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.

	licrosoft Ex	cel - breas	t_for_	classificatio	n.xls								
	<u>Fichier</u> <u>E</u> ditio	on <u>A</u> ffichage	Insert	ion Forma <u>t O</u>	utils <u>D</u> onnées F	e <u>n</u> être <u>?</u>							- 8 ×
	🛩 🖪 🔗	a 🖪 🖓	8 8	🗈 🖻 💅		🔔 Σ f 🖉		I 🚜 🕄	» 10 - 1	i <i>I</i> ≣	≡ 🗉 • 🕭	• <u>A</u> •	»
· · · ·	J502	- :	-										
	A	В	_	С	D	E	F	G	Н	1	J	K	
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		8	· · · ·	1	10	10	3	malignant		
496		5	- 1	1	3	4	1	3	2	1	begnin		
497		5	- 2	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	4	2	1			
509		4	10	10	8	5	8	<u></u>	10	1			
510		4	2	4	4	2	5	2	5	1			+-
512		4	10	4	4	2	1	10	1	1			
512		2	- 10 - 1	2	1	- -	10	10	4	1			
	► ► \dat	taset /						•					
Deg	isin + 🎝 🙆) Formes a	utomatic	jues 🔹 🔨 🔌		1 🖸 🔌 -	<u> A</u> - E	■ ☴ ■	<i>i</i> -				
Prêt								[NU	M	

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

N	licrosoft Exce	el - breast_f	or_classific	ation.xls							
	Fichier Edition	Affichage Ins	sertion Forma	: Outils Donnée:	s Fenêtre ?						_ & ×
	 ~ 🗖 🔊	 Æ []à #8⊊	V 🗈 📾	a 🗤 - 🗠 -	3 7 6		4 0	» 10 -	c 7 =	= . *	- A - »
			<u> 6 43 ∎</u>	× • · · · · · · ·	🚡 Z 3×	Ģ€ Z† A†		• <u>10</u> •	u 1 =	= 🗆 * 🛩	• 📶 • 🗸
	J5U2 -		0	5	-	-	0				
4	A	В		D	E	F	e e e e e e e e e e e e e e e e e e e	H		J	
195	ciump	ucensize	ucelisnape	mgaonesion	sepics 1	DNUCIEI	Denromatin 7	normnuci	mitoses	Class 1 molignont	Status
405	3	1		1 1		3	1 7			1 manynani 1 hognin	learning
400	3	1		1 3	2	2	1 2	1		1 begnin 1 begnin	learning
488	10			́ь́ 10	, 1 i	2 6 1I	י 2	7	1	Ω malignant	learning
489	6		á –	7 4		5 1	З 4	. 2		1 heanin	learning
490	4	1		1		2	1 3	1		1 beanin	learning
491	5	1		1 *		2	1 2	1		1 beanin	learning
492	4	8	3	6 3	3	4 1	J 7	1		1 malignant	learning
493	3	3	3	5 2	2 :	3 1) 7	1		1 malignant	learning
494	8	10)	5 3	3 (в .	4 4	10		3 malignant	learning
495	10	10)	7 8	3	7	1 10	10		3 malignant	learning
496	5	1		1 3	3 .	4	1 3	2		1 begnin	learning
497	5	2	2	1 1	l :	2	1 1	1		1 begnin	learning
498	5	1		2 *	l : : : : :	2	1 1	1		1 begnin	learning
499	4	1		1 1		1	1 2	! 1		1 begnin	learning
500	5	10) .	10 é	5	4 :	54	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	2		2	1 2	1		1	to_classify
504	4			1		2	1 2			1	to_classify
505	10	10		1 10		2	I 3	10		1	to_classify
505	10	10	,	1 4		0 1	1 0			1	to_classify
507	2			2 4	,	2	1 2			1	to_classify
500	10	10	, 1 ·	10 8	-	2 6 1	37	10		1	to_classify
510	4	1	í.	1 1		2	1 2	· 1		1	to_classify
511	2		3	4 4		2 .	5 2	5		1	to_classify
512	4	10	1	8 9		4	1 10	. <u> </u>		1	to_classify
513	5	2	2	3		6 1) 5	1		1	to classify
514	2	1		1		2	1 2	1		1	to_classify
	► ► datas	set /		2		2	· []	1			
Des	sin + 🔓 🙆	Formes autom	atiques 🕶 🔨		4 🛛 🔌	- <u>/</u> - <u>A</u> -		 .			
Prêt	t										

