Subject

Comparison of TANAGRA, ORANGE and WEKA when we build ROC curve on a dataset.

TANAGRA, ORANGE and WEKA are free data mining softwares. They represent the succession of treatments as a stream diagram or a knowledge flow. Sometimes, there is a little difference between these softwares. Nevertheless, we show that in spite of these differences, these softwares often handle the same problems and give a very similar presentation of results. In this tutorial, we try to build a roc curve from a logistic regression.

Regardless the software we used, even for commercial software, we have to prepare the following steps when we want build a ROC curve.

- Import the dataset in the soft;
- Compute descriptive statistics;
- Select target and input attributes;
- Select the "positive" value of the target attribute;
- Split the dataset into learning (e.g. 70%) and test set (30%);
- Choose the learning algorithm. Be careful, the softwares can have different implementation and present a slightly different results;
- Build the prediction model on the learning set and visualize the results;
- Build the ROC curve on the test set.

According the softwares, the progression can be different but it is clear that we must, explicitly or not, process these steps.

Dataset

We use the dataset from the Komarek's website which implement the LR-TRIRLS logistic regression library (<u>http://komarix.org/ac/lr</u>). The DS1-10 dataset contains 26733 examples, 10 continuous descriptors; the frequency of the positive value of the target attribute is 5%. We use ARFF file format for WEKA and TANAGRA, TXT for ORANGE (TANAGRA can also handle TXT file format).

Building a ROC curve with WEKA

The number of methods is impressive in WEKA, but it is also the main weakness of this software, a through initiation is necessary. The needed components for the construction of a roc curve are not obvious.

Execution of WEKA

When we execute WEKA, a dialog box enables to choose the execution mode. We select the KNOWLEDGE FLOW option. We have used the 3.4.7 version in this tutorial; the results can be slightly different on the others versions. The organization of the software is classic. In the top of the window, we find the tools, machine learning components, in some palettes. The KNOWLEDGE FLOW layout allows us to define the succession of data treatments.



Load the dataset

The ARFF LOADER component (DATASOURCES palette) enables to load a dataset. We add it in the workspace, we can select the file with the CONFIGURE menu of the component.



Descriptive statistics

The ATTRIBUT SUMMARIZER component (VISUALIZATION) shows the data distribution; we can quickly detect abnormalities in the dataset. We add this component and we to connect ARFF LOADER to this new component (DATASET connection).

		· · · · · · · · · · · · · · · · · · ·		4				
🖆 W	eka KnowledgeFlow E	nvironment						
	DataSources DataSinks	Filters Classifiers	Clusterers Evaluation	Visualization				
	Visualization							
	E A	A	rta.	, 💼	r 💼 🗌	r 💼 🛛	re.	1
6	1 Lar	1 Q	1 <mark>1</mark>	1 Ar	1 ar	1 <u>10</u>	1 <u>wa</u> r	
	Data	Scatter	Attribute	Model	Text	Graph	Strip	
	Visualiser	FlotMatrix	Summariser	PerformanceChart	Viewer	Viewer	Chart	
-Knov	viedge Flow Layout							2
								^
		dataSet						
	Edit	1	~					
	Delete		Summariser					
	ArffL Configure.							
	Connections							
	instance							
	Actions							
	Start loadi	ng						
								~
<]		>
Statu	15							
Wek	come to the Weka Knowledg	je Flow						Log

The treatment will be executed when we select the START LOADING menu of the ARFF LOADER component. To see the results, we select the SHOW SUMMARIES option of ATTRIBUTES SUMMARIZER.



We see that there are not abnormalities on the descriptors; we see also that we have imbalanced dataset (804 "positive" vs. 25929 "negative").





Define target and input attributes

The last column is the default class attribute for WEKA, but we can also explicitly select the column of the class attribute. In this case, we use the CLASS ASSIGNER component (EVALUATION palette). We add this component in our diagram; we connect ARFF LOADER to this new component (DATASET connection). We select the CONFIGURE menu in order to select the right class attribute.



