# 1 Subject

# Partitioning the dataset in a learning set and a test set. Learning a cost sensitive decision tree.

Error rate evaluation is a key point of the induction process. A usual approach is to partition the dataset in a learning set, which is used for the induction of the classification model, and in a test set, which is used for the performance evaluation. The first subject of this tutorial is to show how to make a partition of the dataset with SIPINA. Then, we build the tree on the first part of the dataset. Later, we classify the examples of the second part of the dataset. We compare the predicted value and the true value. We obtain honest error rate estimation.

The second main subject of this document is to show how to take into account the misclassification costs during the learning process and the evaluation process. We use a slightly modified version of  $C4.5^{1}$ .

# 2 Dataset

We handle a spam detection problem. The learning sample contains 3601 examples; there are 1000 examples for the test set. The goal is to determine if a message is a spam or not, using their characteristics (frequencies of word, length of message, etc.) <sup>2</sup>. An additional column enables to know if the example belongs to the learning or the test sample.

# **3** Tree induction with the C4.5 algorithm

### 3.1 Data importation

The easiest way to import a dataset into SIPINA is to open the file into EXCEL spreadsheet. Using the SIPINA.XLA add-in, we can send directly the selected dataset to SIPINA.

**Note:** For the add-in installation and utilization, see the tutorial "SIPINA ADD-IN FOR EXCEL SPREADSHEET" (<u>http://eric.univ-lyon2.fr/~ricco/sipina\_download.html</u>).

We select the cells corresponding to the dataset into EXCEL. We click on the SIPINA/EXECUTE SIPINA menu. A dialog box appears. We check the cells selection and we validate. The first row of the selection must correspond to the name of the variables (Figure 1).

SIPINA is automatically launched. The dataset is downloaded. We check in the status bar the number of columns (57) and the number of rows (4601).

**Note:** We can edit the dataset into the editor of SIPINA. But this is not really interesting. EXCEL offers more powerful functionalities about data manipulation.

<sup>&</sup>lt;sup>1</sup>See <u>http://www.informatik.uni-freiburg.de/~ml/ecmlpkdd/WS-Proceedings/w10/chauchat\_workshop.pdf</u>

<sup>&</sup>lt;sup>2</sup> <u>http://eric.univ-lyon2.fr/~ricco/dataset/spam.xls</u>; the original database is available on the following website <u>http://archive.ics.uci.edu/beta/datasets.html</u> (spambase)

<b>N</b>	Aicros	oft Exce	l - spam.xls									×
8	Eichier	<u>E</u> dition	Affichage	Insertion	Forma	<u>t O</u> utils	<u>D</u> onnées F	e <u>n</u> être	<u>?</u> Tanagra	Sipina	- 8	x
5	»	Bitstream	Vera Sans 🗸	8 -	G Z	s≣	= = =	9€	% 000 5	s8 Exe	cute Sipina	×
	A1	-	<i>f</i> ∗ wf	f_make						Abo	ut	
	1	AZ 🗌	BA	BB		BC	BD		BE	BF	BG	
1	cf #		capital run le	capital r	un le ca	pital run	e spam	statu	IS I			-
2	-	0	2.22		20	13	l yes	learr	ning			
3		0	1.181		5	2	6 no	learr	ning			
4		0	1.142		4	5	6 no	learr	ning			
5		0	3.212		44	66	5 no	learr	ning			
6		0	1		1		3 no	learr	ning			
7		0.122	4		36	26	4 yes	learr	ning			
8		0	1.3		3	1	3 no	learr	ning 🔓	ast col	umn denotes ti	i
9		0	1.721	Execute Si	pina					X		
10	}	0	1.85							ext to	last column of	
11		0.05	4.623	Datas	et range (ind	luding the name	of the attributes	first row):		nercia	l e-mail. Most o	
12		0	2.166	\$A\$	1:\$BE\$4602	2			-	un-len	gth attributes (	
13		0	1.59					1		attribu	ite, see the end	2
14		0.03	3.774				OK	Ca	ncel			
15		0.132	6.404							ntinuo	us real [0,100]	
10		0	4.111		20	22	2 yes	learr	ing	= percenta	ge of words in t	<u>t</u>
10		0	1 627		10		5 110	learr	ing c	of words in	e-mail. A "word	2
10		0	1.657		17	51	1 no	learr	ing e	end-of-strir	ig.	
20		0	2.407		1/	9	/ 110	learr	ing			
20		0	1.419		20	4	4 no	learr	ning e	5 continuou	s real [0,100] a	•
22		0	2 701		20	18	1 00	learr	ing	= percenta	ge of character	1
23		0	1 997		87	162	0 00	learr	ning			
24		0	2 337		41	50	5 ves	learr	ning	continuou	s real [1,] att	ł
25		0	4.8		19	2	4 no	learr	ning	= average l	ength of uninte	-
26		0	1.375		4	1	1 no	learr	ning			
27		0	4.333		11	10	4 yes	learr	ning	Lontinuou	s integer [1,]	-
H 4	і <b>н</b> н	spamt	oase /				-	•	- 1=	= lendth of	Iongest Uninter	Г
Deg	sin 🕶 🕻	> Forme	es automatiques	• \ )		) 🖾 ᆀ	ः 🛛 🔜	🖉 -	<u>/</u> - <u>A</u> -	= = =	‡ 🗖 🖉 ₊	
Sélec	tionnez	une desti	nation et appuy	ez sur ENT	RÉE ou c	liquez sur Co	ller S	omme=160	08977.758	NL	M	

