1 Introduction

Supported data file format in SIPINA.

The data access is the first step of the data mining process. It is a crucial step. It is one of the main criteria used when we want to assess the quality of a tool. If we do not able to load a dataset, we cannot perform any kind of analysis. The software is not useable. If the data access is not easy and requires complicated operations, we will devote less time to the other steps of the data exploration.

Two points of view are essential when we want to evaluate the data file format: the flexibility and the performance. The flexibility refers to the ability to manipulate the dataset with other tools than the dedicated software, and thus, the ability to exchange the dataset between various tools. The delimited text file format, says also CSV file format, is most probably the easiest to handle. The great majority of tools can handle this format, especially the spreadsheet programs such as Excel or Open Office Calc, but also the DBMS (database management system) such as SQL Server or Oracle.

The performance refers mainly to the quickness of the input/output operations and, to a lesser extent, to the disk occupation. It is above all important when we treat a large dataset (millions of observations and/or hundreds of variables) or when we must repeatedly load and save the dataset.

When we deal with a moderate size dataset (up to 65,535 rows under Office XP; 1,048,576 under Office 2007), it is clear that the spreadsheet takes a particular position. Polls show that many people use Excel (or more generally spreadsheet) with a specialized data mining tools¹. It is not surprising. A spreadsheet program is very suitable to handle a data in the attribute-value format. In this context, the putting up a link between a spreadsheet and the specialized data mining tool is a very useful solution. We show below how to implement it with Sipina.

The first goal of this tutorial is to describe the various file formats that are supported in Sipina. Some of the solutions are more deeply described in other tutorials elsewhere; we indicate the appropriate reference in these cases. The second goal is to describe the behavior of these formats when we handle a large dataset with **4,817,099 instances** and **42 variables**.

Last, we learn a decision tree on this dataset in order to evaluate the behavior of Sipina when we process a large data file.

2 Dataset

First, we use the WEATHER.TXT dataset which a tab-separated values text file format (Quinlan, 1993 -- http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/weather.txt).

03/11/2009

See
 http://www.kdnuggets.com/polls/2008/tools-languages-used-data-cleaning.htm
 and

 http://www.kdnuggets.com/polls/2008/data-mining-software-tools-used.htm
 and

📃 weath	er.txt - Bl	oc-notes	1.00		
Fichier	Edition	Format Aff	ichage ?		
outloo sunny sunny sunny sunny ovcast ovcast ovcast rain rain rain rain	k Temp 75 80 85 72 69 72 83 64 81 71 65 75 68 70	Humid 70 90 85 95 70 90 78 65 75 80 70 80 80 80 96	windy yes no no no yes no yes no yes no no no no	Class Play DontPlay DontPlay Play Play Play Play Play DontPlay DontPlay Play Play Play Play	*
•					

This kind of file format is very used. In most of the cases, we consider that the first row corresponds to the name of the variables.

3 Tab-separated values file format

The starting point is the WEATHER.TXT above. In the following, we use the same process: (1) we load the dataset in a specified format; (2) we check if all the values are imported; (3) we export the dataset in a specified format; (4) we check if we can import it thereafter.

3.1 CSV (*.TXT) text file format

Importing the TXT file format. After we launch SIPINA, we click on the FILE / OPEN menu. A dialog box appears. We select the TEXT FILE FORMAT (*.TXT). We select the WEATHER.TXT data file.



Another dialog box appears, we set the column separator and we indicate that the first row

corresponds to the name of the variables.

Import text file	options						L X
	Outlook	Temp	Humid	Windy	Class		
	sunny	75	70	yes	Play	-	
	sunny	80	90	yes	DontPlay		
	sunny	85	85	no	DontPlay		
	sunny	72	95	no	DontPlay		
	sunny	69	70	no	Play		
	ovcast	72	90	yes	Play		
	ovcast	83	78	no	Play		
	ovcast	64	65	yes	Play		
	ovcast	81	75	no	Play		
Specification	e is name of attribu umn is label of exa ical attribute	utes amples	Delimiters Tabs C Space C Other				
			1			🗙 Annuler	🗸 ОК

We validate by clicking the OK button. The dataset is imported. The values are displayed into the visualization grid.

