1 Theme

Data Mining with R – The "rattle" package.

R (<u>http://www.r-project.org/</u>) is one of the most exciting free data mining software projects of these last years. Its popularity is completely justified (see Kdnuggets Polls – Data Mining/ Analytic Tools Used – 2011). Among the reasons which explain this success, we distinguish two very interesting characteristics: (1) we can extend almost indefinitely the features of the tool with the packages; (2) we have a programming language which allows to perform easily sequences of complex operations.

But this second property can be also a drawback. Indeed, some users do not want to learn a new programming language before being able to realize projects. For this reason, tools which allow to define the sequence of commands with diagrams (such as Tanagra, Knime, RapidMiner, etc.) still remain a valuable alternative with the data miners.

In this tutorial, we present the "Rattle" package which allows to the data miners to use R without needing to know the associated programming language. All the operations are performed with simple clicks, such as for any software driven by menus. But, in addition, all the commands are stored. We can save them in a file. Then, in a new working session, we can easily repeat all the operations. Thus, we find one of the important properties which miss to the tools driven by menus.

To describe the use of the rattle package, we perform an analysis similar to the one suggested by the rattle's author in its presentation paper (G.J. Williams, « Rattle : A Data Mining GUI for R », in *The R Journal*, volume 1 / 2, pages 45—55, December 2009, <u>http://journal.r-project.org/archive/2009-2/RJournal_2009-2_Williams.pdf</u>). We perform the following steps: loading the data file; partitioning the instances into learning and test samples; specifying the types of the variables (target or input); computing some descriptive statistics; learning the predictive models from the learning sample; assessing the models on the test sample (confusion matrix, error rate, some curves).

2 Dataset

We use the «heart »¹ data file. We want to explain the occurrence of the DISEASE from the characteristics of patients. We show here the first instances of the dataset.

B 1	neart_for_rattle.	txt									3
	A	В	С	D	E	F	G	Н	I	J <mark>/</mark> -	Ξ
1	age	sex	chest_pain	trestbps	chol	fbs	restecg	thalach	exang	disease	-
2	31	male	asympt	120	270	f	normal	153	yes	positive	
3	33	female	asympt	100	246	f	normal	150	yes	positive	
4	34	male	typ_angina	140	156	f	normal	180	no	positive	
5	35	male	atyp_angina	110	257	f	normal	140	no	positive	
6	36	male	atyp_angina	120	267	f	normal	160	no	positive	
7	37	male	asympt	140	207	f	normal	130	yes	positive	
8	38	male	asympt	110	196	f	normal	166	no	positive	
9	38	male	asympt	120	282	f	normal	170	no	positive	
10	38	male	asympt	92	117	f	normal	134	yes	positive	
11	41	male	asympt	110	289	f	normal	170	no	positive	
12	43	male	asympt	150	247	f	normal	130	yes	positive	•
H I	(► ► \ hear	t_for_rattle	/				•				1

¹ <u>http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/heart_for_rattle.txt</u>; a description of this data file is available on the following website: <u>http://archive.ics.uci.edu/ml/datasets/Heart+Disease</u>

3 Data Mining with Rattle

3.1 Loading the rattle package

First, we load the rattle package [library()]. Then, we start the GUI with the command rattle().

```
> #loading the package
> library(rattle)
> #lauching the GUI
> rattle()
```

Into the R console, we have...

```
R Console
                                                           - • ×
R version 2.9.2 (2009-08-24)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.
R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.
Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.
> library(rattle)
Le chargement a nécessité le package : pmml
Le chargement a nécessité le package : XML
Rattle: Interface graphique pour data mining utilisant R.
Version 2.5.18 Copyright (c) 2006-2010 Togaware Pty Ltd.
Tapez 'rattle()' pour shake, rattle, et roll vos données.
  rattle()
>
```

From now, we perform all the operations by clicking on the appropriate menu or button. All these operations are recorded as R commands by rattle. The rattle GUI is displayed.