(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".

	hicrosof	t Exce	el - breas	t_1	ior_c	lassific	ation	.xls														
	Eichier E	dition	Affichage	In	sertior	n Forma	i <u>t O</u> ut	ils <u>D</u> onr	nées	Fenêtre	2											_ 8 ×
I D	- 	A.	 []_ **	- 9	¥ſ	3× 🛱	~	- -	× -	- Σ	f.	Q	A∣ Z	1	n 🗖 🗩	»	10 -	6.7	=	= .	& - A -	»
	E 400	••••			90		~			18 Z	,*	₩EI :	Z* A	* 6	u 🛷 🛶	•	10 .	0 1	-		× · •	•
	F498					0		D					г				11		1			
1	A		D			llohong		U rodboo	ion			her	F		hebrometin	norm		mitooo		J	Status	
186	ciump	3	ucensize		1	ensnape	1	gaunes	1011	sepics	-	DHC	Iciel	1	Demontatin	7	1	millose	1	heanin	learning	
400		3			1		1		- 3		- 5			- 1		2	1		- 1	l begnin	learning	
488		10			5		6		10		Ē			- 10	1	7	7		10) malignant	learning	
489		6			9		7		5		Ē			8		4	2)	1	l heanin	learning	
490		4			1		1		1		2			1		3	1		1	beanin	learning	
491		5			1		1		1		2			1		2	1		1	beanin	learning	
492		4			8		6		3		4			10	l i	7	1		1	malignant	learning	
493		3			3		5		2		3	1		10	l - Contra de Co	7	1		1	l malignant	learning	
494		8		1	0		5		3		E	1		- 4		4	10)	3	3 malignant	learning	
495		10		1	0		7		8		- 7			1	1	0	10)	3	3 malignant	learning	
496		- 5			1		1		3		- 4	L .		1		3	2	2	1	begnin	learning	
497		- 5			2		1		1		2	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		1	0		10		- 5		- 4	ļ.		- 5		4	4	L	1	malignant	learning	
501		3			1		1		1		2	2		1		1	1		1	begnin	learning	
502		3			1		1		1		2	2		3	l	3	1		1	begnin	to_classify	
503		4			2		2		1		2	2		1		2	1		1	begnin	to_classify	
504		4			1		1		1		2	2		1		2	1		1	begnin	to_classify	
505		6			1		1		1		2			1		3	1		1	begnin	to_classify	
506		10		1	0		10		4		E	•		1		3	10)	1	begnin	to_classify	
507		2			1		1		1		2	2		1		2	2	2	1	begnin	to_classify	
508		1			3		3		2		2	2		1		7	2	2	1	begnin	to_classify	
509		10		1	0		10		8		E			8		7	10)	1	begnin	to_classify	
510		4			1		1		1		2	2		1		2	1		1	begnin	to_classify	
511		2			3		4		4		2	2		- 5		2	5	5	1	begnin	to_classify	
512		4		1	0		8		-5		4			1	1	כ	1		1	begnin	to_classify	
513		- 5			2		3		1		E			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	-
		datas	et /		-				-								10				· · · ·	
Des	sin + <table-cell></table-cell>	G	Formes au	iton	natique	es • 🔨	\mathbf{X}			4 🗕	<u>ی</u>	- 🦽	- A	-	= ≕ ≓ ।	1	-					
Prêt	t																				NUM	

R.R.

