



RAPIDMINER
FREE SOFTWARE FOR DATA
MINING, ANALYTICS AND
BUSINESS INTELLIGENCE

Luigi Grimaudo (luigi.grimaudo@polito.it)
DataBase And Data Mining Research Group (DBDMG)

Summary

- RapidMiner project
- Strengths
- How to use RapidMiner
- Operator highlights
- RapidMiner GUI
- References

RapidMiner Project

- A fully integrated environment for machine learning, data mining, text mining, predictive analytics and business intelligence
- It is distributed under the AGPL open source license and has been hosted by SourceForge since 2004
- It can be used as a stand-alone application for data analysis or as a data mining engine for the integration into own code

How to use RapidMiner

- RapidMiner can be used in several ways:
 - ▣ As a standalone tool by means of the simple GUI, connecting the requested operators to build your process, executing it and getting its result directly in the RapidMiner environment
 - ▣ As a batch process one can build the workflow by means of the GUI and then execute it running the RapidMiner script with the XML process as input
 - ▣ As a Java API one can integrate the RapidMiner facilities in your own data mining or business intelligence code building the requested process directly inside the java code
 - ▣ As an hybrid solution one can build the process with the GUI to executing and to managing it inside a Java code

Operator highlights (1)

- Data mining modeling:
 - Support Vector Machines (SVM),
 - Rule learners
 - Decision trees
 - Bayes
 - Gaussian Processes
 - Neural Networks
 - Evolutionary optimization
 - Boosting
 - Apriori
 - FPGrowth
 - Clustering
 - and many others

Operator highlights (2)

- Data Transformations:
 - Aggregation
 - Discretization
 - Normalization
 - Filter
 - Sampling
 - PCA
 - Missing value replenishment
 - Lot more

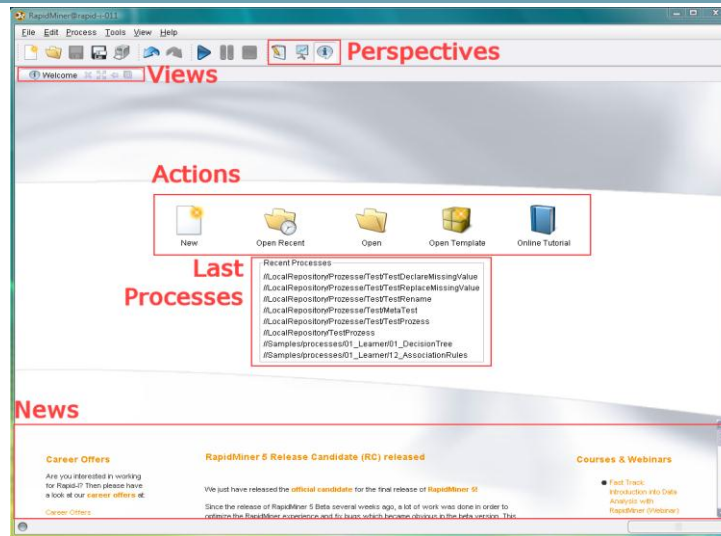
Operator highlights (3)

- Evaluation:
 - ▣ Cross-validation
 - ▣ Leave-one-out
 - ▣ Sliding time windows
 - ▣ Back testing
 - ▣ Significance tests
 - ▣ ROC
 - ▣ Etc.

Download and launch RapidMiner

- Download:
 - <http://sourceforge.net/projects/rapidminer/files>
- Launch:
 - Double click:
rapidminer-5.*\rapidminer\lib\rapidminer.jar
 - Or from command line prompt, in the RapidMiner root directory:
java -jar ./lib/rapidminer.jar

Welcome Perspective



Repository (1)



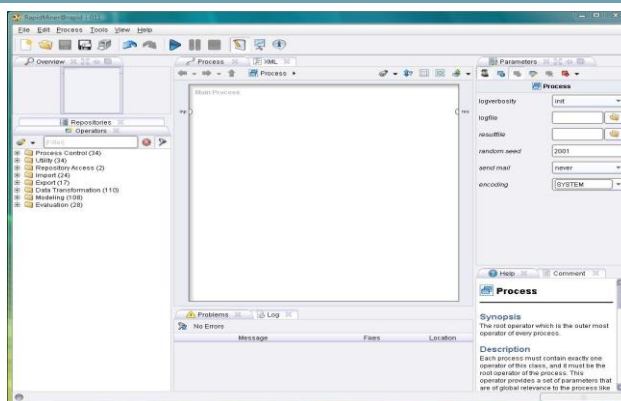
When you launch RapidMiner for the first time, it asks you to create a new **Repository** to store/load processes and data.