The use of rattle is always the same: we define the command by working in the appropriate tab (Data: load the dataset; Explore: some descriptive statistics; Test: some statistical tests, etc.); then, we launch the calculations by clicking on the EXECUTER button into the toolbar.

3.2 Importing the data file

Into the "Data" tab, we click on the FILENAME button. We select the" heart_for_rattle.txt" data file.

@ R Data Miner - [Rattle]	1 (1) 🙀 (1) 💼 anna		
<u>P</u> roject <u>I</u> ools <u>S</u> ettings <u>H</u> elp	 Rattl 	le Version 2.5.18 <u>togaware.com</u>	
Exécuter Nouveau Ouvrir Enregistrer Report Expert	3 払 rêter Quitter		
Data Explore Test Transform Cluster Associate Model Evaluate Log	<u>\</u>		
Source: Spreadsheet ARFF ODBC R Dataset RData File I	Library 🔿 Corpus 🔘 So	cript	
Filename: (Aucun) Separator: , V Header	R Sélectionner un fichier	r	
Partition Parti	D:\ DataMinir	ng Databases_for_mining	benchmark_datasets
Input Ignore Weight Calculator:	Raccourcis No	om	▲ Modifié
Welcome to Rattle (rattle.togaware.com).	🛞 Récemment uti	heart.txt	22/02/2006
Rattle is a free graphical user interface for Data Mining, software environment for statistical computing and graphic sophisticated environments for data mining, statistical ar	Documents Bureau C	heart_continuous.txt heart_for_rattle.txt heart_h_from_weka.txt	21/04/2004 Hier à 17:29 17/03/2010
See the Help menu for extensive support in using Rattle. I Survival Guide includes Rattle documentation and is availa	Donnees (D:\)	pred_disqual.txt pred_disqual_cart.txt	26/02/2005 26/02/2005 ≡
Rattle is licensed under the GNU General Public License, V ABSOLUTELY NO WARRANTY. See Help -> About for details.	📾 F:\		
Rattle version 2.5.18 Copyright 2006-2010 Togaware Pty Ltd Rattle is a registered trademark of Togaware Pty Ltd	G:∖ 		
To Begin: Choose the data source, specify the details, then click the Execute button.			
	۰ III ا		-
	Ajouter Enlever		TXT Files
			<u>O</u> uvrir A <u>n</u> nuler

We specify the column separator: « SEPARATOR = \t ». Then we click on EXECUTER.