Figure 1 - Transferring the dataset from EXCEL to SIPINA

# 3.2 Selecting the learning method – C4.5

Now we have to select the learning method. We click on the INDUCTION METHOD / STANDARD ALGORITHM menu. A dialog box which describes the available methods appears. We select the C4.5 approach (QUINLAN, 1993).

🕂 Sipina Research V	ersion - [Le	arning set	editor]						(		
🔝 File Edit Data Sta	atistics Induc	tion method	Analysis Vi	ew Window	Help					- 8 ×	
🗅 👸 🕞 📥 🌬	🖸 🛍 🛤 🛤 Standard algorithm										
	<u>×</u>		wf_make	wf_address	wf_all	wf_3d	wf_our	wf_over	wf_remove	wf_interi	
Attribute selection		1	0.00	0.52	62	0.00	0.52	0.00	0.00	0.00	
		2	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	
		3	0.00	0.00	0.66	0.00	0.00	0.66	0.00	0.00	
		4	0.08	0.00	0.16	0.00	0.00	0.08	0.00	0.08	
	Select an indu	ction methor							00	0.00	
	Jacket an mut	enon marine		. Let		i rilou I	1		47	0.00	
	_ 00	0.00									
	00	0.00									
	00	0.00									
	ChAID (Kass - 19	Cestnik, Kononer 80) 🥏	NKO & Bratko - I	986)					06	0.33	
	00	0.00									
	00	0.00									
	00	0.00									
	One-varial Decisi	0111166							41	0.82	
									00	0.00	
-									_00	0.00	
				C4.5 (Quin	lan - 1993)			-	_00	0.00	
							V CKM	🗙 Annule	1 00	0.00	
-		13	0.00	0.00	0.00	0.00	0.00		00.00	0.00	
		20	0.00	0.00	0.00	0.00	0.00	00	0.00	0.00	
		21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
J		22	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.20	
Learning method		23	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	
MethodName=Improved C	hAID (Tsc 🔺	24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
MethodUlassName=TArbre	eDecision	25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Merge=0.05	=	26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Split=0.001		27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
LypeBonterroni=1		28	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
Sampling=0	~	29	0.67	0.00	0.00	0.00	1.01	0.00	0.00	0.00	
Examples selection		30	0.00	0.00	0.00	0.00	0.64	0.00	0.64	0.00	
4601 examples selected 31 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0											
0 examples idle											
Editing NEW.FDM Attributes : 57 Examples : 4601											
Improved ChAID (Tschupro	ow Goodness of	Split)								1	

When we have selected the learning method, another dialog box enables us to specify the algorithm parameters. We use the default parameters. We validate by clicking OK.

C4.5 paramet	ers  🛛
C.L. for pessimi	stic pruning
	25 🚖
Size of leaves :	2
<ul> <li>Sampling</li> <li>All dataset</li> </ul>	
C Simple	Size :
C Balanced	300
Judulieu	
	🗸 08💬

We choose the C4.5 algorithm because this is a state-of-the-art method. An improved version which takes into account misclassification costs is also available into SIPINA

The parameters appear in the middle part of the project explorer (Figure 2).