Sipina Research Version							
File Edit Data Statistics Inducti	on method	Analysis	View Wi	ndow Help			
Y ¥ 📴 ⊾							
		1				· · · · · · · · · · · · · · · · · · ·	
Attribute selection	🔊 Learnin	ng set editor				1. A.	Â
		Outlook	Temp	Humid	Windy	Class	
	1	sunny	75.00	70.00	yes	Play	· -
	2	sunny	80.00	90.00	yes	DontPlay	-
	3	sunny	85.00	85.00	no	DontPlay	
II /	4	sunny	72.00	95.00	no	DontPlay	
1	5	sunny	69.00	70.00	no	Play	1
il il	6	ovcast	72.00	90.00	yes	Play	
II	7	ovcast	83.00	78.00	no	Play	
Learning method	8	ovcast	64.00	65.00	yes	Play	i 1
MethodName=Improved ChAID (Tsc 🔺	9	ovcast	81.00	75.00	no	Play	1
MethodClassName=TArbreDecision	10	rain	71.00	80.00	yes	DontPlay	
Merce=0.05	11	rain	65.00	70.00	yes	DontPlay	1
Split=0.001	12	rain	75.00	80.00	no	Play	
TypeBonferroni=1	13	rain	68.00	80.00	no	Play	
ValueBonterroni=1	14	rain	70.00	96.00	no	Play	
Europhics coloction						1	
		1					
			_				-
	•	11			-		•
Improved ChAID (Tschuprow Goodnes	s of Split)					Exec.Time :	16 ms. //

Note: Specifying the types of the variables. Sipina refers to the first row of the data (the second row if the first row corresponds to the name of the variables) to determine the type of the variable. If we have a numerical value, Sipina considers that we have a continuous attribute; if we have alphabetical value, it sets the column as a discrete attribute. We can modify these default settings by clicking on the DATA / DEFINE SPECIFICATIONS menu. Sipina can modify the type of variables; it can scan the set of values for each column, etc.

Caution: decimal point. For Sipina, the decimal point is always «.» whatever our OS configuration.

Exporting the dataset into the TXT format. In order to export a dataset in the TXT file format, we click

~	•	٠	
5	١r	וור	na
-	۱ŀ		iu

on the FILE / EXPORT menu. We select the TXT format. A dialog box asks us the list of variables to export, and other specifications.

Text file format options	
Variables to export	Export with variables names examples labels Delimiters is Tabs Space Other V OK

3.2 Weka file format (*.ARFF)

Weka is a very popular data mining tool (<u>http://www.cs.waikato.ac.nz/ml/weka/</u>). It has its own format (ARFF) which is a text file format with additional information. The first part is the data dictionary; the second part corresponds to the values.

Exporting into the ARFF format. To export the previous dataset into the ARFF file format, we click on the FILE / EXPORT menu. We select the Weka files format. We set WEATHER.ARFF as file name.

🔉 Si	pina Research Version							
File	Edit Data Statistics Induction r New	nethod Analysis	View Windo	ow Help				
	Open Save Save s	Learning set editor Outlook sunny ouppy	Temp 75.00	Humid 70.00	Windy yes	Class Play		
	Export	Stregistrer so	us	90.00	ves	Domeav		×
Learn Meth Hdl= Split- Sam Exa	Subsample management * Exit * ining method * iod/Name=Improved ChAID (Tsc 8 ige=0.05 10 90.001 8 Bonferroni=1 13 Bing=0 * mples selection * sved ChAID (Tschuprow Goodness of State *	Enregistrer dans : Emplacements récents Bureau Maison Ordinateur	ipina file f	onet Date dam Aucun e	III Type	Taille	The second seco	
		Réseau	Nom du fichier : Type :	Wea	ther.aff KA Files		•	Enregistrer Annuler