CIS6 R	R Data Miner	- [Rattle (h	eart_for_rat	tle.txt)]				
Pro	ject <u>T</u> ools	<u>S</u> ettings	<u>H</u> elp					Rattle Version 2.5.18 togaware.com
	Exécutor	Nouveau	E Ouvrir	Enregistre	er Repo	ort Exp	ort	🚫 🐳 Arrêter Quitter
Dat	a Explore	Test Trans	form Clust	er Associa	te Model	Evaluate L	og	
Sou	irce: 🔘 S	preadsheet	ARFF	ODBC	R Data	iset 🔘 RE	ata File	🔘 Library 🔘 Corpus 🔘 Script
	2006							
File	name: 📋 h	eart_for_rat	tle.txt 🛛 📇	Separator:	\t 📙 🔽	Header		
	Partition	70/30 🖊	Seed:	42	View	Edit		
	Input 🧲	Ignore W	/eight Calcu	lator:]	Target Data Type
No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Comment
						autorite .	ignore	comment
1	age	Numeric	۲	0	0	O	O	Unique: 38
1 2	age sex	Numeric Categoric	0	© ©	0	0	© ©	Unique: 38 Unique: 2
1 2 3	age sex chest_pain	Numeric Categoric Categoric	© ©	0	0	© © ©		Unique: 38 Unique: 2 Unique: 4
1 2 3 4	age sex chest_pain trestbps	Numeric Categoric Categoric Numeric	0 0 0	0	© © ©	© © ©		Unique: 38 Unique: 2 Unique: 4 Unique: 31
1 2 3 4 5	age sex chest_pain trestbps chol	Numeric Categoric Categoric Numeric Numeric	0 0 0 0					Unique: 38 Unique: 2 Unique: 4 Unique: 31 Unique: 150
1 2 3 4 5 6	age sex chest_pain trestbps chol fbs	Numeric Categoric Categoric Numeric Numeric Categoric	0 0 0 0 0					Unique: 38 Unique: 2 Unique: 4 Unique: 31 Unique: 150 Unique: 2
1 2 3 4 5 6 7	age sex chest_pain trestbps chol fbs restecg	Numeric Categoric Categoric Numeric Numeric Categoric Categoric	0 0 0 0 0 0					Unique: 38 Unique: 2 Unique: 4 Unique: 31 Unique: 150 Unique: 2 Unique: 3
1 2 3 4 5 6 7 8	age sex chest_pain trestbps chol fbs restecg thalach	Numeric Categoric Categoric Numeric Numeric Categoric Categoric Numeric	0 0 0 0 0 0 0					Unique: 38 Unique: 2 Unique: 4 Unique: 31 Unique: 150 Unique: 2 Unique: 3 Unique: 69
1 2 4 5 6 7 8 9	age sex chest_pain trestbps chol fbs restecg thalach exang	Numeric Categoric Categoric Numeric Categoric Categoric Numeric Categoric Categoric	0 0 0 0 0 0 0 0 0					Unique: 38 Unique: 2 Unique: 4 Unique: 31 Unique: 150 Unique: 2 Unique: 3 Unique: 69 Unique: 2
1 2 3 4 5 6 7 8 9 10	age sex chest_pain trestbps chol fbs restecg thalach exang disease	Numeric Categoric Categoric Numeric Numeric Categoric Numeric Categoric Categoric Categoric	0 0 0 0 0 0 0 0 0 0 0 0 0					Unique: 38 Unique: 2 Unique: 4 Unique: 31 Unique: 150 Unique: 2 Unique: 3 Unique: 69 Unique: 2 Unique: 2
1 2 3 4 5 6 7 8 9 10	age sex chest_pain trestbps chol fbs restecg thalach exang disease	Numeric Categoric Categoric Numeric Categoric Categoric Categoric Categoric Categoric Categoric	0 0 0 0 0 0 0 0 0 0 0					Unique: 38 Unique: 2 Unique: 4 Unique: 31 Unique: 150 Unique: 2 Unique: 3 Unique: 69 Unique: 2 Unique: 2

The dataset is loaded. The variable type is automatically detected from the distinct values into each column (discrete or continuous). We can define the TARGET attribute and the INPUT ones. Last, we specify the size of the training (70% of instances, drawn randomly) and test (30%) samples.

3.3 Dataset description

R Data Miner	r - [Rattle (beart for rattle tyt)]
Project Tools	Settings Help Rattle Version 2.5.18 toggware com
Exécuter	Nouveau Ouvrir Enregistrer Report Export Arrêter Quitter
Data Explore	Transform Cluster Associate Model Evaluate Log
Type: 🔘 Sui	🗖 💮 Distributions 🔘 Correlation 💿 Principal Components 🔘 Interactive
Summary	🗌 Describe 🔄 Basics 🔄 Kurtosis 🔄 Skewness 🔄 Show Missing
Below is a	summary of the dataset.
The data is	s limited to the training dataset.
Data Irame:	crssdataset[crsssample,] 200 observations and 10 variables Maximum # MAS:0
	Levels Storage
age	integer
chest pain	4 integer
treathns	integer
chol	integer
fbs	2 integer
restecg	3 integer
thalach	integer
exang	2 integer
disease	2 integer
+	++
Variable	LEVELS
sex	female, male
+	++
chest_pain	n asympt,atyp_angina,non_anginal,typ_angina
+	++
fbs	lf,t
+	laft want human normal at t ware shrarmality
+	ieic_venc_nyper,normal, st_t_wave_abnormality
exang	no, yes
disease	negative, positive
+	++
For the sim	mple distribution tables below the 1st and 3rd Ou.
refer to th	he first and third quartiles, indicating that 25% of the observations have values of
or greater	than (respectively) the value listed.
	any sheat win tweather
Min 100	Sex cnest_pain trestpps cnoi
1et On .42	0.00 remare: 55 dSympt : 61 Min. : 95.0 Min. : 152.0
Median .40	2.00 marc .17/ acyp_angina.// 150 gu.:120.0 130 gu.:211.0
Mean :49	3.27 tvp angina : 8 Mean :133.7 Mean :252.0
3rd Ou.:54	4.00 3rd Qu.:140.0 3rd Ou.:277.5
Max. :65	5.00 Max. :190.0 Max. : 603.0
fbs	restecg thalach exang disease
f:184 le	aft vent hyper : 5 Min. : 82.0 no :142 negative:135
t: 16 no	ormal :157 1st Qu.:122.0 yes: 58 positive: 65
st	t_wave_abnormality: 38 Median :140.0
	Mean :139.6
	3rd Qu.:155.2
	Max. :190.0
Generated h	Dy Rattle 2010-06-15 09:50:20 Maison
	т т
Circle .	Daskasha Cuinat
Find:	Kechercher Suivant
Data summary g	enerated.

