## 1 Theme

#### Studying the behavior of 64-bit version of some tools when they handle a very large dataset.

Because I have recently updated my operating system (OS), I am wondering how the 64-bit versions of Knime 2.4.2 and RapidMiner 5.1.011 could handle a very large dataset, which cannot be loaded into main memory on a 32-bit OS. This article completes a previous study where we deal with a moderate sized dataset with 500,000 instances and 22 variables. Here, we handle a dataset with 9,634,198 instances and 41 variables. We have already used this dataset in another tutorial. We showed that we cannot perform a decision tree induction on this kind of database without a swapping system, which is implemented into the SIPINA, on a 32-bit OS. We note that Tanagra can handle the dataset, but this is because it encodes the values of the categorical attributes with a byte. The memory occupation remains moderate.

In this tutorial, we analyze the behavior of the 64-bit Knime and RapidMiner on this database. We use a 64-bit OS, but we have "only" 4 GB of available memory on our personal computer.

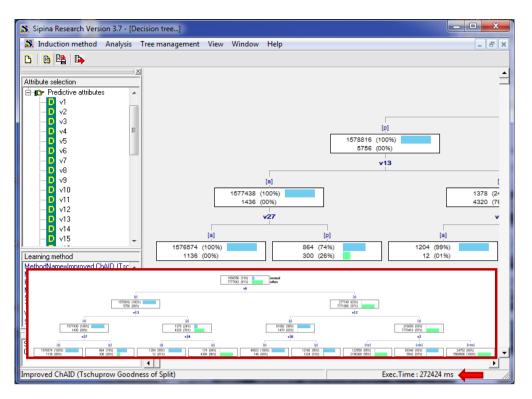
## 2 SIPINA

To ensure that SIPINA can handle this dataset, we must modify the settings into the **SIPINA.INI** initialization file<sup>1</sup>.

💦 File Edit Data Statistics	Induction r	method A	nalysis Vie	w Window	Help				-	8
<u>) 🐺 🙀 📄 </u>										
12	4	v1	v2	v3	v4	v5	<b>v</b> 6	v7	v8	
Attribute selection	. 1	а	tcp	http	SF	р	р	a	а	
	2	а	tcp	http	SF	р	p	а	а	
	3	а	tcp	http	SF	р	p	а	а	
	4	а	tcp	http	SF	р	p	а	а	
	5	а	tcp	http	SF	р	p	а	а	
	6	а	tcp	http	SF	р	p	а	а	
	7	а	tcp	http	SF	р	p	а	а	
	8	а	tcp	http	SF	р	p	а	а	
	9	а	tcp	http	SF	p	p	a	а	
	10	а	tcp	http	SF	p	p	а	а	
	11	а	tcp	http	SF	р	р	a	а	
	12	а	tcp	http	SF	р	р	а	а	
	13	a	tcp	http	SF	p	p	а	а	
	14	а	tcp	http	SF	р	p	а	a	
	15	a	tcp	http	SF	p	p	а	а	
	16	a	tcp	http	SF	p	p	а	а	
	17	a	tcp	http	SF	p	p	а	а	
earning method	18	a	tcp	http	SF	p	p	а	а	
ethodName=Improved ChAID (Tsc 🔺 ethodClassName=TArbreDecisionI 🚽	19	a	tcp	http	SF	p	p	а	а	_
dl-8	20	a	tcp	http	SF	p	p	а	a	_
erge=0.05	21	a	tcp	http	SF	p	p	a	a	
plit=0.001 vpeBonferroni=1	22	a	tcp	http	SF	p	p	а	а	
ypeBonrerroni=1 alueBonferroni=1	23	a	tcp	http	SF	p	p	a	а	
ampling=0	24	a	tcp	http	SF	p	p	a	а	
xamples selection	25	a	tcp	http	SF	p	p	a	a	
	26	a	ten	http	SF	n	n	8	a	
	•									١Þ.
	Editing	NEW.FDN	Λ			At	tributes : 41	Example	s:9634198	

The data file is imported into 232 seconds. The processing time for the decision tree induction is 272 seconds (about 5 minutes).