We can check the data file on the disk. We see here the contents of the file into the standard Windows notepad.

weather.arff - Bloc-notes		
Fichier Edition Format Affichage ?		
<pre>@relation weather.arff @attribute Outlook {sunny,ovcast,rain} @attribute Temp REAL @attribute Humid REAL @attribute Windy {yes,no} @data winny,75,70,yes,Play winny,85,85,no,DontPlay sunny,85,85,no,DontPlay sunny,69,70,no,Play ovcast,72,90,yes,Play ovcast,64,65,yes,Play ovcast,64,65,yes,Play ovcast,64,65,yes,DontPlay rain,75,80,no,Play rain,75,80,no,Play rain,75,80,no,Play rain,75,80,no,Play rain,75,80,no,Play rain,70,96,no,Play</pre>		*
<	۲	зđ

In the data dictionary part, a continuous variable is pointed out by the REAL keyword; the set of values are simply enumerated for a discrete attribute. The "@data" part is similar to the CSV file format.

Importing the ARFF file format. Sipina can load a data file in the ARFF format. To do this, we click on the FILE / OPEN and we select the Weka File Format (*.ARFF). No additional settings are required.

3.3 Text file for the old Sipina - Version 2.5

The old version of Sipina (2.5 and older) uses a text format which is very similar to the ARFF format. But, it uses two separate files: the first corresponds to the data dictionary; the second contains the values. The main drawback is that the user had to handle simultaneously two files. It was not really convenient.

Exporting into the Sipina version 2.5 file format. We click on the FILE / EXPORT format. We select the SIPINA V2.5 FILES option. We set WEATHER.DAT as file name.

Sipina Research Version					
File Edit Data Statistics Induction m	nethod Analysis	View Window Help			
New					
Open Save	Learning set editor	Temp Humid	Windy Class		
Save as	sunny	75.00 70.00	yes Play	=	
Export	🔉 Enregistrer sou	us			×
Subsample management	Enregistrer dans :	📔 sipina file format			* 💷 🕶
Exit	Ca	Nom Date d	e m Type	Taille	
Total 7 Learning method 9 MethodName=Improved ChAID (Tsc ~ MethodClassName=TArbreDecision) 9 Hdl=8 Merge=0.05 Spit=0.001 TypeBorferroni=1 ValueBonferroni=1 Sampling=0 11 12 13 Examples selection 14 Improved ChAID (Tschuprow Goodness of S	Piplacements récents Bureau Maison Ordinateur	Auc	un élément ne corres	nond à votre reche	rche.
	Réseau	Nom du fichier : Type :	weather.dat SIPINA v2.5 Files 🛛 🕌		Enregistrer Annuler
	(L				

A dialog box appears. We set the list of variables to export, ALL VARIABLES here.

Variables to export	
Select variables to export	
All Variables Dutlook Temp Humid Windy Class	
🗸 ок	

In the following dialog box, we choose the target attribute for a supervised learning scheme.

Sipina file format options	
Title	Select class attribute
[🗸 ОК

We can see that two new files are created on the disk.

A dataset_for_soft_dev_and_comparison > sipina file	format	↓ 49	Rechercher	× 0
Fichier Edition Affichage Outils ? ↓ Organiser マ 🚍 Affichages マ 🗂 Ouvrir マ 🛞 Graver				0
Liens favoris Documents Images Musique Autres » Dossiers	Nom	Date de modifie weather.bt Document texte 351 octets weather.arff Fichier ARFF 496 octets weather.dat	cation Type	e Taille
 non_parametric_test normality_test outliers partial_correlation pls regression regression_logistique_pour_la_sante sipina file format 		Fichier DAT 299 octets weather.par Fichier PAR 245 octets		
weather.arff Date de modificati 16/02/2009 07:18 ARFF Data File Taille : 496 octets Date de création : 16/02/2009 07:18				

We can open them in the NOTEPAD. For WEATHER.PAR, we have the description of the attributes and their statuses in the learning scheme.