Into the Explore tab, we obtain some descriptive statistics indicators about the variables (SUMMARY / SUMMARY option). For the discrete variables, rattle lists the values (levels). For the continuous ones, we have the min, max, mean, quartiles. **All the indicators are computed on the learning sample**.

With the SUMMARY / DESCRIBE option, we obtain a more detailed description. Among others, for the continuous variables, the indications are useful to detect unusual values (outliers).



Into the Explore tab still, with the DISTRIBUTIONS option, we obtain some graphical representations of the distributions. We have for instance the conditional box plots of AGE and CHOL according to the values of DISEASE.

😰 R Data Miner - [Rattle (heart_for_rattle.txt)]		x
<u>P</u> roject <u>T</u> ools <u>S</u> ettings <u>H</u> elp	Rattle Version 2.5.18 togaware.	<u>:om</u>
Exécuter Nouveau Ouvrir Enregistreu Report	t Export Arrêter Quitter	
Data Explore Test Transform Cluster Associate Model Eva	valuate Log	
Type: O Summary O Distributions O Correlation O P	Principal Component: 🔘 Interactive	
Numeric: Effacer Plots per Page: 4 📮 🔲 Annotate	Target: dise	ase
Benford Bars Benford Digit: 1 Benford Bars Attack	Rattle: Plot 2	
Na Variable Bas Dist. Lista area Consulting Bas food . Ma		
1 age 28.	Distribution of age (sample)	ĥ
4 trestbps 92.0		
5 chol 🗸 📄 85.0		
8 thalach 🗌 📄 82.0		=
	All negative positive	
	disease	
	Rattle 2010-Juin-15 10:10:06 Maison	-
< [Distribution of chol (sample)	
Categoric: Effacer		
No. Variable Bar Plot Dot Plot Mosaic Levels		^
2 sex 2 2		
6 fbc 2		
	All negative positive	
9 exang 2	disease	=
10 disease 2	Rattle 2010-juin-15 10:10:07 Maison	
	Euderer Jukuner Teiner	
•	III	
All 2 plots have been generated.		a di

We can obtain also the conditional distribution functions.



About the discrete variables, we can obtain the "Mosaic" of the variables, according still to the values of the target attribute.

R Data Miner - [Rattle (heart_for_rattle.txt)]	
Project Tools Settings Help	Rattle Version 2.5.18 togaware.com
Exécuter Nouveau Ouvrir Enregistrer Kepon	rt Export Arêter Quitter
Data Explore Test Transform Cluster Associate Model E	valuate Log
Type: 🔘 Summary 💿 Distributions 🔘 Correlation 🔘	Principal Components OInteractive
Numeric: Effacer Plots per Page: 4	Target: disease
Benford Bars Benford Digit: 1 🗎 @ abs @ -	Rattle: Plot 2
No Variable Box Blat Histogram Cumulative Benford M	Mosaic of sex (sample)
	by disease
4 trestbps 9	maletemale
5 chol 8	ase e transference e transfer e transference e transfer e transference e transference
8 thalach 🗌 📄 8	
	sex Rattle 2010-juin-15 10:32:29 Maison
	Mosaic of resterg (sample)
Categoric: Effacer	by disease
No. Variable Bar Plot Dot Plot Mosaic Levels	
3 chest pain 4	ase e
6 fbs 2	
7 restecg 🔲 🔽 3	
9 exang 🔲 🔲 2	resteca
10 disease 2	Rattle 2010-juin-15 10:32:29 Maison
	Enregistrer Imprimer Eermer
All 2 plots have been generated.	