<sup>&</sup>lt;sup>1</sup> http://data-mining-tutorials.blogspot.com/2010/01/dealing-with-very-large-dataset-in.html

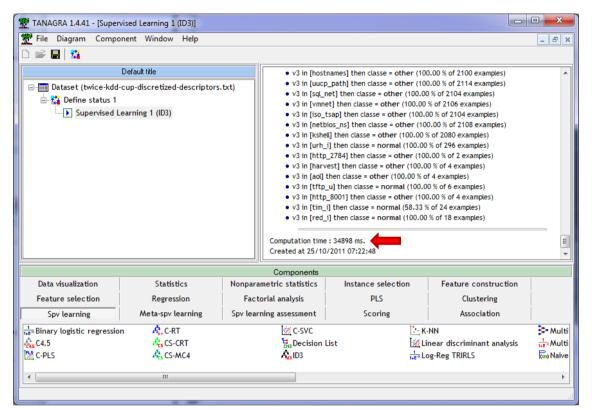


The whole process is already described elsewhere. We note only that the processing time is not influenced by the OS. On the same PC, we have the same performance whatever the OS version (Windows Vista 32-bit vs. Windows 7 64-bit). The memory occupation (about 232 MB) remains moderate in relation to the database size. This is the main interest of a swapping system. We can handle a very large database. The disk size is not a limitation. We note also that the solution is all the more efficient that the disk is performing. If we use a <u>solid state disk</u> for the temporary files, the computation time will be very likely reduced.

oplications	Processus	Services	Perf	formance	Mise	en réseau Utilisateurs		
Nom de l'in	nage N	om d'u	P	Mémoire .		Description	*	
Sipina_Res	sear M	laison	00	237 832	к	Sipina_Research		
firefox.ex	e*32 M	laison	01	118 252	к	Firefox		
explorer.e	xe M	laison	00	38 704	ĸ	Explorateur Windows		
wmdc.exe	M	laison	00	36 308	к	Gestionnaire pour appareils Win		
OUTLOOK	.EX M	laison	00	28 844	к	Microsoft Office Outlook		
plugin-con	tain M	laison	02	18 232	к	Plugin Container for Firefox	=	
dwm.exe	M	laison	02	15 400	к	Gestionnaire de fenêtres du Bur		
POWERPNT.E Ma		laison	00	13 868	к	Microsoft Office PowerPoint		
vlc.exe *3	32 M	laison	00	12 000	к	VLC media player		
csrss.exe			00	11 752	к			
IAStorIcor	n.ex M	laison	00	7 368	к	IAStorIcon	_	
WmdHost.	exe M	laison	00	6 272	к	Windows Mobile Device Center		
splwow64	exe M	laison	00	4 864	ĸ	Print driver host for 32bit applic		
avgtray.e	xe M	laison	00	4 704	ĸ	AVG Tray Monitor		
SnippingTo	ool M	laison	00	3 864	ĸ	Outil Capture		
taskhost.e	exe M	laison	00	3 212	к	Processus hôte pour Tâches Wi	_	
•			~~			4		
							_	

## 3 Tanagra

On the same computer, the performance of Tanagra - which runs under 32-bit mode - under Windows 64-bit version is not modified. The importation time is 87 seconds, and the tree is built in 34 seconds.

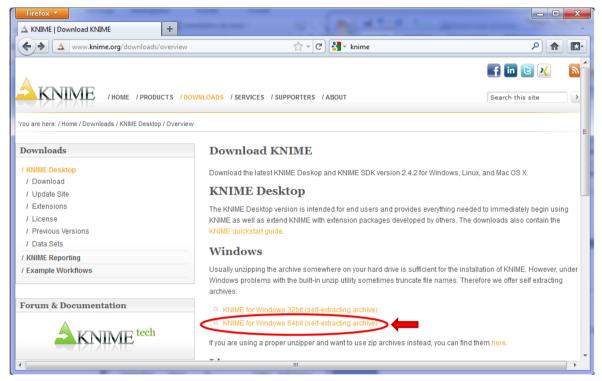


Because, the values of the categorical attributes are encoded with a byte (the number of values for one attribute is limited to 255), the memory occupation ( $\approx$  560 MB) makes the calculations tractable.