In the following step, we must specify the "positive" value of the class attribute. We use the CLASSVALUE PICKER (EVALUATION palette) component and select the "TRUE" value in the parameter dialog box.



Subdivision of the dataset into "learning" and "test" set

We want to build our prediction model on the 70% of the whole dataset, and compute the ROC curve on the remaining. So, we set the TRAINTEST SPLIT MAKER (EVALUATION) in the diagram and configure its parameters.



Logistic regression

The logistic regression component is in the CLASSIFIERS palette. We set it in the diagram, we connect **twice** the TRAIN TEST SPLIT MAKER to this new component: twice because we must use together the training and the test set which are produced by the same component.

🍰 Weka KnowledgeFlow Environment										×			
	DataSources	DataSinks F	ittees Classifiers	Justerers	Evaluation Vi	sualization						ſ	R
			\sim			functions		_					6
								LODISTIC	MULTI-PERC				5
	Complement	Naive	NaiveBayes Welsis esial	NaiveBayes Nicela	NaiveBayes Wednessble	Least	Linear	Teninaie	Multilayer	Pace	RB F	Simpl	
	< A livebayes	Dayes	Mulcihomiai	Simple	opdateable	neasq	Regression	Logistic	Priception	Regression	Network	Kegr >	
Knov	vledge Flow Layo	out											
								I					^
				•				1					
			data 4	<u></u>		`	``	1					=
	dat dat		- ' L	Þ¢ ′	a con	traini testSe	ngSet	1					
		Class	Class	sValue	TrainTest		- 🚧	1					
-		Assigner	Pie	:ker	SplitMaker		LOGISTIC						
ls1_	10.arff	apet											
							Logistic						
		àttr i huta											
		Summarizer											
<												>	
-Statu	IS												
Web	come to the Weka	Knowledge F	low									Lo	g

The results

To see the results of the regression, we connect the LOGISTIC component to TEXT VIEWER (VISUALIZATION palette) that we set in the diagram.



We execute again the diagram (START LOADING of ARFF LOADER component). The SHOW RESULTS of TEXT VIEWER opens a new window with the results of the learning process.



We obtain the regression coefficients.

👉 Text Viewer		X
Result list	_ Text	
16:01:40 - Model: Logistic	=== Classifier model ===	^
	Scheme: Logistic	
	Relation: dsl_10.arff	
	Logistic Regression with ridge parameter of 1.0E-8	
	Coefficients	
	Variable Coeff.	
	1 0.1294	
	2 0.3301	
	3 0.7174	
	4 0.3972	
	5 0.164	
	6 -0.0313	
	7 -0.0689	
	8 0.7603	
	9 -0.2524	
	10 0.0683	
	Intercept -4.2745	
		~

ROC curve

In order to evaluate the learning process, we add the CLASSIFIER PERFORMANCE EVALUATOR component (EVALUATION). We connect the BATCH CLASSIFIER output of LOGISTIC to this new component.



We set MODEL PERFORMANCE CHART (VISUALIZATION) in the diagram; we use the THRESOLD DATA (!) output of the CLASSIFIER PERFORMANCE EVALUATOR when we connect the two components.



We run again the diagram with the START LOADING menu of ARFF LOADER. We select the SHOW PLOT menu of the last component (MODEL PERFORMANCE CHART). We obtain the ROC curve.



WEKA is a very powerful tool, but knowing all its features is not obvious.

ROC curve with TANAGRA

There are three parts in the main window of TANAGRA. A the left, we have the stream diagram where we define our treatments; at the bottom, the data mining tools in various palettes; at the center, the window for displaying the results.

Load the dataset

We select the FILE / NEW menu in order to create a new diagram.