💭 D:\DataMining\Databases_for_mining\datase 👝 💷 💽
TITRE : ;
VARIABLE
Outlook : sunny=1,ovcast=2,rain=3
Temp : MINI=0, MAXI=2
Humid : MINI=0, MAXI=2
Windy : yes=1,no=2
Class : Play=1,DontPlay=2;
ENDOGENE = 5 ;
EXOGENE = 1 ,2 ,3 ,4 ;
SEGMENTATION ;
LAMBUA = 51; IAILLE = 2;
٠ • //

Into WEATHER.DAT, we have the description of the values. Sipina uses the encoding values for the discrete attributes.

💭 D:\DataMining\Databases_for_mining\dataset_for_soft_dev_and_compariso 💼 💷 🗮	X
1 1 75.00 70.00 1 1	1
2 1 80.00 90.00 1 2	
3 1 85.00 85.00 2 2	
4 1 72.00 95.00 2 2	
5 1 69.00 70.00 2 1	
6 2 72.00 90.00 1 1	
7 2 83.00 78.00 2 1	
8 2 64.00 65.00 1 1	
9 2 81.00 75.00 2 1	
10 3 71.00 80.00 1 2	
11 3 65.00 70.00 1 2	
12 3 75.00 80.00 2 1	
13 3 68.00 80.00 2 1	
14 3 70.00 96.00 2 1	
	_
	111

3.4 HTML file format

The HTML file format is also a text file format. The dataset can be visualized in a browser. It can be also imported into a spreadsheet (Excel or Open Office Calc). We note however that HTML is a very verbose format. The size of the file can increase rapidly for large datasets.

Exporting into the HTML file format. We click on the FILE / EXPORT menu. We select the HTML file format. We set WEATHER.HTML as file name.



The dataset can be displayed into a browser, Firefox here.

🕲 м	lozilla Firef	рх									x
<u>F</u> ichi	<u>Fichier Édition</u> <u>Affichage Historique Marque-pages Qutils ?</u>										
R		C :	X 🏠		fil 🟠 🔹	. (ئ	Wikipe	édia	(fr)		ρ
					. You .						
	Bureau Lyon 2 Moteurs de recherche 📸 Prévisions météo de M										
Goo	Google G Rechercher • • >> O Paramètres•										
ß	file:///D:/DataMat/weather.html ×										
		_									
ID	Outlook	Temp	Humid	Windy	Class						
1	sunny	75.00	70.00	yes	Play						
2	sunny	80.00	90.00	yes	DontPlay						
3	sunny	85.00	85.00	no	DontPlay						
4	sunny	72.00	95.00	no	DontPlay						
5	sunny	69.00	70.00	no	Play						
6	ovcast	72.00	90.00	yes	Play						
7	ovcast	83.00	78.00	no	Play						
8	ovcast	64.00	65.00	yes	Play						
9	ovcast	81.00	75.00	no	Play						
10	rain	71.00	80.00	yes	DontPlay						
11	rain	65.00	70.00	yes	DontPlay						
12	rain	75.00	80.00	no	Play						
13	rain	68.00	80.00	no	Play						
14	rain	70.00	96.00	no	Play						
Term	niné										

The dataset can be also imported into a spreadsheet, Open Office Calc for instance.