🔊 Sipina Research Version - FLe	earning set	editor]							
S File Edit Data Statistics Indu	tion method	Analysis Vie	w Window	Heln					a x
×		wf_make	wf_address	wf_all	w/f_3d	wf_our	wf_over	wf_remove	wf 🔨
Attribute selection	1	0.00	0.52	0.52	0.00	0.52	0.00	0.00	0.0
	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
	3	0.00	0.00	0.66	0.00	0.00	0.66	0.00	0.0
	4	0.08	0.00	0.16	0.00	0.00	0.08	0.00	0.0
	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
	6	0.00	0.36	0.00	0.00	0.00	0.36	1.47	0.0
	7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
	8	0.00	0.00	1.85	0.00	0.00	0.00	0.00	0.0
	9	0.00	0.00	0.53	0.00	0.00	0.53	0.00	0.0
	10	0.40	0.40	0.26	0.00	0.13	0.20	0.06	0.3
	11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
	12	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.0
	13	0.67	0.16	1.35	0.00	1.01	0.00	0.00	0.0
	14	0.82	0.00	0.82	0.00	0.41	0.00	0.41	0.8
	15	0.00	0.00	0.78	0.00	1.17	0.00	0.00	0.0
J	16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
Learning method	17	0.00	0.00	0.11	0.00	0.11	0.00	0.00	0.0
MethodName=U4.5 [Uumlan - 1993] MethodClassName=TArbraDesisionC45	18	0.66	0.66	0.66	0.00	0.00	0.00	0.00	0.0
Hdl=5	19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
Confidence level=25	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
Leaf size=2	21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
Sampling=0 SamplingPart=300	22	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.2
	23	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.0
Examples selection	24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0
4601 examples selected	25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0 💙
0 examples idle	<								>
1	Editing N	EW.FDM				Attribut	tes : 57	Examples : 46	501
C4.5 (Quinlan - 1993)									

Figure 2 – C4.5 is the selected method

### 3.3 Defining the class attribute

We click on the ANALYSIS / DEFINE CLASS ATTRIBUTE menu in order to define the class attribute (SPAM) and the predictive attributes (the other ones, excepting STATUS). In the dialog box which appears, using drag and drop, we do the selection.

🔉 Sipina Research Version - [Le	arning set	editor]								
🔉 File Edit Data Statistics Induc	tion method:	Analysis V	iew	Window	Help				-	Ξ×
15 🏔 📴 🐜 🖿		Define da	ass at	tribute						
		Select ac	tive e	xamples		,				
X						wf_3d	wf_our	wf_over	wf_remove	wf 🔨
Attribute selection	1	Set weigh	nt field	d		0.00	0.52	0.00	0.00	0.0
	2	Set priors	5			0.00	0.00	0.00	0.00	0.0
	3	Set costs				0.00	0,00	0.66	0.00	0.0
	4	Set positi	ve cla	ass value.		0.00	0.00	0.08	0.00	0.0
	5	Loorping				0.00	0.00	0.00	0.00	0.0
	6	charming.	in in the second se			0.00	0.00	0.36	1.47	0.0
	7	Stop ana	iysis			0.00	0.00	0.00	0.00	0.0
	8	Classifica	tion	Attribute	selection					
	6	Test		Classic	_			- Variables		
	10		_	Liass				wf_pm		
	10	LIFT R	OC d	Japan	-			wf_direct		
	10	Error mo		- Attributes				wf_meeting		
	12	Errormea	sure	wf_make	1		^	wf_original wf_project		
	13	Feature s	selec	wf_addre	188			wf_re		
	14			wf_3d		_		wf_table		
	15	Personna	l tes	wf_over				wf_conferenc		
	16	0.00	0.0	wf_remov	/e ·	<b>-</b>		cf_(		
Learning method	17	0.00	0.0	wf_order				ct_l cf_!		
MethodDassName=C4.5 (Quinian - 1993) MethodDassName=TArbreDecisionC45	18	0.66	0.6	wf_mail wf_receiv	/e		✓	cf_\$		
Hdl=5	19	0.00	0.0					capital_run_le	ngth_averag	
Confidence level=25	20	0.00	0.0	C Only	y discrete			capital_run_le	ngth_longes noth_total	
Leaf size=2	21	0.00	0.0	C Only	y continuous			spam		_
Sampling=U SamplingPart=200	22	0.00	0.1	@ Bot	h			status		~
SampingFatt=300	23	0.00	0.0						1	
Fyamples selection	24	0.00	0.0							Annuler
4601 evemples selected	25	0.00	0.0	n	0.00	0.00	0.00	0.00	0.00	0.0 -
0 examples idle		10.00	0.00	-	0.00	10.00	10.00	0.00	0.00	>
	Edition						Attack	utor i 57	Examples : 4	601
	jealang (h		_				ACCID	ules : 37	Lixamples : 4	JO1 //
C4:2 (Quinian - 1993)										11.