Didacticiel - Etudes de cas

TANAGRA 1.4.4					
Elle Diagram Window Help					E AL
Choose your dataset and start download					2
Diagram title :		1			20
Data mining diagram file name :		a			200
JD.MempleXedeladicbdm		•			K-AD
Dataset (*.txt,*.arff,*.xls) :		<u>.</u>			1. 1.8
D:\DataMining\Databases_for_mining\dataset_for_soft_de	ev_and_compa 🔀	Γ] - ``			5 J
		1 V	(10 A 10
	Tanagra				? 🔀
ОК	Regarder dans :	🗀 from_lr_trirls_roc_	curve	S 🕫 📂 🛄 -	
		Total 10.arff			
	Mes documents récents				
Die teller	Bureau				
Data visualization statistics Nonparam					
Coulographica Moto coulographica Coulographica					
	Mes documents				
Correlation scatterplot 🔤 Export dataset 🧖 Scat					
	Poste de travail				
The second second second					
		Nom du fichier :	ds1_10.arff	×	Ouvrir
	Favoris réseau	Fichiers de type :	WEKA dataset	~	Annuler

Descriptive statistics

We add a DEFINE STATUS in the diagram (click on the toolbar icon); we set as INPUT all continuous attributes.

TANAGRA 1.4.4 - [Datas	et (ds1_10.arff)] Window Help			- 7
Dataret (dr.1.10 arff)	efault title		Dataset (ds1_11	.arff)
Define status 1	Define attribute statuses Parameters Attributes : C V1 C V2 C V3 C V4 C V5 C V5 C V5 C V5 C V5 C V6 C V7 C V8 C V8 C V8 C V8 C 10 D CLASS	Target V1 V2 V3 V4 V5 V6 V7 V7 V7 V7 V7 V7 V7 V7 V7 V7	Input Illustrative	v_and_comparison\from_ir_triris_ro
Data visualization Feature selection Spv learning	Regression Fa Meta-spv learning Spv le	OK ctorial analysis arning assessment	Cancel Help PLS Scoring	ure construction Clustering Association
➡Correlation scatterplot	Export dataset 🔐	Scatterplot	🕅 View dataset	🧱 View multiple scatterplot

We set the MORE UNIVARIATE CONT STAT (STATISTICS) in the diagram, we execute the component (VIEW menu), detailed results are displayed.

🕎 TANAGRA 1.4.4 - [More Univariate cont stat 1]												
The Diagram Component Window Help												
D 📽 🔜 🎎												
Default title			Stati	Values	Count	Percent	Histogram	<u>^</u>				
🖃 🏢 Dataset (ds1_10.arff)		Average	-6.529	x_<12.1089	127	0.48%						
🖃 🚰 Define status 1		Median	-6.7182	-12.1089 =< x < -10.7646	734	2.75%						
- Kore Univariate co	int stat 1		Std dev. [Coef of variation]	2.4774 [-0.3794	-10 7646	2242	0.45%					
4			MAD [MAD/STDDEV]	2.0503 [0.8276	-10.7646_=<_X_<9.4204	2312	0.05%					
			Min*Max (Full range)	-13.45*-0.01 [13.44	-9.4204_=<_x_<8.0762	4724	17.67%					
		V1	1st * 3rd quartile [Range]	-8.39*-4.70 [3.69	-8.0762_=<_x_<6.7319	5422	20.28%					
			Skewness (std-dev)	0.1588 (0.0150	-6.7319_=<_x_<5.3877	4556	17.04%					
			Kurtosis (std-dev)	-0.5790 (0.0300	-5 3877 - < × < -4 0435	3008	1/1 96%					
	· · ·				-5/30///1_/_10435	5770	14.30%					
					-4.0435_=<_x_<2.6992	2978	11.14%					
					-2.6992_ - <_x_<1.3550	1419	5.31%					
1					x>=1.3550	463	1.73%					
									•			
		-		Components								
Data visualization	Statistics	Nonp	parametric statistics	Instance selection	eature construction Feature selection			Regression				
Factorial analysis	PLS		Clustering	Spv learning	Meta-spv learning	5pv learning	assessment	Scoring				
Association												
🕁 Bartlett's test 🛛 🚟	Levene's test		🔩 One-way MANOVA	🚻 Univariate discrete	stat							
중 Brown - Forsythelastest 분 Linear correlation 프 Paired T-Test												
A Fisher's test La More Univariate cont stat T-Test												
Group characterization 🖄 Normality rest 🛄 T-Test Unequal Variance												
Crodp exploration M	One-way HNOV	•	Ers onwarrate continu002 51	. a .								
<u> </u>												

Subdivision of the dataset into "learning" and "test" set

We select again the root of the diagram and insert the SAMPLING component (INSTANCE SELECTION); we set the right parameters (PARAMETERS menu).