诸 we	eather.html - Op	enOffice.org Ca	alc					۲.	
<u>F</u> ichie	<u>Fichier Édition Affichage Insertion Format Outils Données Fenêtre Aide</u>								
1 🗃	- 🔰 📄 🗠	- 🕑 i 🗟 🕯	🖹 🕓 🍪 🛛	😴 🄀 🖷 I	🛍 • 🎸 🧐) • @ • 🥃	3 🛃	>> •	
	$\boxed{ \mathbb{R} \text{Times New Roman}} \boxed{ \mathbb{R} $								
HTM	1L_tables	• 7 2	= [ID						
	A	В	С	D	E	F	G		
1	ID	Outlook	Temp	Humid	Windy	Class			
2	1	sunny	75.00	70.00	yes	Play Play			
3	2	sunny	80.00	90.00	yes	DontPlay			
4	3	sunny	85.00	85.00	no	DontPlay		Ξ	
5	4	sunny	72.00	95.00	no	DontPlay			
6	5	sunny	69.00	70.00	no	Play			
7	6	ovcast	72.00	90.00	yes	Play			
8	7	ovcast	83.00	78.00	no	Play Play		_	
9	8	ovcast	64.00	65.00	yes	Play Play		_	
10	9	ovcast	81.00	75.00	no	Play			
11	10	rain	71.00	80.00	yes	DontPlay			
12	11	rain	65.00	70.00	yes	DontPlay		-	
13	12	rain	75.00	80.00	no	Play		-	
14	13	rain	68.00	80.00	no	Play		-	
15	14	rain	70.00	96.00	no	Play Play			
16									
	Feuille		·	•			•		
Feuil	lle 1 / 1 Standa	rd 100	1%	STD		Somme=0)		

4 The binary file formats of Sipina

Sipina has a proprietary binary file format. In the simplest version (FDM), the vectors of values are directly dumped into a file. The input/output operations are very fast for this format. It is especially intended to data files containing a high number of instances and a moderate number of variables. It is less adapted for the dataset containing a high number of variables.

About the size of the file, we can easily compute it:

- The disk occupation of each value is 4 bytes.
- To each row, we have a label described on 25 bytes.
- Each name of variable is encoded on 25 bytes.
- We can add eventually the description of the values for a discrete attribute.

Thus, for a dataset with 100,000 instances and 10 continuous attributes, the file size is:

100.000 x 10 x 4 + 100.000 x 25 + 10 x 25 = 6.500.250 bytes # 6.2 MB

4.1 Handling the FDM file format

Exporting in the FDM file format. We click on the FILE / SAVE AS menu. The default file format is FDM, we set WEATHER.FDM as file name.

Si Si	pina Research Version	thod A too	Window Hel			
	New			P		
	Open Save	ea Enregistrer so	us			X
	Save as		Nom Da	ate de m Type	Taille	
	Subsample management	Emplacements récents		Aucun élément ne conse	spond à votre recherc	he.
J Learr	Exit	Bureau				
Meth Meth Hdl=8 Merg Split=	odCName=Improved ChAID (Tsc A 9 odClassName=TArbreDecisionI 8 e=0.05 0.001	Maison				
Value Samp	Bonferroni=1 Bonferroni=1 Jing=0	Ordinateur				
Impro	ved ChAID (Tschuprow Goodness of Sp	lit Réseau	Nom du fichier : Type :	weather.fdm Data Manager File (*.F	DM)	Annuler

As we see into the directory, the file size (WEATHER.FDM) is higher than the others (TXT, ARFF).

A dataset_for_soft_dev_and_comparison + sipina file	format 👻 🖣	Rechercher	٩
Fichier Edition Affichage Outils ?			
🆣 Organiser 🔻 📰 Affichages 🔻 📑 Ouvrir 🔞 Graver			•
Liens favoris	Nom Date de modification	Type Taille	
Documents Images	weather.txt Document texte 351 octets	weather.arff Fichier ARFF 496 octets	
Musique Autres »	weather.dat Fichier DAT 299 octets	Fichier PAR 245 octets	
Dossiers	weather.html Firefox Document 1.92 Ko	weather.fdm Fichier FDM 2.01 Ko	←
weather.fdm Date de modificati 16/02/2009 08:14 Fichier FDM Taille : 2.01 Ko Date de création : 16/02/2009 08:14			A

Reading the FDM file format. Only Sipina can read the FDM file. We click on the FILE / OPEN menu and we select the appropriate file (WEATHER.FDM).