Applications	Proces	sus Service	s Per	formance Mise en rése	au Utilisateurs	
Nom de l'i	mage	Nom d'u	P	Mémoire (jeu de tr	Description	
Tanagra.	exe	Maison	00	573 464 K	Tanagra.exe	
firefox.ex	(e *32	Maison	00	196 836 K	Firefox	
AcroRd32	2.ex	Maison	02	79 916 K	Adobe Reader	
plugin-co	ntain	Maison	00	52 772 K	Plugin Container for Firefo	
explorer.	exe	Maison	00	20 260 K	Explorateur Windows	
dwm.exe		Maison	00	13 708 K	Gestionnaire de fenêtres 🗮	
csrss.exe			00	11 160 K		
IAStorIcon.ex		Maison	00	10 876 K	IAStorIcon	
splwow64	ł.exe	Maison	00	6 152 K	Print driver host for 32bit	
AcroRd32	2.ex	Maison	00	5 248 K	Adobe Reader	
avgtray.	exe	Maison	00	4 576 K	AVG Tray Monitor	
POWERP	VT.E	Maison	00	4 504 K	Microsoft Office PowerPoi	
SnippingT	ool	Maison	00	3 976 K	Outil Capture	
taskmgr.e	exe	Maison	00	3 216 K	Gestionnaire des tâches d	
nvvsvc.e	xe		00	3 152 K		
wisptis.ex	æ	Maison	00	2 972 K	Composant de saisie tacti 🖕	
				III 0.000 W	<u>►</u> 101 <del>-</del> - 1	
<u> </u>	i ele en la			les utilisateurs	Arrêter le processus	

## 4 Knime

On the Knime website, we can choose between the 32-bit and the 64-bit version of the software.



We select this last one for this tutorial. We do not describe the whole definition of the workflow here because we have already done this <u>previously</u> (section 6.2).

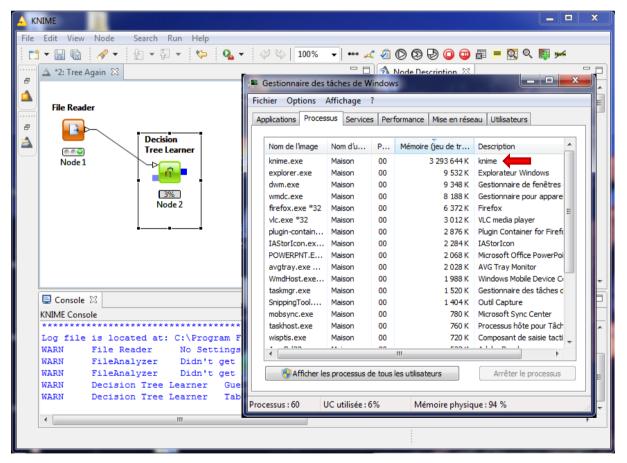
Let us see what happens when we launch the processing. Knime displays an enigmatic error message.

		CLASS CONTRACTOR	
File Edit View Search Run Node	e Help		
📑 🖬 🗟 🛷 🕶 🖓 🕶	🔸 🌾 🗛 🕶 🥪 🔛		
🛀 🔬 *0: Tree on large dataset 🛛	- 8	🔕 Node Description 🔀	
			*
		Decision Tree Learn	er 👘
File Reader	cision ee Learner	This node induces a classification decision tree in main m attribute must be nominal. The other attributes used for deci- either nominal or numerical. Numeric splits are always binar dividing the domain in two partitions at a given split point. No either binary (two outcomes) or they can have as many ou values. In the case of a binary split the nominal values ar subsets. The algorithm provides two quality measures for sp gini index and the gain ratio. Further, there exist a post prunin the tree size and increase prediction accuracy. The pruning the minimum description length principle. The algorithm can be run in multiple threads, and thu	sion making can be ry (two outcomes), ominal splits can be toomes as nominal re divided into two plit calculation; the g method to reduce method is based on
📮 Console 🖂			🛃 🗉 🔻 📑 🖛 🗖
KNIME Console Ctrl+N			
**************************************	*****	**************************************	*
	4.2.0030673 - the Konstan 2011, Uni Konstanz and KN		
*****	*****	*****	
ERROR BufferFromFileIt ERROR BufferFromFileIt	eratorVersion20 E eratorVersion20 C	ts\Knime\.metadata\knime\knime.log rrors while reading row 1180576 from file "knim ODING PROBLEM OutOfMemoryError caught, implemen C overhead limit exceeded	
			~
•	III		۱

After some search (e.g. <u>http://www.petefreitag.com/item/746.cfm</u>), we understand that we can overcome this memory limitation by modifying the heap allocation memory. If we suppose that a double precision coding is used for each value, the memory occupation is (roughly, we do not take into account the eventual meta-information and the memory used for the calculations) about 2.94 GB [(9.634.198 x 41 x 8) / (1024 x 1024 x 1024)  $\approx$  2.94 Go]. So, we set the following setting into the **Knime.ini** initialization file.