Target and input attributes

With a new DEFINE STATUS component, we set CLASS as TARGET attribute, the continuous attributes as INPUT. We obtain the following result when we select the VIEW menu.

TANAGRA 1.4.4 - [Defi	ne status 2]									
Tile Diagram Component	Window Help						_ @ ×			
Default tit	le					Parameters	<u>~</u>			
🖃 🏢 Dataset (ds1_10.arff	i) i i i i i i i i i i i i i i i i i i	Target : 1								
🖃 🏠 Define status 1		input : 10 Illustrative	· 0							
More Univar	iate cont stat 1									
Define statu	e 2					Results				
	· (Attribute	Target	Input	Illustrative	•				
		V1	-	yes	-					
		V2	-	yes	-					
		V3	-	yes	-					
		V4	-	yes	-					
		V5	-	yes	-		=			
	·-?\	V6	-	yes	-					
		V7	-	yes	-					
		V8	-	yes	-					
		//9	-	yes	-					
		V10	-	yes	-					
		CLASS	yes	-	-					
				_						
		Computatic Created at	n time : 19/02/2	0 ms. 2006 17	7:06:25					
							×			
Data visualization	Ctatictics	Nopr	Compon	ents rio sta	tiction	Instance colorition				
	Feature colocition	Non	Pear	nu sta	icistics	Factorial application				
	Clusteries		Ceute	avair -		Mete ceu leaveire				
PL3	Clustering		spv le	arning Jatia -		meta-spv tearning				
Spy learning assessment Scoring Association										
🔨 Recover examples 📝 S	Sampling 🖉 🖉 Sel	ect examp	oles	🖋 Str	atified san	որնոց				
L										

Supervised learning

We want to add the logistic regression component in the diagram. There are two steps: first, we add a "meta-supervised" component (SUPERVISED LEARNING from META-SPV LEARNING palette); second, we embed the LOG-REG TRIRLS component (palette SPV LEARNING). This component comes from the Komarek's library (http://komarix.org/ac/lr). Its implementation seems very fast and robust. The results are displayed in a new window.





The confusion matrix is not relevant in our context because our dataset is unbalanced. The most important is that our classifier can set the "positive" examples before the "negative" ones according a computed "score" which is proportional to the "positive" conditional probabilities.

Computing the "score"

We set the SCORING component (SCORING) in the diagram; we select the PARAMETERS menu in order to define the "TRUE" value of the class attribute as "positive".

R.R.



ROC curve

We add a new DEFINE COMPONENT in the diagram; we set as TARGET the class attribute, and the generated attribute SCORE_1 as INPUT. We can select several INPUT attributes here and compare the performances of various learning algorithms.



Last, we add a ROC CURVE component in the diagram. We select "TRUE" as the "positive" label; the computation must be realized on unselected examples (test set).



We obtain the results (VIEW menu). We do not build the curve but a table with all necessary values (TRUE POSITIVE rate, FALSE POSITIVE rate). Several indicators are available.