The main argument in favor of FDM is the quickness. The main drawback is the file size. In this context, we had developed a compressed format for the FDM binary file format. There are two compressed format in Sipina.

4.2 The FDZ file format

FDZ is also a binary file format. It is the FDM file which is compressed with the LZW algorithm (Lempel Ziv Welch - <u>http://en.wikipedia.org/wiki/Lempel–Ziv–Welch</u>). Rather the quality of the compression, we give priority to the quickness of the processing. The resulting file size is thus larger than those obtained with dedicated tools such as 7-Zip.

We note that this strategy creates a temporary file on the disk during the processing. It is of course removed when the final FDZ file is created.

Creating a file in a FDZ format. We click on the FILE / SAVE AS menu. We select the FDZ format and we set WEATHER.FDZ as file name.

🔊 S	ipina Research Version					
File	Edit Data Statistics Induction me	thod Analysis V	iew Window	Help		
	New	Enregistrer so	us			×
	Open 🕅 🕺	Enregistrer dans :	📗 sipina file for	mat		
_	Save	a m	Nom	Date de m Type	Taille	
	Save as			Aucun élément ne corr	espond wotre recherche	
	Export	 Emplacements récents 		Addit clement ne con	espond to one recirciterent.	
	Subsample management					
	Exit	Rumou				
<u>, </u>	1 7	Dureau				
Lear	ning method 8					
Met	nodName=Improved ChAID (Tsc 🔺 📗 9					
Hdl=	·8 = 11	Maison				
Split	=0.001 12					
Type	eBonferroni=1					
Sam	pling=0 - 14	Ordinateur				
Exa	mples selection					
	I = 1	Keseau	Nom du fichier :	weather.fdz		arregistrer
Impr	oved ChAID (Tschuprow Goodness of Sp		Type :	Compressed Data Ma	anager File (*.FDZ)	• Annuler

Reading the FDZ file format. We click on the FILE / OPEN menu when we want to read a FDZ file.

4.3 The ZDM

The ZDM format is also a compressed version of the FDM file. But, unlike the FDZ format, the compression process is performed "on the fly". None temporary file is created.

We must give priority to this format when we work on a support with poor performances (such as USB memory bar for instance).

Creating a ZDM file. We click on the FILE / SAVE AS menu and we select the ZDM format. We set WEATHER.ZDM as file name.



Reading a ZDM file. Such as the other binary formats, we click on the FILE / OPEN menu and we select the appropriate format when we want to read the data file.

5 Summary – The size of the various file format

We summarize here the file size according to the selected file format. For the old Sipina 2.5 format, we must add up the size of PAR and DAT files.



We note that the ZDM format is smallest; at the opposite, the FDM is the largest. This is fairly typical situations that we encounter in practice.

6 A link between Excel and Sipina

Many data miner use a spreadsheet during the data preparation phase. Because it is easy to use and it is rather efficient for the moderate sized datasets.

We can send a dataset from Excel to Sipina using the SIPINA.XLA add-in. The installation of the add-in is described here: <u>http://eric.univ-lyon2.fr/~ricco/doc/sipina_xla_installation.htm</u>; its use is described here: <u>http://eric.univ-lyon2.fr/~ricco/doc/sipina_xla_processing.htm</u>.

For our dataset, after the installation of the add-in, we perform the following steps in order to send the dataset towards Sipina. Sipina is automatically launched.



7 Performance comparison

We have various ways to handle a dataset. What is the file format that we must choose?

For a moderate sized dataset, the connexion between Excel and Sipina must be used. It is a good compromise between flexibility and quickness. In addition, we have benefit of all the data processing capabilities of a spreadsheet.