	-startup
	plugins/org.eclipse.equinox.launcher_1.1.0.v20100507.jar
}	launcher.library
ł	plugins/org.eclipse.equinox.launcher.win32.win32.x86_64_1.1.1.R36x_v20100810
5	-vmargs
5	-Xmx4096m
1	-XX:MaxPermSize=192m
}	-server
)	-Dsun.java2d.d3d=false

Now, Knime can import the dataset. We can launch the decision tree induction process.



The Knime memory occupation is about 3.14 GB during the calculations. Because the amount of available memory is only 4 GB, Windows begins to use intensively its swap file. The computation time is dramatically increased. So, after about 30 minutes (the tree induction is jammed to 3%, see the screenshot above), I stopped the calculations.

I think if we have more memory, we can increase the Xmx parameter of the KNIME.INI configuration file. And so, we can lead the analysis on the whole dataset. A user reports to me that he can achieve the process with a PC with 8 GB RAM. The memory occupation is about 7 GB at the end of the decision tree construction.

From that perspective, because the theoretical limitation is alleviated, the main bottleneck becomes the machine characteristics (the computer mother card capacity and the OS specifications – e.g. Windows 7 - <u>http://en.wikipedia.org/wiki/Windows 7#Physical memory limits</u>).

# 5 RapidMiner

Firefox *						
C Rapid - I +						
rapid-i.com/content/view/26/84/la	ing,en/ 🟫 * 🖒 🚼 rapidminer	P 🏦 🖬				
HOME SEARCH SITEMAP LEGAL CONTACT U	S DEUTSCH					
	Data A					
	Analysi					
	and as loss					
	Data Analysi easy as Lego	REPORT THE FUTURE				
PRODUCTS DOWNLOADS DOCUMENTS	SERVICES COMMUNITY PARTNER SHOP ABOUT US					
Newsletter	Download RapidMiner					
Name	RapidMiner is available in different editions:					
E-mail	<ul> <li>Community Edition&lt; (download on this page)</li> </ul>					
Subscribe	Enterprise Edition = Community Edition + Services + Guarantees					
C 1 552 personnes aiment ça.	It is also possible to purchase a proprietary license (OEM license) of Rapidl Rapidl/liner into own products in cases where open-source license does n (commercial) license of Rapidl/liner.					
	RapidMiner 5.1					
	System	Download Link				
Get Certified!	Windows 32	DOWNLOAD				
Learn more about accelerating your career by becoming a	Windows 64					

We download the 64 bit version from the RapidMiner website. We use the "Community Edition" which is freely downloadable. It is fully functional. The "Enterprise Edition" incorporates some services and guarantees (see <a href="http://rapid-i.com/content/view/181/190/">http://rapid-i.com/content/view/181/190/</a>).

### 5.1 Processing with the standard settings

We do not detail the construction of the process here. For the reader not familiarized with RapidMiner, they can see the tutorial available on our website: <u>http://data-mining-tutorials.blogspot.com/2011/09/new-gui-for-rapidminer-50.html</u>

🚱 Tree Kdd 99 - RapidMiner@Maison-PC	
<u>File E</u> dit <u>P</u> rocess <u>T</u> ools <u>V</u> iew <u>H</u> elp	
🕒 📦 🖬 🖬 🗊 🖍 🔦 🕨 🔳 🛐 🥰	
Process 🕱 👍 XML 🕱	🛛 🗒 Parameters 🛛 💥 🚭 🖘 🔟
🔎 Overview 🛛 🚰 🖅 🔟 🖕 🖛 🖵 🛉 Process 🕨	að • \$? 🔲 🖄 🚽 💈 🖏 🦻 🦻 👳 🛼 •
Repositories X     Main Process	Secision Tree
Imain Process       Imain Process       Imain Process       Read CSV       Set Role	Decision Tree criterion gain_ratio V
Process Control (34)     Process Control (34)     Out	exa or tra mod exa c minimal size for 4
Generation (26)	minimal leaf size 2
🖶 🦳 Export (17) 🕀 🦳 Data Transformation (11	minimal gain 0.1
⊕ 🦳 Modeling (116) ⊕ 🦳 Evaluation (28)	maximal depth
	🚽 \Lambda 3 hidden expert parameters
	Comment X
🔥 Problems 🗶 💦 Log 🗶	(2) Help X
No problems found	• • •
Message	Fixes Location
	💡 Decision Tree
	Synonsis
•	