Variable selection appears in the top part of the project explorer.

🛿 Sipina Research Version - [Learning set editor]										
💦 File Edit Data Statistics Indu	uction method	Analysis Vi	ew Window	Help				-	Ξ×	
🖰 🗎 🐂 🐂										
12	A	wf_make	wf_address	wf_all	wf_3d	wf_our	wf_over	wf_remove	wf 🔨	
Attribute selection	_ 1	0.00	0.52	0.52	0.00	0.52	0.00	0.00	0.0	
🖃 📭 Class attribute 🛛 🔼	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	
spam	3	0.00	0.00	0.66	0.00	0.00	0.66	0.00	0.0	
Predictive attributes	4	0.08	0.00	0.16	0.00	0.00	0.08	0.00	0.0	
u ut_make	5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	
wt_address	6	0.00	0.36	0.00	0.00	0.00	0.36	1.47	0.0	
		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	
wr_3d	8	0.00	0.00	1.85	0.00	0.00	0.00	0.00	0.0	
wi_our	9	0.00	0.00	0.53	0.00	0.00	0.53	0.00	0.0	
wi remove	10	0.40	0.40	0.26	0.00	0.13	0.20	0.06	0.3	
wf_internet	11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	
wf order	12	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.0	
wf mail	13	0.67	0.16	1.35	0.00	1.01	0.00	0.00	0.0	
🗖 wf receive 🧔	14	0.82	0.00	0.82	0.00	0.41	0.00	0.41	0.8	
	15	0.00	0.00	0.78	0.00	1.17	0.00	0.00	0.0	
	16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	
Learning method	17	0.00	0.00	0.11	0.00	0.11	0.00	0.00	0.0	
MethodName=C4.5 (Quinlan - 1993)	18	0.66	0.66	0.66	0.00	0.00	0.00	0.00	0.0	
MethodClassName=TArbreDecisionC45	19		0.00	0.00	0.00	0.00	0.00	0.00	0.0	
Confidence level=25	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	
Leaf size=2	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	
Sampling=0	22	-0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	
SamplingPart=300	23	0.00	0.10	0.00	0.00	0.00	0.00	3.00	0.2	
F La calcation	23	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.0	
Examples selection	- 24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.0	
0 examples idle	<	10.00	0.00	0.00	0.00	0.00	0.00	0.00	>	
	Editing 1	NEW.FDM				Attrib	utes : 57	Examples : 46	501 /	
C4 5 (Ouiplan - 1993)										

The acronym "D" is used for discrete variable ("C" for continuous attribute).

### 3.4 Partitioning the dataset into learning and test sets

We use the additional column (STATUS) for this operation. We click on the ANALYSIS / SELECT ACTIVE EXAMPLES menu. A dialog box appears. We select the RULE SELECTION option.

🔉 Sipina Research Version										
File Edit Data Statistics Induct	ion method	Analysis View Window	/ Help							
🏹 🚯 📴 🦏 🖪		Define class attribute.								
	X	Select active example:	i 🗕	_	-					
Attribute selection	- 🔊 Le	Set weight field								
🖃 🕼 Class attribute		Set priors		out all	wet 3d	last our	wf over	wf remove	w/f_internet	wrf. orde
D spam	1	Set costs		1.52	0.00	1.52	0.00	0.00	0.00	0.00
🖻 🗊 Predictive attributes	2	Set positive class valu	B	100	0.00	0.00	0.00	0.00	0.00	0.00
-C wf_make	3			).66	0.00	0.00	0.66	0.00	0.00	0.00
	4	Learning		1.16	0.00	0.00	0.08	0.00	0.08	0.73
wf_all	5	Stop analysis		100	0.00	0.00	0.00	0.00	0.00	0.00
wf_3d	6	Classification	•	100	0.00	0.00	0.36	1 47	0.00	0.00
wf_our	7	Test		100	0.00	0.00	0.00	0.00	0.00	0.00
wt_over	8			85	0.00	0.00	0.00	0.00	0.00	0.00
wr_remove	8	LIFT ROC curve								
wf_order	10	Filter active examples								
wf mail	11	Luf make		1	- Introduce a	numerical value				
wf receive	12	wf_address	^	<	Introduce o	Thanking a value	, ,	- 11		
e of will	13	wf_all		<=						
wf neonle	14	wr_3a wfour		=	Pick one or	several values				
wf report	15	wf_over		>=				- 11		
wf_addresses	16	wf_internet		>						
wf_free	17	wf_order		$\diamond$						
C wf_business	18	wf_receive		,						
C wf_email	19	- wf_will		lin						
<	20	wf_report		notin	Generate p	roposition				
I a surface sure the set	21	wf_addresses	~		And	C Or	valuate			
MethodName-C4 5 (Quislan, 1992)	- 22	-								
MethodClassName=TArbreDecisionC	45 23									
Hdl=5	24									
Confidence level=25	25							Filt	er method	
Leaf size=2 Sampling=0	20							0	All	
SamplingPart=300	20							0	List	
J	27								mandom samplir Rule celection	
Examples selection	20	-							TOP SEIECUONS	W.
	29	List of examples Rand	iom samp	ling Ru	e filter				🖊 ОК 🛛 🗙	Anneler
	30	· · · · · · · · · · · · · · · · · · ·						····		<u> </u>
1										>
C4.5 (Quinlan - 1993)										