For a large sized dataset, we must choose between the flexibility of the text file format and the quickness of the binary file. It depends on the needs of data exchange between users or tools

(flexibility) and the need to read and save repeatedly the dataset (quickness).

In order to compare the behavior of the various file formats, we use another data file with **4,817,099** instances and **42** variables (<u>http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/kdd-cup-discretized-descriptors.txt.zip</u>). It is a cleansed version of the « KDD Cup 1999 Data² ». We use two criteria in order to compare the performances of the file formats: the disk occupation; and the processing time during the reading and the writing. Our PC is a Quad Core Q9400 at 2.66 GHz, running under Windows Vista.

	Text file (TXT)	ARFF (Weka)	Binary FDM	Binary FDZ (compressed)	Binary ZDM (compressed)
File size on the disk (KB)	447.917	447.919	808.055	107.846	15.037
File reading (seconds)	107	105	8	57	7
File writing (seconds)	188	120	8	15	15

We note several results:

- It is obvious that the memory occupation of the dataset is the same when it is loaded, whatever the file format used.
- About the file size, TXT are ARFF are similar. It is not surprising.
- The FDM file is voluminous. The compressed format enables to dramatically reduce the FDM file size, especially for the ZDM format.
- About the processing time, as expected the reading and the writing of the text file (TXT and ARFF) are not fast. Notably because alphabetical value must be encoded during the reading process. The writing seems slower than the reading because it was not really optimized into Sipina.
- On the other hand, the binary file format is very fast, especially for the FDM format. None intermediate operation comes to slow down the input/output of the vectors of values.
- As expected, FDZ and ZDM are a bit slower than FDM. With the exception of ZDM during the reading. I think it is the consequence of the reduced size of the file.

We show here the disk occupation of the various formats of the dataset. We note that when we use an external program in order to compress the text files, we obtain a smaller file size (TXT.ZIP). We use the ULTRA option of 7-ZIP. Perhaps it is another way to improve the disk occupation of the data files. But in this case, the processing is very slow (about 120 seconds on my computer).

		I COM		100.00 (M)		x
COO V III « concou	urs-cup-etc 🕨	kdd-cup-99 🕨 test big dataset for sipina 🕨	▼ 4 ₂	Rechercher		٩
Fichier Edition Affich	age Outils	?				
🆣 Organiser 👻 🏢 Aff	fichages 🔻 🄇	🔋 Graver		_	_	0
Liens favoris		Nom	Date de modificati	Туре	Taille	
Documents Images Musique Autres >> Dossiers Eureau	×	 kdd-cup-discretized-descriptors.txt.zip kdd-cup-discretized-descriptors.zdm kdd-cup-discretized-descriptors.fdz kdd-cup-discretized-descriptors.txt kdd-cup-discretized-descriptors.aff kdd-cup-discretized-descriptors.fdm 	15/02/2009 20:29 16/02/2009 08:44 16/02/2009 08:40 27/01/2009 21:40 16/02/2009 08:55 16/02/2009 08:35	Dossier compressé Fichier ZDM Fichier FDZ Document texte Fichier ARFF Fichier FDM	2 118 Ko 15 037 Ko 107 846 Ko 447 917 Ko 447 919 Ko 808 055 Ko	
6 éléments					A	/

8 Decision tree learning on KDD-CUP 1999 dataset

Even if it is not the main subject of this tutorial, we want to observe the behavior of Sipina when we learn a decision tree from our large dataset. With the dataset loaded, and before any learning process, the memory occupation is 882 MB.

We specify the TARGET attribute and the INPUT ones by clicking on the ANALYSIS / DEFINE CLASS ATTRIBUTE menu.