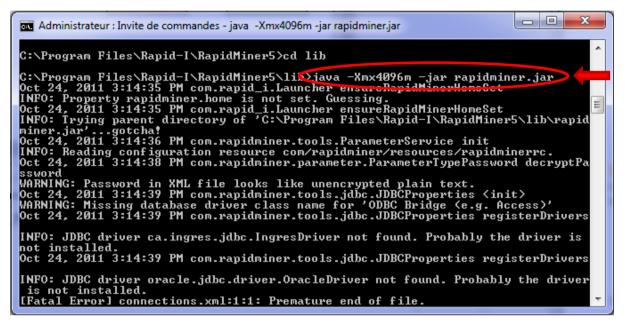
We launch RapidMiner using the shortcut on the desktop. We set the maximum depth of the tree to 4. We start the process, including the importation of the data file. After a few seconds, an error message appears.

🚯 Tree Kdd 99 - RapidMiner@Maison-PC	
Eile Edit Process Tools View Help	
🗋 😋 📰 🚰 💷 💊 🦔 🕨 🖿 🛐 😤 💿	
Process X 🗄 XML X	📑 Parameters 💥 💱 🖨 🔯
🔁 🖽 🖽 - 🗰 - 🛊 - 🛔 💭 Process > 🧳 - 🗊 🔟 🌛 -	🍒 😼 👞 🤛 🗮 🗮 🕶
	Process
Repositories      Main Process      Read CSV     Set Role     Decision Tree	logverbosity init 🔻
	loafile
Ciliar     Ciliar	
Fichier Options Affichage ?	
Applications Processus Services Performance Mise en réseau Utilisateurs	
Nom de limage Nom d'u P Mémoire (jeu de tr Description	
javaw.exe Jaison 2895 572 K Java(TM) Platform SE bin:	
Inclowexe 32 Maison 00 69 672 K Firefox dwm.exe Maison 02 26 320 K Gestionnaire de fenêtres	
dwm.exe Maison 02 26 320 K Gestionnaire de fenêtres POWERPNT.E., Maison 00 17 504 K Microsoft Office PowerPoi	🔼 4 hidden expert parameters
explorer.exe Maison 00 12/304X industrial windows	
at the same in the second se	×
solvové 4.exe Maison 00 4 676 K Print driver host for 32bit	
IAStoricon.ex Maison 00 4 488 K IAStoricon	
argtray.exe Malson 00 3040 K AVG Tray Monitor bg 🦳 This process would need more than the	e maximum amount of available
taskmgr.exe Maison 01 2 628 K Gestionnaire des tâches c memory. You can either leave the process	
taskinot.exe Maison 00 1584K Processus hole pour Tach with more memory, reduce the amount of	·
RapidMiner.e Maison 00 1488 K RapidMiner.exe Operators, optimize the process by using	
csrss.exe 00 1096K schemes, or directly work on databases	systems, e.g. by using the cached
wmdc.exe Maison 00 828 K Gestionnaire pour appare database example source operators.	-
nvvsvc.exe 00 788 K	
Carrêter le processus de tous les utilisateurs Arrêter le processus	<b>√</b> Ok
Processus : 57 UC utilisée : 5% Mémoire physique : 97 %	

The available memory is not enough to achieve the process. We observe that the memory occupation of RapidMiner (through JAVAW.EXE) is 2.67 GB.