We intend to specify that the examples which are STATUS = LEARNING correspond to the learning sample. We work first in the top part of the window. We have to do the following steps: (1) We select the STATUS attribute in the list box. (2) We select the IN condition. (3) Then, we select the LEARNING value in the left part of the window.

We validate this rule selection by clicking the VALIDATE button. The rule is available in the bottom part of the window.

Filter active examples	
wf_re wf_edu wf_table wf_conference df_; cf_f cf_i cf_i cf_i cf_i cf_i capita_run_length_averag capita_run_length_longes capita_run_length_total spam status       Introduce a numerical value         Pick one or several values       Image: Conference cf_i cf_i cf_i capita_run_length_averag capita_run_length_total spam status       Image: Conference capita_run_length_total spam capita_run_length_total spam       Image: Conference capita_run_length_total spam       Image: Conference capita_run_length_total spam	Filter method C All C List C Random sampling C Rule selection
List of examples Random sampling Rule filter	🗸 OK 🛛 🗶 Annuler

We can count the covered examples by using the contextual menu. We observe that 3601 examples are available for the learning process.



We click the OK button. The selection is displayed in the bottom part of the project explorer.

Sipina Research Version										
File Edit Data Statistics Inducti	ion n	nethod Analy	sis View '	Window Help						
🖸 🖄 🛤 🐚										
	×									<u>^</u>
Attribute selection Attribute selection										
🖃 📭 Class attribute 🖉			wf_make	wf_address	wf_all	wf_3d	wf_our	wf_over	wf_remove	wf_interner
📃 🕒 spam		1	0.00	0.52	0.52	0.00	0.52	0.00	0.00	0.00
Predictive attributes		2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
wf_make		3	0.00	0.00	0.66	0.00	0.00	0.66	0.00	0.00
wf_address	-11	4	0.08	0.00	0.16	0.00	0.00	0.08	0.00	0.08
wt_all		5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00 =
wt_3d		6	0.00	0.36	0.00	0.00	0.00	0.36	1.47	0.00
wr_our		7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
wi_over		8	0.00	0.00	1.85	0.00	0.00	0.00	0.00	0.00
wf_internet		9	0.00	0.00	0.53	0.00	0.00	0.53	0.00	0.00
wf_order		10	0.40	0.40	0.26	0.00	0.13	0.20	0.06	0.33
wf mail		11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
wf receive		12	0.00	0.00	0.00	0.00	0.30	0.00	0.00	0.00
- C wf will		13	0.67	0.16	1.35	0.00	1.01	0.00	0.00	0.00
	-	14	0.82	0.00	0.82	0.00	0.41	0.00	0.41	0.82
		15	0.00	0.00	0.78	0.00	1.17	0.00	0.00	0.00
Learning method		16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
MethodName=C4.5 (Quinlan - 1993)	-	17	0.00	0.00	0.11	0.00	0.11	0.00	0.00	0.00
Hdl=5	io	18	0.66	0.66	0.66	0.00	0.00	0.00	0.00	0.00
Confidence level=25		19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Leaf size=2		20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sampling=U SamplingPart=300		21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Sampingrak-300		22	0.00	0.10	0.00	0.00	0.00	0.00	0.00	0.20
Fxamples selection		23	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00
3601 examples selected		24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1000 examples idle		25	0		0.00	0.00	0.00	0.00	0.00	0.00 🔽
	•[[	<	•							>
C4.5 (Ouinlan - 1993)										

### 3.5 Decision tree learning

We start the learning phase by clicking the ANALYSIS / LEARNING menu. The window seems strangely empty. This is because the tree is very large. We must use the TREE MANAGEMENT / ZOOM OUT menu in order to show the whole tree.