Repository (2)



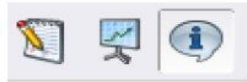
You have to set the **name** of the new Repository and the **path** of its physical location on the disk.

RapidMiner GUI



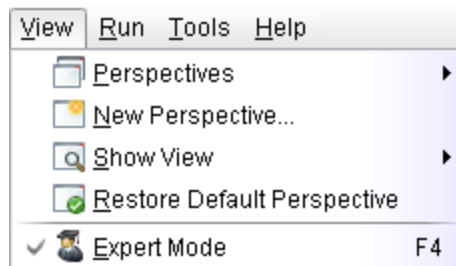
Creating a new process, the GUI generates an **XML file** that defines the analytical processes the user wishes to apply to the data. This file is then read by the RapidMiner engine to run the analyses automatically. While these are running, the GUI can also be used to interactively control and inspect running processes.

RapidMiner GUI – Perspectives



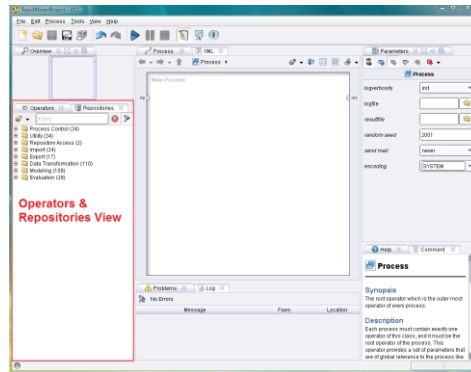
- **Design Perspective:** is the central RapidMiner perspective where all analysis processes are created and managed
- **Result Perspective:** If a process supplies results then RapidMiner takes you to this Result Perspective
- **Welcome Perspective:** first perspective when RapidMiner is lunched, where you can see the last executed processes and some logs.

Expert mode



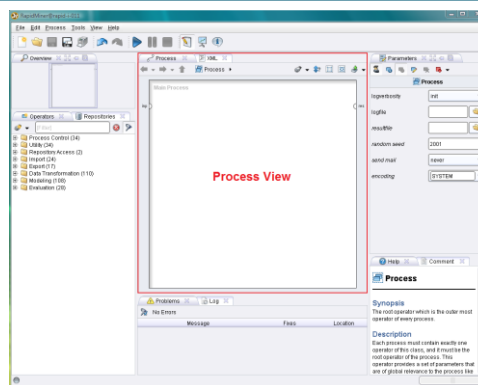
To set your custom parameters for the models you need to enable the **expert mode**.

Design Perspective



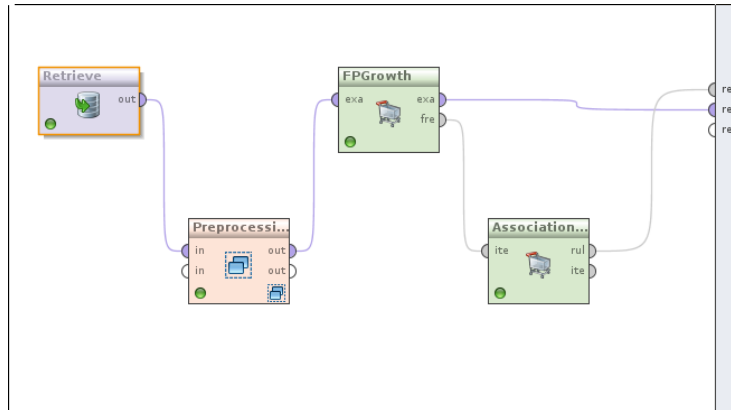
In this view all the work steps (called **operators**) available in RapidMiner are presents and they are used as building block for every process. The **repository** section serves for the management and structuring of your analysis processes into projects and at the same time as both a source of data as well as of the associated metadata.

Design Perspective



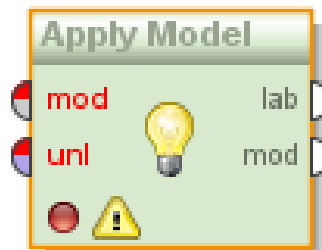
The process view shows the individual steps within the analysis process as well as their connections. New steps can be added to the current process. Connections between them can de defined and detached.

