EVALUATOR component (EVALUATION), one for each learning method. The type of the connection must be BATCH CLASSIFIER.



#### Visualization component

We set the TEXT VIEWER (VISUALIZATION) component in order to display the results. Only one component is necessary, it makes it possible to join together the results in the same window. We use the TEXT connection.



#### Diagram execution

The execution of the diagram is done into two steps: [1] we select the START LOADING menu of the first data source (learning set), the prediction models are computed; [2] we select the START LOADING menu of the second data source, the test set, the accuracy of the models is computed.



When we select the SHOW RESULTS menu of the TEXT VIEWER component, we can see the detailed results for each learning algorithm.

Did	acticiel - Etuc	des de cas	R.R
	👙 Text Viewer		
{	Result list 17:54:07 - Logistic 17:54:07 - SMO 17:54:07 - J48	Text Scheme: Logistic Relation: breast_test.arff	
		Correctly Classified Instances 191 95.5 % Incorrectly Classified Instances 9 4.5 %	
	÷	Kappa statistic 0.0597   Mean absolute error 0.0569   Root mean squared error 0.1898   Relative absolute error 13.1678 %   Root relative squared error 40.8686 %   Total Number of Instances 200   === Detailed Accuracy By Class ===	III
		TP Rate FP Rate Precision Recall F-Measure ROC Area Class 0.956 0.048 0.978 0.956 0.967 0.994 begnin 0.952 0.044 0.909 0.952 0.93 0.994 malignant === Confusion Matrix ===	
		a b < classified as 131 6   a = begnin 3 60   b = malignant	~

We obtain the following accuracy rate:

- Decision tree: 93.5% (error rate 6.5%);
- Logistic regression: 95.5%
- Linear SVM: 95.5%.

We note that SVM and Logistic regression have the same accuracy rate but not the same confusion matrix; the structure of the error is not the same one.

## Algorithms comparison with TANAGRA

Compared to the two other packages, TANAGRA uses a tree to represent the treatments. That simplifies its structure, but induced a strong constraint; it is not possible to specify two data sources. It is consequently necessary to prepare the data differently.

### Data preparation

We use BREAST\_ALL.XLS<sup>2</sup>. All the examples are gathered in the same sheet; we add a new column, which enables us to distinguish training set and test set (STATUS).



#### Data importation

We close EXCEL<sup>3</sup> and execute TANAGRA. We create a new diagram and import BREAT\_ALL.XLS.

<sup>&</sup>lt;sup>2</sup> TANAGRA can read XLS format. The dataset must be in the first sheet of the workbook.

<sup>&</sup>lt;sup>3</sup> EXCEL locks the file that it is handling; we must close the file before executing TANAGRA.



#### Training and test subdivision

We use the SELECT EXAMPLES component (INSTANCE SELECTION) in order to select the active (training) examples.

TANAGRA 1.4.4 - [Dat	aset (breast_all.xl	s)]				
🕎 File Diagram Componen	t Window Help					_ 8 ×
D 📽 🖪   👪			ite-value	examples selectio		
	Default title	_				
🖃 🏢 Dataset (breast_all	xls)	Para	ameters			
🔤 🚵 Select example	s 1					
Para	meters		Attribute :	status	~	_TOW\sp
/ Exec	ute		- and a constant			
View						
			Value :	train	~	
i i						
						-
				ОК	Cancel Help	<u>,</u> ≚
			<u></u>			
Data visualization	Statistics		Nonpar	ametric statistics	Instance selection	<u>ר</u> ן
Feature construction	Feature select	ion		Regression	Factorial analysis	-
DIS	Clustering			ny learning	Meta-spy learning	
Sou learning assessment	Scoring		-	Association	meta spo tearning	
	Scoring					
Kecover examples	Select examples					
🔪 Samhring	orrarinieu sampung					

When we execute the component (VIEW menu), we see that we have 499 selected examples for the following computations.

#### Select attributes

We add a DEFINE STATUS component in order to select the TARGET attribute (CLASS); the continuous attributes are INPUT. We do not need use STATUS attribute here.