### 5.2 Launching from the Windows command line

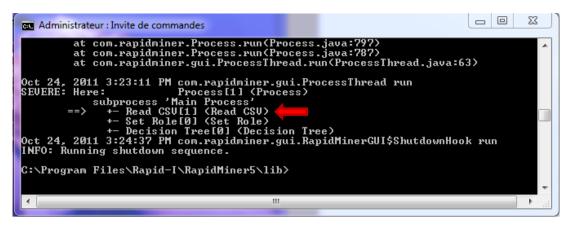
We can modify the settings files into the SCRIPT directory. But, after unsuccessful attempts, I preferred to use the command line in order to control finely the launching settings.



After few minutes, the process has failed again.

Tree Kdd 99 – F									
<u>Eile E</u> dit <u>P</u> roc	ess <u>T</u> ools	View	<u>H</u> elp						
😬 🛸 🔚	<b>F</b>			1 🔳 🛐 🛒 🤅	D				
🔎 Overview	x 🗄 🕁 🖻		Proces	s 🛛 🕞 XML 🗶				🛛 🛃 Par	ameters 🗙 🚰 🖘 🔟
1 million	N.M. JANNIN,				r	a) – 🏫	II 🛛 📣 🗸	- 🕱 🔥	
📲 Gestionnaire d	les tâches de V	Vindows				· · ·			
Fichier Options	Affichage	?							Process
Applications Pro	cessus Service	es Perfor	mance Mise en résea	au Utilisateurs	ole	Decision		logverbosi	ty init .
Nom de l'imag	e Nom d'u	P 1	Mémoire (jeu de tr	Description	exa ori	tra	exa	logfile	
java.exe iexplore.exe explorer.exe POWERPNT.E dwm.exe	Maison	2: 01 01 00 01	28 644 K 9 536 K 8 976 K	Java(TM) Platform SE bin: Internet Explorer Explorateur Windows Microsoft Office PowerPoi Gestionnaire de fenêtres		0			
plugin-contair firefox.exe *	Maison	00	Process Faile		(	-		x	]
IAStorIcon.ex SnippingTool. avgtray.exe	Maison Maison	00 02 00		This process would need memory. You can either h					) expert parameters
iexplore.exe iexplore.exe taskmgr.exe		00 00 00		with more memory, reduc operators, optimize the pi	rocess by usir	ng other learr	ning or preproc	essing	Comment X
iexplore.exe wisptis.exe taskhost.exe	Maison Maison Maison	00 00 00 00 00 00 00 00 00 00 00 00 00		schemes, or directly work database example sourc		systems, e. <u></u>	, by using the	cacheu	8
C Affiche	er les processus	de tous l					(	V Ok	CESS S erator which is the oute
Processus : 63	UC utilisée :	32%	Mémoire physiq	ue : 96 %				most oper	ator of every process.
) [1] Process 3:	37 🔓 [1] Read	d CSV 3	:37						

We note that the memory occupation is 3.17 GB when the process was interrupted. The command window shows that the dataset was not able to be completely imported.



RapidMiner rests on the same technology than Knime (JAVA). No doubt that if we had more memory, he could lead to their term the calculations. Let us note also that RapidMiner proposes us the other strategies to go beyond to the error: the sampling, the using of other data mining method, or working directly inside of a DBMS system. This last suggestion seems really relevant. I think I will study it in a future tutorial. I know that tools such as R propose also this possibility. But, I never tried seriously this kind of solution until now.

### 5.3 Internal coding of the values

The **DATAMANAGEMENT** option of the **READ CSV** operator is very useful for the handling of very large datasets. It enables us to choose the internal coding used for the data representation. DOUBLE\_ARRAY is the default setting. Each value is described by a double precision real. The memory occupation is 8 bytes. For our dataset where we have categorical attributes only, each of them having few values, this choice is oversized. We can dramatically reduce the memory occupation by using a suited coding strategy. For instance, if we use the BYTE\_ARRAY encoding, that is fully enough for our dataset, we can divide by 8 the memory occupation!

### 6 Conclusion

The shift to the 64-bit version of the data mining tools allows to benefit of the capacities of more powerful computers. We can load a larger dataset and perform the calculations into main memory. About our experiments, if we have a computer with 8 GB RAM, we can complete the analysis.

But, on other hand, this shift from a 32-bit system to a 64-bit one is not really useful if the characteristics of the computer are not improved. Especially, on a low-power machine, the solution such as the one implemented into Sipina, where we write the table into temporary files on the disk, remains a valuable solution.