For instance, about SEX, the men (MALE) are more numerous than women (FEMALE) into the sample; and the proportion of disease is higher for the men.

We can also obtain the correlations about the continuous input attributes. The correlations are described in a hierarchical structure. It is useful for instance for the detection of the redundant variables.



3.4 Data transformation

The "Transform" tab is dedicated to the variable transformation. Some usual operators are available (e.g. logarithm, rank, etc.).

R Data Miner - [Rattle (heart_for_rattle.txt)]	
Project Tools Settings Help	Rattle Version 2.5.18 togaware.com
🖉 📄 🔚 🗐 🥥 🦂	tter
Data Explore Test Transform Cluster Associate Model Evaluate Log	
Type: Rescale Impute Remap Cleanup	
Normalize: Recenter Scale [0-1] -Median/MAD	Natural Log 💿 Matrix
Order: 💿 Rank 💿 Nolan Groups	
No. Variable Data Type and Number Missing	*
1 age Numeric [28 to 66; unique=38; mean=47; median=49].	
2 sex Categorical [2 levels].	
3 chest_pain Categorical [4 levels].	
4 trestbps Numeric [92 to 200; unique=31; mean=132; median=130].	
5 chol Numeric [85 to 603; unique=150; mean=249; median=248].	
6 fbs Categorical [2 levels].	=
7 restecg Categorical [3 levels].	
8 thalach Numeric [82 to 190; unique=69; mean=139; median=140].	
9 exang Categorical [2 levels].	
10 disease Categorical [2 levels].	
۰ (m	•

3.5 Supervised learning

This step is at the heart of our analysis. We select the "Model" tab. We want to evaluate three methods: decision tree induction, random forest, logistic regression.

R Data Miner - [Rattle (heart_for_rattle.txt)]
Project Tools Settings Help
Secuter Nouveau Ouvrir Enregistrer Report Export Arrêter Quitter
Data Explore Test Transform Cluster Associate Model
Type: Tree Boost Forest SVM Linear Keural Net Survival All
Target: disease (a) Traditional (b) Conditional Model Builder: rpart
Priors: Min Split: 20 🙀 Max Depth: 30 👘 Include Missing
Loss Matrix: Min Bucket: 7 Complexity: 0.0100 Rules Draw
Summary of the Decision Tree model for Classification (built using 'rpart'):
n= 200
node), split, n, loss, yval, (yprob) * denotes terminal node
<pre>1) root 200 65 negative (0.67500000 0.32500000) 2) chest_pain=atyp_angina,non_anginal 111 10 negative (0.90990991 0.09009099) * 3) chest_pain=asympt,typ_angina 89 34 positive (0.38202247 0.61797753) 6) exang=no 42 17 negative (0.59523810 0.40476190) 12) thalach>=131 25 6 negative (0.76000000 0.24000000) * 13) thalach< 131 17 6 positive (0.35294118 0.64705882) * 7) exang=yes 47 9 positive (0.19148936 0.80851064) 14) sev=female 7 3 negative (0.57142857 0.42857143) *</pre>
15) sex=male 40 5 positive (0.12500000 0.87500000) *
١
The corresponding rules have been listed. You may need to scroll the textview to view them.

About the decision tree, rattle uses the rpart command from the rpart package. We note the default parameters used. We click on the EXECUTER button. We obtain the rules associated to the tree by clicking on the RULES button.