#### Learning method

We must insert the three learning methods in the diagram. We present the detailed operation for the logistic regression. There are two steps when we want to add a supervised algorithm in the diagram: first, we insert a meta-supervised component that defines the aggregation strategy (SUPERVISED LEARNING – META SPV LEARNING tab)

🌋 TANAGRA 1.4.4 - [Sel	ect examples 1]					
🏆 File Diagram Componen	t Window Help		_ @ ×			
D 📽 🖬   👪						
Default	title	Select example	es 1			
🖃 🏢 Dataset (breast_all	.xls)	Parameters				
😑 💉 Select example	s 1	Attribute selection : status				
😑 🎎 Define statı	us 1	Value selection : train				
Supervis	ed Learning 1					
		Results				
		499 selected examples from 699				
L		Computation time : 0 ms.	~			
<b>I</b>	Co	mponents				
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				
🕨 Arcing [Arc-x4]	Boosting					
🕑 Bagging 🛛 🔍	Supervised Learning					

Second, we embed in this component the learning strategy BINARY LOGISTIC REGRESSION (SPV LEARNING tab). This implementation of logistic regression is slightly slower than the others, but it has the advantage of providing a series of additional statistics.



We insert the other learning methods in the diagram: SVM (C-SVC) and the decision tree (C-RT). We obtain the following diagram.

💇 TANAGRA 1.4.4 - [De	fine status 1]			
🏆 File Diagram Componer	nt Window Help			_ 8 ×
D 📽 📮 👫				
	Default title			
🕞 🥅 Dataset (breast al	Lxk)		Define status	1
Select example	er 1		Parameters	
		Target :	1	
		Input : 9	<b>10</b>	
Supervi	ised Learning 1 (Binary logistic	regression)	Ve:U	
Sup	ervised Learning 2 (C-SVC)		Desults	
	Supervised Learning 3 (C-RT)		Results	
		Attribu	rte Target Input Illustrative	
		status		
		clump	- yes -	~
		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
<sup>1</sup> / <sub>1+E*</sub> Binary logistic regressi	on 🗽 C-SVC	🕒 K-NN	> M	ultilayer perceptron
- <mark></mark> C4.5	Decision List	🎆 Linear di	scriminant analysis 🔒 👖 🕯	ultinomial Logistic Regression
A, C-RT	👬 ID3	1 1+8**Log-Reg	TRIRLS 🐻 N	aive bayes
<			)	>

We start the execution of the whole diagram by selecting the VIEW menu of the last component. **The models are built using only the selected examples**.

#### **Comparison of performances**

To compare the performances, we must insert again the DEFINE STATUS component in the diagram by clicking on the short cut into the toolbar. We set the CLASS attribute as TARGET; the predictions of each method are the INPUT attributes. We note that **these predictions are computed for the whole dataset, including the unselected examples**.



We add the TEST (SPV LEARNING ASSESMENT tab) in the diagram. We do not forget to specify that the confusion matrix computation must be done on the unselected examples, which represents the test set.



The view menu displays the following results:

						Test 1	
						Parameters	
aluation se	t : <b>unse</b>	lected examp	les				
Docute							
pred Spulptance 1							
Error este 0.0450							
Values prediction				Confusion waters			
Value	Recall	1-Precision		heanin	malignant	Sum	
begnin	0.9562	0.0224	begnin	131	6	137	
malignant	0.9524	0.0909	malignant	3	60	63	
			Sum	134	66	200	
		n	red Soviost	ance 2			
	Freor ra	P.	od_opinise		1550		
Values prediction			Confusion matrix				
Value	Decall	1.Precision	bognin walignant Cum			Sum	
hegnin	0.9416	0.0227	heanin	179	8	137	
malignant	0.9524	0.1176	malignant	3	60	63	
Ū			Sum	132	68	200	
		D	red Souloct	ance 3			
	<b></b>	р.	ieu_spainste	ance_J	750		
Valu	error ra	iction		Confusi	0750 op matrix		
Value	n n n n n n n n n n n n n n n n n n n		Contrasion matrix				
value	Recall	1-Precision	<b>b</b> a sector	Degnin	maugnant	Sum	
Degnin	0.9051	0.0159	Degnin maligna-t	124	13	137	
maugnant	0.9683	0,1757	maugnant	124	61	200	
			Sum	120	74	200	

The classification accuracy rates are:

- Decision tree: 92.5% (error rate 7.5%);
- Logistic regression: 95.5%
- Linear SVM: 94.5%.

### Conclusion

We see in this tutorial that it is easy to perform a comparison of algorithms using a predefined test set with ORANGE, WEKA and TANAGRA.

The results can be slightly different between the packages. This is not surprising because of the heuristic nature of learning algorithms. The effect of the implementation choices also is not negligible. Nevertheless, very large differences would have been alarming.