Stpina Research Version		
Induction method Analysis Tree manage	ement View Window Help	
🗅 👸 🚉 🍖 🐚		
X		
Attribute selection	Decision tree	
🖃 🗊 Class attribute 🛛 🔨		
D spam	1	
Predictive attributes		
🔤 wf_make		
wf_address		
WI_all =		
wi_ou		
wf over		
- C wf remove		
- C wf_internet		
- C wf_order		
🚽 🚾 wf_mail		
wf_receive		
wt_will		
wi_people		
wf addresses		
Learning method		
MethodName=C4.5 (Quinlan - 1993)		
MethodClassName=TArbreDecisionC45		
Hdl=5 Confidence level=25		1
Leaf size=2		
Sampling=0		•
SamplingPart=300		1
Evamples selection		
3601 examples selected		•
1000 examples idle		N 💌
	(	>
C4.5 (Quinlan - 1993)		Time : 1344

We can use also the shortcut CTRL + Q.



The tree is very large. We cannot really extract useful knowledge from these rules. But it is not the goal in this tutorial. We want only to elaborate an accurate classifier.

### 3.6 Error rate measurement

#### 3.6.1 Confusion matrix and resubstitution error rate

First, we want to compare the true and the predicted values of the class attribute on the learning set. We obtain a confusion matrix and the resubstitution error rate.

In order to obtain this matrix, we activate the ANALYSIS /TEST menu. In the dialog box, we select first the learning set.

🔉 Sipina Research	Version - [Decision tree	.]		
🔉 Induction method	Analysis Tree management N	iew Window	Help	_ 8 ×
🕂 🖳 👹 📦 🖻	Define class attribute Select active examples			
Attribute selection	Set weight field Set priors Set costs Set positive class value Learning Stop analysis Classification Test LIFT ROC curve Error measurements	>	Apply classifier on Apply on © Learning set © Inactive examples of Databases	
Learning method MethodName=C4.5 (Qi MethodClassName=TArt Hdl=5 Confidence level=25 Leaf size=2 Sampling=0 SamplingPart=300 Examples selection 3601 examples selected 1000 examples idle	Personal tests			
, C4.5 (Quinlan - 1993)	NU			e : 1344

The confusion matrix and the error rate are displayed in a new window (Figure 3). We read:

- 80 "spam" are classified as regular messages;
- 54 regular messages are classified as "spam";
- The other messages are properly classified. (1342 "spam", 2125 regular messages).

In the following confusion matrix:

Row : true values	Positive	Negative
Column : predicted values	prediction	prediction
+ (« positive » is "spam")	а	b
-	С	d

We compute the error rate as follows:  $e = \frac{b+c}{a+b+c+d}$ 



Figure 3 – Confusion matrix and resubstitution error rate

The resubstitution error rate is

$$e_r = \frac{80 + 54}{3601} = 3.7\%$$

The classification tree seems accurate. But, this result is not really credible. We use the same dataset in order to learn and evaluate the model. The resulting error rate is often optimist, especially about the decision tree with many rules (Figure 4).

ew Window Help	
5	
	~
visualization F5	
Ctrl+C	
Ctrl+W	=
Ctrl+Q	
tation Ctrl+A Nodes : 147	
ntation Ctrl+Z Veaves : 74 Max depth :16	
t file	
114	
	24 24 14
	w Window Help

Figure 4 – Tree characteristics

#### 3.6.2 Confusion matrix and test error rate

In order to obtain an unbiased error rate evaluation, we must use the test set.

Again, we click on the ANALYSIS / TEST menu. But we choose now the INACTIVE EXAMPLES OF DATABASES option.

🔉 Sipina Research	Version - [Decision tree]		
🔉 Induction method	Analysis Tree management Vie	w Window Help	_ 8 ×
<mark>B 🗄 📑 🛉 🖻</mark>	Define class attribute Select active examples		
Attribute selection Class attribute Predictive attribute Momentary Class attribute Predictive attribute Momentary Class attribute Predictive attribute Momentary Momen	Set weight field Set priors Set costs Set positive class value Learning Stop analysis Classification Test LIFT ROC curve	Apply classifier on X Apply on C Learning set Inactive examples of Databases	
wif order	Error measurements	V OK X Annuler	5 0.00 5 0.00 7 0.00 7 0.00
Learning method MethodName=C4.5 (Q) MethodClassName=TA Hdl=5 Confidence level=25 Leaf size=2 Sampling=0 SamplingPart=300 Examples selection 3601 examples selected 1000 examples idle	Personnal tests		
(0.5 (0.00)			e : 1344

If we compute the sum of the values of the matrix, we have 1000, the test sample size.