Data Explore Test	Transform CI	uster Associate M	odel Evaluate	Log		
Type: 💿 Tree 💿 E	Boost 🔘 Fores	t 🔘 SVM 🔘 Line	ar 🔘 Neural N	et 🔘 Survival 🔘	All	
Target: disease 🔘 T	Fraditional 🔘 🤇	Conditional				Model Builder: rpart
Priors:		Min Split:	20	Max Depth:	30	Include Missing
Loss Matrix:		Min Bucket:	7	Complexity:	0.0100	Rules
Tree as rules: Rule number: chest_pain= exang=yes sex=male Rule number: chest_pain= exang=no thalach< 13 Rule number: chest_pain= exang=yes sex=female	15 [yval=po asympt,typ] 13 [yval=po asympt,typ] 11 14 [yval=no asympt,typ]	ositive cover= _angina ositive cover= _angina egative cover= _angina	=40 (20%) p: =17 (8%) pro	cob=0.88] ob=0.65]		
Rule number: chest_pain= exang=no thalach>=13	12 [yval=ne asympt,typ]	egative cover= _angina	=25 (12%) pi	cob=0.24]		
Rule number: chest_pain=	2 [yval=neo atyp_angina	gative cover=1 a,non_anginal	.11 (56%) pi	cob=0.09]	J	
[1] 9 7 6 3 8	4152					
Generated by R	attle 2010-	-06-15 10:59:1	.6 Maison			

We can obtain also a graphical representation of the tree with the DRAW option.



About the **random forest** approach, rattle uses the randomForest command from the randomForest package. We obtain the following results with the default settings.

R Data Miner - [Rattle (heart_for_rattle.txt)]	
Project <u>T</u> ools <u>S</u> ettings <u>H</u> elp	Rattle Version 2.5.18 togaware.com
🔗 📄 🔚 🗐 🛁 🧐 Exécuter Nouveau 🗗 Enregistrer Report Export Arrêter Quit	ter
Data Explore Test Transf Cluster Associate Model Evaluate Log	
Type: Tree Boost Forest SVM Linear Neural Net Survival All	
Target: disease 💿 Traditional 🔘 Conditional	Model Builder: randomForest
Number of Trees: 500 🛓 Sample Size: Importance Rules	
Number of Variables: 3	
Summary of the Random Forest model:	<u>^</u>
Call: randomForest(formula = disease ~ ., data = crs\$dataset[crs\$sample, Type of random forest: classification Number of trees: 500 No. of variables tried at each split: 3], ntree = 500, mtry =
Confusion matrix:	
negative positive class.error	
negative 123 12 0.08888889	
positive 21 44 0.32307692	
Variable Importance	
negative positive MeanDecreaseAccuracy MeanDecreaseGini	
chest pain 1.77 3.80 2.17 19.89	=
thalach 0.70 2.46 1.40 14.69	
exang 1.68 2.89 1.85 13.66	
chol 0.66 0.54 0.56 12.60	
age -0.07 0.39 0.11 9.78	
trestbps -0.46 0.67 -0.02 7.44	
sex 0.33 2.82 1.44 4.36	
restecg 0.36 -0.73 -0.12 1.93	
Display the Model	
To view model 5, for example, execute the command printRandomForests(crs\$rf, 5) in the R console. Generating all models will take quite some time.	
Time taken: 0.42 secs	
Generated by Rattle 2010-06-15 13:19:14 Maison	
< []	· •
The Random Forest model has been built. Time taken: 0.42 secs	

The OOB (out-of-bag) error estimation is 16.5%. We will compare this value to the one obtained on the test set below.

About the logistic regression, we use the glm() command. It automatically transforms the discrete predictors using dummy variables. We obtain the following results.