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

🔀 Micr	osoft Excel - br	east_for_classification.xls		
📳 Eichi	ier Edition Affich	ige Insertion Format Qutils Données Fenêtre ?		_ 8 ×
D 🚅		🖤 🔏 🛍 🐔 🛷 🖙 🖙 🍓 Σ 🏂 🖓 🖓 🕼 🚜 🤋 👋 🕺 10 🗸 G	I ≣ ≣	•• <u>A</u> • *
F5				
	Enregistrer so	IS 📝 🚺	l J	K 🔺
1 clu	Enregistrer <u>d</u> ans :	🛅 breastcancer 💽 🔶 🔁 🔞 🗙 📸 🖬 🗸 Outils 🗸	oses class	Status
492		E breast.txt	1 malignant	learning
493		E breast malignant scored.txt	1 malignant	learning
494	Historique	E breast noisy.txt	3 malignant	learning
495	misconique	E breast sorted on mitoses.txt	3 malignant	learning
496		🗐 output breast classified.txt	1 begnin	learning
497			1 begnin	learning
498	Mes documents		1 begnin	learning
499			1 begnin	learning
500			1 malignant	learning
501			1 begnin	learning
502	Bureau		1 begnin	to_classify
503			1 begnin	to_classify
504			1 begnin	to_classify
505	<u>*</u>		1 begnin	to_classify
506	Favoris		1 begnin	to_classify
507			1 begnin	to_classify
508		New dy Siskier , breast for classification byt	1 begnin	to_classify
509	Eavoris réseau		1 begnin	to_classify
510		Type de fichier : Texte (séparateur: tabulation) (*.txt) (*.txt) 💌 Annuler	1 begnin	to_classify
511			1 begnin	to_classify
			i nomin	
De <u>s</u> sin •	- 🔓 🍪 Forme	s automatiques • 🔨 🌂 🖸 🔿 🎽 📲 💆 • 🚣 • 🚍 🗮 🛱 🎒 •	1	
Prêt				

Classifier deployment on a new dataset

Dataset importation in TANAGRA

Import the dataset in TANAGRA and define a stream diagram.

·			0.0511 100100 0 10 10011510	
TANAGRA 1.1.1				
File Diagram Window Help				
D 🚅 🔳 🗱				
Data min	ing diagram (empty)			
	Choose your dataset an	d start download		
	Diagram title :			
	Delault title			
	Data mining diagram	n file name :		
	JD: DataMining Data	pases_for_mininglbenchman	K_datasetsibreastcance 📷	
	Dataset (text file) :			
	r_mining\benchmar	k_datasets\breastcancer\brea	ast_for_classification.txt 崖	
				_
		ОК	Cancel Help	
		Components		
Data visualization	Descriptive stats	Instance selection	Feature construction	Feature selection
Regression	Factorial analysis	Clustering	Spv learning	Meta-spv learning
Spv learning assessment	Association			
Correlation scatterplot	🖺 Export dataset	🧖 Scatterplot	🔛 View dataset	🤽 View multiple scatterp
<				>
				.::

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).

TANAGRA 1.1.1 - [Dataset (breast_for_classification.txt)]					X
Tile Diagram Component Window Help					n ×
🗅 🛩 🖬 🍇					
Default title	-			••••••••••••••••••••••••••••••••••••••	~
🖃 🎹 Dataset (breast_for_classification.txt)				Results	
🦳 者 Select examples 1	11 attribute(: 699 example	s) (s)			
	Attribute	Category	Informations		
	clump	Continue	-		
Attribute-value examples selection	ucellsize	Continue			
	ucellshape	Continue	-		
Parameters	mgadhesion	Continue	-		
	sepics	Continue			
Amilia da Chata	bnuclei	Continue	-		=
Attribute : Jatus	bchromatin	Continue			
	normnucl	Continue	-		
Value : learning 💌 🖌	mitoses	Continue	-		
	class	Discrete	2 values		
	Status	Discrete	2 values		
	Execution tim	e:0ms.			
OK Cancel Help	Created at 23	8/12/2004	13:49:49		~
,	C	omponer	ıts		
Data visualization Descriptive stats Instance	e selection	Fe	ature constru	uction Feature selection Regression	T
Factorial analysis Clustering Spv I	earning	1	Weta-spv leari	ning Spv learning assessment Association	
Recover examples Sampling Stratified sampling					

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.

Case studies

Classifier deployment on a new dataset



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.