🔉 Sipina Research Version - [Learnir	ng set editor]		
🔉 File Edit Data Statistics In	nduction method	Analysis View Window Help	_ 8 ×
🔥 🖄 📴 📭		Define class attribute	
X	U v1	Select active examples	V8 V.A
Attribute selection	1 a	Attribute selection	
	2 a	Set wei	izbles
	3 a	Set pric Class	2
	4 a	Set cos	3
	5 a	24 y24 y24 y24 y24 y24 y24 y24 y24 y24 y	4
	6 a	SET POS Attributes v26	5
	7 a	Learnin V1	7
	8 a	v3	3
	9 a	- Stop ar V4 V3)
	10 a	Classifi v6 v3	2
	11 a	V/ V3	3
Learning method	12 a	V3	5
MethodName-Improved ChAID (Tast	13 a	V10 v30	5
MethodClassName=TArbreDecisionI	14 a	V3.	/ 8
Hdl=8	15 a	Error m C Only discrete	9
Merge=0.05	16 a	C Only continuous	J 1
TypeBonferroni=1	17 a	Feature 6 Both	sse
ValueBonferroni=1	18 a		
Sampling=0	19 a	Person	
Examples selection	20 a	tcp	VK Annuler
	₹ □	ian hite CE n n a	P → F
	Editing D:\Dat	aMining\Databases for mining\concours-cup-et Attributes : 41 Exar	nples : 4817099 🛛 🖉
Improved ChAID (Tschuprow Goodne	ess of Split)	Exec.Time : 67	54 ms.

² http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

~ ·	•	
51	nına	
2	pina	

We want to use the half of the dataset for the learning phase, the other half for the evaluation of the classifier. We click on the ANALYSIS / SELECT ACTIVE EXAMPLES menu; then we choose the RANDOM SAMPLING option. The learning sample size is 2,408,549 cases (50% of the dataset).

💦 Sipina Research Version - [Learnin	ng set editor]						×
🔉 File Edit Data Statistics In	nduction method	Analysis View Window Help				- 8	×
🖰 📑 👺 🕒		Define class attribute					
×	v1	Select active examples	v5	v6	7	v8	V ~
Attribute selection	1 a	Set weight field	p	р	8	a	ε 🛄
Predictive attributes	2 a	Set weight heid	p	p	a	а	8
	3 a	Set priors	p	p	а	а	8
- D v3	4 a	Set costs	р	р	а	а	8
D v4	5 Filter a	ctive examples				-	
<mark>D</mark> v5	7					1	
<mark>-</mark> 2 ∨6_	8	Size of sample					
	9	50 🗢 % ie 2408549	of 4817	7099 exam	noles		
	10						
- D v10	11	с. г					
Learning method	12	Sampling strategy					
MethodName=Improved ChAID (Tsc	13	• Handom					
MethodClassName=TArbreDecisionI	14	C Stratified					
Merce=0.05	15						
Split=0.001	16						
TypeBonterroni=1	17						
Sampling=0 -	10				· · · ·		
Examples selection	20						
	24				· · · · ·	Filter method	
						C List	
	Editing					 Random sa 	ampling
Improved ChAID (Tschuprow Goodne	ess of Split)					C Rule selec	tion
	List	of examples Random complian Bule	filter				4 Annual

We can launch the learning process. We click on the ANALYSIS / LEARNING menu. The tree is obtained after 50 seconds.



The memory occupation of Sipina is 920 MB at this step.

We want to assess the tree on the test set. We click on the ANALYSIS / TEST menu. Into the dialog settings, we select the INACTIVE EXAMPLES OF DATABASE option. Sipina displays the confusion matrix and the test error rate which is 1.61%.



9 Conclusion

Our idea in this tutorial was to give a quick overview of file formats that could be processed with Sipina. They can be classified into two categories: text format, with the main asset flexibility; and binary formats, crucial for processing times and disk occupation when compressed.

Of course, there is no format which could be uniformly better than the others. It depends, on the one hand, to the characteristics of the dataset; on the other hand, to the objective and the constraints of our analysis. The flexibility and the performances (disk occupancy and processing quickness) seem to be good criteria to compare the advantages and the drawbacks of the various formats.