🔉 Sipina Research Version				
Induction method Analysis View Windo	ow Help			
🖰 🗃 📲 h 📭				
	S. Decision to	ee		· •
⊡-¶‡⇒ Class attribute D spam	🔉 Confusi	on matrix : Test	set on D: \Data 🔳	
🖃 🗊 Predictive attributes 🛛 🔍 🎵	spam —			
		yes	no	
	yes	332	59	
Learning method	no	27	582	
MethodName=C4.5 (Quinlan - 1993) MethodClassName=TArbreDecisionC45 Hdl=5				/
Confidence level=25				
Leaf size=2	Cost : 0.0860	E.		111
Sampling=U SamplingPart=200				1
Sampingrak-300				
Examples selection				
3601 examples selected				Ì
1000 examples idle			1.241	`
	( )			>
C4.5 (Quinlan - 1993)				

The test error rate is

$$e_t = \frac{59 + 27}{1000} = 8.6\%$$

The true probability of misclassification of the tree on unseen cases is 8.6%.

#### 3.6.3 Other indicators: recall, precision and F-Measure

In our study and it is the case very often, the two values of the class attribute has not the same importance, nor the same significance. Predicting a spam as a normal message when the message is automatically filtered on the server is not really problematic. The user must delete them manually. On the other hand, predicting a normal message as a spam is very harmful. In fact, the message is automatically deleted, the user is not informed about this deletion, even if it is a very important message. The error rate does not take into account this difference. This measure gives the same importance to the two kinds of misclassification.

For this reason, in the text categorization context, the users prefer two more relevant measures to the domain:

- The recall  $r = \frac{a}{a+b}$  is the fraction of "spam" which is recovered. In our test set, we have r = 332/(332 + 59) = 84.9%.
- The precision  $p = \frac{a}{a+c}$  is the fraction of true positive among the message predicted as "spam". In our study, p = 332/(332 + 27) = 92.5.

A good classifier must have a high recall AND a high precision.

To handle two values simultaneously is never quite easy, especially when we want to compare the performances of several models. One proposes in the literature a synthetic criterion: the F-MEASURE. The formula is the following; it is a weighted harmonic mean of recall and precision.

$$F_{\alpha} = \frac{(1+\alpha^2) \times r \times p}{\alpha^2 \times p + r}$$

In the best case: r = 1, p = 1, then F = 1.

 $\alpha$  is the weight. If we set the same weight to recall and precision, we will choose  $\alpha = 1$ , then the F-Measure will be  $F_1 = \frac{2 \times 0.925 \times 0.849}{1 \times 0.925 + 0.849} = 0.885$ .

If we set  $\alpha = 0.5$ , we give twice more importance to precision compared with recall; if we set  $\alpha = 2$ , we give twice more importance to recall compared with precision (see <a href="http://en.wikipedia.org/wiki/Information-retrieval">http://en.wikipedia.org/wiki/Information-retrieval</a>).

The good value of  $\alpha$  depends on the constraints of the domain and the goal of the study. We can also define a range of values to study the behavior of the models on various scenarios.

# 4 Using non-symmetrical misclassification costs

In our spam detection problem, we would like to give ten times more importance to precision compared with recall. How can we insert this information in the decision tree induction process?

Concerning the evaluation, we modify the parameter of F-Measure by adopting the weight  $\alpha = 0.1$ . But how can we incorporate this information in the induction process?

The simplest approach is to integrate this new constraint in the form of misclassification costs matrix, and to use a method which takes into account this matrix explicitly. SIPINA implements a cost-sensitive decision tree induction approach, derived from C4.5 algorithm (Chauchat et al., 2001). In the following, we select this approach; then, we insert the misclassification cost matrix.

### 4.1 Stopping the preceding analysis

First, we must interrupt the preceding analysis. We click on the WINDOW / CLOSE ALL menu. All the windows of analysis are closed.