TANAGRA 1.4.4 - [Roc curve 1]												
Tile Diagram Component Window Help							_ 8 ×					
Default title	1r				Par	rameters	~					
🖃 🎹 Dataset (ds1_10.arff)		Positive class value : true										
🖨 🏰 Define status 1		Used examples : Unselected										
🖃 🌶 Sampling 1						Results						
🖃 🎦 Define status 2		ROC Curve	•									
Supervised Learning 1 (Log-Reg TRIRLS)		Sample cize (202		ר								
E ↓ Scoring 1		Positive example:	; ;:242	⊢	<u> </u>		=					
		Negative example	s : 777	8								
		score Attribute		Score 1		L						
	K	AUC		0.8208								
		Target size (%)	Scor	e FP-Rate	TP-Rate							
		0	0.43	12 0.0000	0.0000							
		5	0.11	75 0.0406	0.3512							
		10	0.073	79 0.0869	0.5207							
		15	0.05	79 0.1349	0.6364							
		20	0.04	44 0.1835	0.7314							
		25	0.03	52 0.2339	0.7686							
		30	0.028	0.2845	0.7975							
		35	0.023	37 0.3354	0.8182							
<		40	0.019	97 0.3862	0.8430							
	JL	AF	0.04	20 0 4075	0.0540							
Data visualization Statistics	Non	Co narametric static	mpon Hies	ents Instance	selection	Feature construction						
Fosture coloction Pogression	1400	parametric statis	tics	Instance	i selection	Clustering						
	C	ractoriat analysis		, 	·L5	Clustering						
weta-spy tearning	spv	rearring assessm	enc	1 20	oring	ASSOCIATION						
∠ Lift curve ↓ Scoring Rec curve												
Vice came												
							.:					

TANAGRA enables to compute quickly the ROC curve. We note the importance of DEFINE STATUS before the ROC component. It enables to compare various input columns for the examples ranking: they can be computed by supervised learning algorithm, but they can be also supplied by another process.

ROC curve with ORANGE

The utilization of ORANGE is very simple. It is particularly well suited for our subject. But, sometimes, some default options are not clear; we must study carefully the parameters of the software.

Orange

The main window of ORANGE is similar to WEKA interface: tools are in the top of the window; there is a workspace where we can define the stream diagram.



Load the dataset

ORANGE can handle TXT (tab separator) file format. We select the FILE tool, it is automatically added in the workspace. We can select the dataset with the OPEN menu.



Descriptive statistics

ORANGE offers nice graphical tools for descriptive statistics. We select the ATTRIBUTE STATISTICS component. We connect FILE to this new component; the execution is automatically started. We can see the result with the OPEN menu.



Target and input attributes

By default, the last column is the class attribute. We can modify the selection with the SELECT ATTRIBUTES component (DATA). We see that the default selection is right.



Logistic regression and learning evaluation

First, we add the LOGISTIC REGRESSION from the CLASSIFY palette; second, we add the TEST LEARNER component (EVALUATE). We set as "train – test" the evaluation process in this last component, the learning set size is 70% (OPEN menu).



We connect the LOGISTIC component to TEST LEARNERS, and SELECT ATTRIBUTES to TEST LEARNERS. In this last connection, a dialog box appears, it enables us to select the right dataset.



When we confirm this connection, the computation is automatically started. We can display the results with the OPEN menu of TEST LEARNERS.



Various indicators are computed, including the AUC ratio. But we do not find which value of the class attribute defines these results.

ROC curve

ORANGE has an interactive graphical tool for ROC curve computation. We set this component (ROC ANALYSIS from EVALUATE) in the diagram. We connect the TEST LEARNERS to this new component.



When we activate the OPEN menu of ROC ANALYSIS, we obtain the following chart. We can interactively select the "positive" class.



ORANGE is very easy to use. If we want to compare the performances of various learning algorithm, we add the learning components in the diagram and connect them to TEST LEARNERS: the new ROC curves are automatically included in the ROC ANALYSIS tool. Impressive!

Conclusion

So that the results are really comparable, it would have been necessary to subdivide the file in training set and test set, then launching the softwares on the same datasets.

ORANGE and WEKA use the same strategy. The dataset is subdivided into two files. Because they use graph, their diagram can handle several data sources.

TANAGRA adopts another strategy. We must add a new column in the dataset; it defines the role of each example. Then we replace the SAMPLING component with the SELECT EXAMPLES component in the previous diagram, the other components of the diagram are the same one.