R Data Miner - [Rattle (heart_for_rattle	e.txt)]				x
<u>P</u> roject <u>T</u> ools <u>S</u> ettings <u>H</u> elp			0	Rattle Version 2.5.18 <u>togaware</u>	e.com
Exécuter Nouveau Ouvrir	Enregistrer Report	Export Arrêter	Quitter		
Data Explore Test Transform Cluste	Associate Model Evalua	ate Log			
Type: 🔘 Tree 🔘 Boost 🔘 Forest 🔘) SVM 💿 Linear 🔘 Neu	ral Net 🔘 Survival 🔘 A	JI		
Numeric Generalized Poiss	on 💿 Logistic 🔘 Probit	t 💿 Multinomial		Model Builder: glm (logit)	
Plot					
Summary of the Logistic Reg	ression model (buil	t using glm):			~
Call: glm(formula = disease ~ ., data = crs\$dataset[crs\$	family = binomial(1 sample,])	link = "logit"),			
Deviance Residuals:					
Min 1Q Median	3Q Max				
-2.5490 -0.4750 -0.2661	0.4889 2.8434				
Coefficients:					
	Estimate Std. Er	ror z value Pr(>	z)		
(Intercept)	0.294096 4.036	0.073 0.94	1916	1	Ξ
age	-0.007997 0.033	8876 -0.236 0.81	3390		
sexmale	1.189818 0.587	423 2.025 0.04	2818 *		
chest painatyp angina	-2.186895 0.575	810 -3.798 0.00	0146 ***		
chest_painnon_anginal	-1.586384 0.627	181 -2.529 0.01	1426 *		
chest paintyp angina	0.517539 0.923	208 0.561 0.57	5079		
trestbps	-0.004770 0.012	892 -0.370 0.71	1408		
chol	0.005226 0.003	447 1.516 0.12	9502		
fbst	1.251407 0.833	696 1.501 0.13	3346		
restecgnormal	1.098985 2.277	026 0.483 0.62	9351		
restecgst t wave abnormalit	y 0.426301 2.314	418 0.184 0.85	3862		
thalach	-0.023186 0.011	624 -1.995 0.04	6071 *		
exangyes	1.874358 0.494	388 3.791 0.00	0150 ***		
 Signif. codes: 0 '***' 0.0	01 '**' 0.01 '*' 0.	05 '.' 0.1 ' ' 1		•	
(Dispersion parameter for b	inomial family take	en to be 1)			
Null deviance: 252.23 Residual deviance: 144.08 AIC: 170.08	on 199 degrees of on 187 degrees of	freedom freedom			
Number of Fisher Scoring it	erations: 6				
Log likelihood: -72.039 (13 Null/Residual deviance diff Chi-square p-value: 0.00000 Pseudo R-Square (optimistic	df) erence: 108.153 (12 000): 0.70482172	? df)			-
		III			F

3.6 Measuring the generalization performance

Last step of our analysis, we want to evaluate the performances of the classifiers on the test sample (30% of the whole dataset).

We activate the "Evaluate" tab. First, we want to obtain the confusion matrix and the associated error rate. We select the "Error Matrix" option. For the "Data" item, we must select the "Testing" option. Only the models learned into the "Model" tab are available here.

We click on the EXECUTER menu. We observe that the logistic regression is the better here with a test error rate equal to 18.18%.

We note also that the OOB error rate (16.5%) seems underestimate the error rate for the random forest (20.45% on the test set). But, because the test set size is small, and the test error rate being also an estimation of the "true" error rate, we consider with many cautions this result.