😹 Sipina Research Version											
Induction method Analysis View	Window Help										
<mark>ŭ 🖹 🕞</mark> 🐂 🖡	Cascade Tile										
Attribute selection	Arrange icons		www.								
🖻 📭 Predictive attributes	Close All		Ew?								
C wf_address C wf_address C wf_all C wf_3d C wf_our wf_our	1 Learning set e 2 Analysis Repo 3 Decision tree. 4 Confusion mai 5 Confusion mai	editor rt  trix : Lea trix : Tea	rning set on D:\Dai it set on D:\DataMi	taMining\Dat ning\Databa	abases_for_ ses_for_mini	mining\datase ng\dataset_fo	t_for_s r_soft_	oft_dev_and_comp dev_and_compariso	arison\f-measur	asure for e for dec	r decisio cision tre
wf_remove wf_internet	6 7 	= 9	State Confusion	matrix : 1	est set on	D: WataMin	ning\D	atabases_for_n	1 💶 🕻		
wf_order	<u>°</u>	ye	spam								— EI
wt_mail	10	no		yes		no					
wr_receive	11		yes	332		59					
C wf_people C wf_epont C wf_addresses C wf_tree C wf_business C wf_enail C wf_you C wf_credit	12 13 14 15 16 17 18 19										
	20										
Learning method MethodName=C4.5 (Quinlan - 1993) MethodClassName=TArbreDecisionC4 Hdl=5 Confidence level=25 Leaf size=2 Sampling=0	- 22 5 23 24 25 26	K									
SamplingPart=300	27	0.0 0.0 Cos									<u> </u>
Examples selection	20	0.67	Cost : 0.0860							11	
3601 examples selected 1000 examples idle	30	0.00	0.00	0.00	0.00	0.64	0.00	0.64	0.00	0.00	~
r C4.5 (Quinlan - 1993)										Time : :	1047

# 4.2 Selecting the cost sensitive approach

We click again on the INDUCTION METHOD / STANDARD ALGORITHM menu in order to choose the appropriate method. In the dialog box, we choose COST SENSITIVE C4.5 (CHAUCHAT & RAKOTOMALALA, 2001).

Select an induction method				
Induction Graph Rule Induction Neural network Discriminant analysis Decision list Other				
A limited search induction tree algorithm (Catlett - 1991) ID3-IV (Quinlan - 1986)				
GID3 (Cheng, Fayyad, Irani & Qian - 1988) ASSISTANT 86 (Cestnik, Kononenko & Bratko - 1986)				

In the parameter settings, we can insert the costs matrix during the growing phase (not really efficient) or/and the pruning phase (efficient).

Cost sensitive C4.5
Learning parameters
Lambda : 3
Size of leaves : 🙎 🚖
Handling Costs □ Growing ▼ Pruning
Sampling All dataset C Simple Size : C Balanced C Stratified 300

The LAMBDA parameter is used for the laplacian estimation of the class distribution within nodes. It can be understand as a smoothing parameter for the probability estimation.

# 4.3 Defining the misclassification cost matrix

We must define the misclassification costs. We click on the ANALYSIS / SET COSTS menu. In the dialog box, we insert the values. We validate by clicking on the OK button.



### 4.4 Training and testing the tree

We start the learning phase with the ANALYSIS / LEARNING menu. The tree is shorter than the preceding, with 36 leaves however.



We apply now the tree on the test set (ANALYSIS / TEST / INACTIVE EXAMPLES OF DATABASES). We obtain the following confusion matrix.

8.	Sipina Research	Version
In	duction method	Analys

SIPINA



We compare this result with the "standard" C4.5 method (Tableau 1).

Indicator	C4.5 « standard » (M1)	C4.5 « cost sensitive » (M2)
True positive	332	262
False positive	27	9
Error rate	8.6%	13.8%
Recall	84.9%	67.0%
Precision	92.5%	96.7%
F-Measure (0.1)	0.924	0.963

Tableau 1 - Comparison of the two approaches

According the error rate, M2 seems worst. But our main preoccupation is the precision. We observe that M2 is significantly better; it is also better according the F-Measure.

#### Conclusion 5

This tutorial had a double goal. The first was to show that it is possible to carry out the "train - test" sequence on a predefined subdivision of the dataset with SIPINA. That makes it possible to compare the performances of several algorithms implemented in SIPINA, but also implemented in other software which can handle this kind of subdivision.

The second important subject treated is the integration of the misclassification costs during the construction of the models. This functionality is unfortunately unusual in the Data Mining Software. It is implemented in SIPINA. It appeared interesting to show how to implement it, what to think about it, and how to read the results in this configuration.

- O X