R.R.



VIEW DATASET

💯 TANAGRA 1.1.1 - [View dataset 1 [All] (199 exa	mples, 12 att	ributes)]					
Tile Diagram Component Window Help							_ 8 ×
D 🛩 🖬 🎎							-
Default title	bnuclei	bchromat	bchromatin normnucl		class	Status	pred_SpvInsta
🖃 🥅 Dataset (breast_for_classification.txt)	3	3	1	1	begnin	to_classify	begnin
🚊 🎤 Select examples 1	1	2	1	1	begnin	to_classify	begnin
🖮 🔁 Define status 1	1	2	1	1 1	begnin	to_classify	begnin
Supervised Learning 1 (C-RT)	1	3	1		begnin	to_classify	begnin
😑 🌂 Recover examples 1	1	8	10	1	begnin	to_classify	malignant
View dataset 1	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		begnin	to_classify	begnin
	1	10	1		begnin	to_classify	malignant
	10	5	1		begnin	to_classify	begnin
	1		1	1	begnin	to_classify	begnin
	1	2	1	1	begnin	to_classify	begnin
	10	9	10		begnin	to_classify	malignant
	10	8	10		begnin	to_classify	malignant
	10	5	4	4	begnin	to classify	malignant ⊻
<u> </u>							2
		Com	ponents				
Data visualization Descriptive stats	Instance	selection	Feature co	nstruction	Feature s	election	Regression
Factorial analysis Clustering	Spv le	arning	Meta-spv	learning	Spv learning	assessment	Association
⊕Correlation scatterplot 🖉 Scatterplot Export dataset 🔯 View dataset	<u>K</u> it View (nultiple scatte	rplot				

Export results

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

Case studies

Classifier deployment on a new dataset

TANAGRA 1.1.1 - [View dataset 1 [All] (199	xamples, 12 attributes)]	
Tile Diagram Component Window Help		_ 8 ×
Default title	bnuclei bchromatinnormnucl mitoses class Statu	s pred_SpvInsta
😑 🏢 Dataset (breast_for_classification.txt)	lgOpPrmExportDataset	assify begnin 📃
🚊 🥢 Select examples 1		assify begnin
🖶 🏪 Define status 1	Parameters	assify begnin
Supervised Learning 1 (C-RT)	1	assify begnin
-X Recover examples 1	Examples selection 21	assify malignant
👯 View dataset 1	O all examples	assify begnin
Export dataset 1	Selected examples	assify begnin
		assify malignant
	Attributes selection	assify begnin
	💿 all attributes	assify begnin
	O target attributes	assify malignant
		assify begnin
	Filename :	assify begnin
	D:\DataMining\Databases_for_mining\benchmark_datasets\brea	assify begnin
		assify malignant
-		assify malignant
	OK Cancel Help	assiry marignant
	Components	
Data visualization Descriptive stats	Instance selection Feature construction Feature selection	Regression
Factorial analysis Clustering	Spv learning Meta-spv learning Spv learning assessmen	nt Association
⊕Correlation scatterplot	<u>[* :</u> View multiple scatterplot	
		.:

We can view the dataset in a spreadsheet.

	Aicrosoft Exc	el - output_b	reast_classifi	ed.txt					(\times
	Eichier Edition	Affichage Inse	ertion Forma <u>t</u>	<u>O</u> utils <u>D</u> onnées	Fenêtre ?					_ 8	X
	🚔 🖪 🙈	a 🛯 🖤 🗌	X 🖻 🛍 🚿	10 · 01 ·	🎑 Σ f*	-			?	» 6	**
	K1 -	= SI	atus						-	.11	
	Н		J	К	L			М		N	-
1	normnucl	mitoses	class	Status	pred Spyli	nstan	ce 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			<u> </u>			
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		_				-
	())) outp	ut_breast_cla	assified /		4					•	
De	ssin + 🎝 🙆	Formes automa	itiques 🕶 🔨 🦄		4 🗕 🔌	- 🥒	- <u>A</u>	• 🔳 📼	Ħ		0.
Prê	t							NUM			/

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.