@ R Data Miner - [Rattle (heart_for_rattle.txt)]
Project Tools Settings Help
👸 📄 🔚 🗐 🥖 Kaport Superior
Data Explore Test Transform Cluster Associate Model Evaluate Log
Type: 💿 Error Matrix 🔘 Risk 🔘 Cost Curve 🔘 Hand 🔘 Lift 🔘 ROC 🔘 Precision 🔘 Sensitivity 🔘 Pr v Ob 🔘 Score
Model: 🖉 Tree 🗌 Boost 🖉 Forest 🗍 SVM 🖉 Linear 🗌 Neural Net 🗍 Survival 🗍 KMeans 🗍 HClust
Data: 🔘 Training 🔘 Validation 💿 Testing 🔘 CSV File 🛛 (Aucun) 🛛 📄 🔘 R Dataset 📃 🗸
Risk Variable: Report: Class Probability Include: Identifiers All
Error matrix for the Decision Tree model on heart_for_rattle.txt [test] (counts):
Actual Predicted negative positive negative 23 6 positive 3 12
Error matrix for the Decision Tree model on heart_for_rattle.txt [test] (%):
Actual Predicted negative positive negative 52 14 positive 7 27
Overall error: 0.2045455
Generated by Rattle 2010-06-15 13:48:39 Maison
Error matrix for the Random Forest model on heart_for_rattle.txt [test] (counts):
Actual Predicted negative positive negative 25 8 positive 1 10 Error matrix for the Random Forest model on heart_for_rattle.txt [test] (%):
Actual
Predicted negative positive
positive 2 23
Overall error: 0.2045455
Generated by Rattle 2010-06-15 13:48:39 Maison
Error matrix for the Linear model on heart_for_rattle.txt [test] (counts):
Actual Predicted negative positive negative 24 6 positive 2 12
Error matrix for the Linear model on heart_for_rattle.txt [test] (%):
Actual Predicted negative positive negative 55 14 positive 5 27 Overall error: 0.1818182 Generated by Rattle 2010-06-15 13:48:39 Maison
······································
Generated Error Matrix.

Actually, the error rate is not a good criterion here. We note that the differences between the methods are based only on one misclassified instance. In our context, it is perhaps more interesting to use the ROC curve which highlights the ability of the methods to assign higher score to the positive instances compared with the negative ones (see http://data-mining-tutorials.blogspot.com/2008/11/roc-curve-for-classifier-comparison.html or http://data-mining-tutorials.blogspot.com/2008/11/roc-curve-for-classifier-comparison.html or http://data-mining-tutorials.blogspot.com/2008/11/roc-curve-for-classifier-comparison.html or http://data-mining-tutorials.blogspot.com/2008/10/computing-roc-curve.html).

We select the ROC option under rattle.



According the AUC criterion, the decision tree is definitely the worst compared with the two other classifiers, which are similar in terms of performance. It is not surprising. We know that the decision tree is not well adapted to the <u>scoring</u> process.

3.7 R commands associated to the treatments

Image: R Data Miner - [Rattle (heart_for_rattle.txt)]
Project Tools Settings Help Rattle Version 2.5.18 togaware.com
Exécuter Nouveau Ouvrir Enregistrer Report Export Arrêter Quitter
Data Explore Test Transform Cluster Associate Model Evaluate Log
Export Comments Rename Rattle Variables: From crsS to MY
 # little effort the log can be used to score a new dataset. The logical variable # 'building' is used to toggle between generating transformations, as when building # a model, and simply using the transformations, as when scoring a dataset.
building <- TRUE scoring <- ! building
The colorspace package is used to generate the colours used in plots, if available.
library(colorspace)
Rattle timestamp: 2010-06-15 14:34:33 i386-pc-mingw32
Load the data.
crs\$dataset <- read.csv("file:///D:/DataMining/Databases_for_mining/benchmark_datasets/Heart/heart_for_rattle.txt", sep="\t", na.strings=c(".", '
#=====================================
▼ M_4_4_46

One of the main criticisms which we make for the software driven by menu is that once the process is finalized, when we close the software, we have no recollection of the sequence of operations we performed. In the next working session, it is complicated to reproduce them as before. It is necessary to have an excellent memory, or to have taken care of noting all that we made.

Rattle allows to overtake this drawback by translating all the operations (corresponding to a click on the EXECUTER menu) performed by the user in a sequence of R commands. We can visualize them in the "Log" tab. We can store these commands (and the comments) into a file. In the next working session, it is very easy to perform the same data processing by loading these commands.

4 Rattle under Linux (Ubuntu)

The installation of the Rattle package under Linux is not easy. We must follow carefully the description available on the website. In case of problem, a troubleshooting procedure is proposed. This is the one that I used (see http://datamining.togaware.com/survivor/Install_GNU_Linux.html).





5 Conclusion

In this tutorial, we showed that it was possible to use R without knowledge about its programming language with the help of the rattle package. This package is rather specialized about the data mining methods. For the statisticians, there are other packages such as "<u>R Commander</u>".