Operators (1)



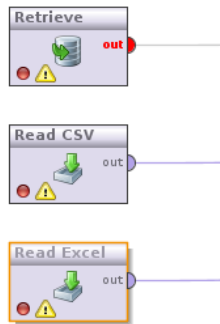
Working with RapidMiner fundamentally consists in defining analysis process by indicating a **succession of operators**.

Operators (2)



The inputs and outputs of operators are generated and consumed by **ports**. Every operator is defined by its **inputs, outputs, action performed** and **parameters**.

Operators – Load data

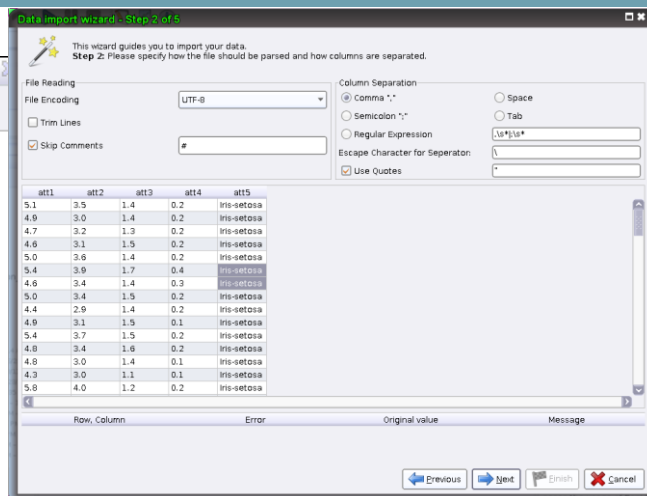


You can load data in different ways: from **repository**, **csv**, **excel**, etc..

Store data into Repository (1)



To store a dataset into Repository you can use the **Data import Wizard**.



Store data into Repository (2)

Data import wizard - Step 4 of 5

This wizard guides you to import your data.
Step 4: RapidMiner uses strongly typed attributes. In this step, you can define the data types of your attributes. Furthermore, RapidMiner assigns roles to the attributes, defining what they can be used for by the individual operators. These roles can be also defined here. Finally, you can rename attributes or deselect them entirely.

Reload data Guess value types Preview uses only first 100 rows.

	att1	att2	att3	att4	att5
	real	real	real	real	binco...
	[tribute]	[tribute]	[tribute]	[tribute]	[tribute]
5.100	3.500	1.400	0.200	Iris-setosa	
4.900	3	1.400	0.200	Iris-setosa	
4.700	3.200	1.300	0.200	Iris-setosa	
4.600	3.100	1.500	0.200	Iris-setosa	
5	3.600	1.400	0.200	Iris-setosa	

0 errors. Ignore errors Show only errors

Row, Column	Error	Original value	Message
-------------	-------	----------------	---------

Data import wizard - Step 5 of 5

This wizard guides you to import your data.
Step 5: Please specify a repository location.

- Juneteierarchical (load)
- april_exp (load)
- article_exp_dir (load)
- clustering (load)
- feature_selection (load)
- find_best_feature_selection_method_73761_FILES (load)
- http (load)
- new_feature (load)
- p2p (load)
- piera_process (load)
- prov_more (load)
- unk (load)
- brackOS (load - v1, 10/20/11 9:14 AM - 0.8 MB)
- data_1_0_08 (load - v1, 6/20/11 6:39 PM - 340 MB)
- data_1_0_08_h2 (load - v1, 6/20/11 9:50 PM - 148 MB)
- data_1_9 (load - v1, 6/20/11 10:30 PM - 313 MB)
- flat (load - v1, 6/20/11 11:17 PM - 146 MB)
- prov (load - v1, 7/13/11 6:02 AM - 88 MB)
- seed1 (load - v1, 10/25/11 2:30 PM - 6.9 MB)
- temp (load - v1, 8/22/11 10:17 PM - 289 MB)
- unk1 (load - v1, 7/18/11 8:15 PM - 305 MB)
- unk2 (load - v1, 7/18/11 9:48 PM - 276 MB)

Name: newData
 Location: /MyRepository/newData

Process – Discretization

FrequencyDiscretization (Discretize by Frequency)

create view

attribute filter type: all

invert selection

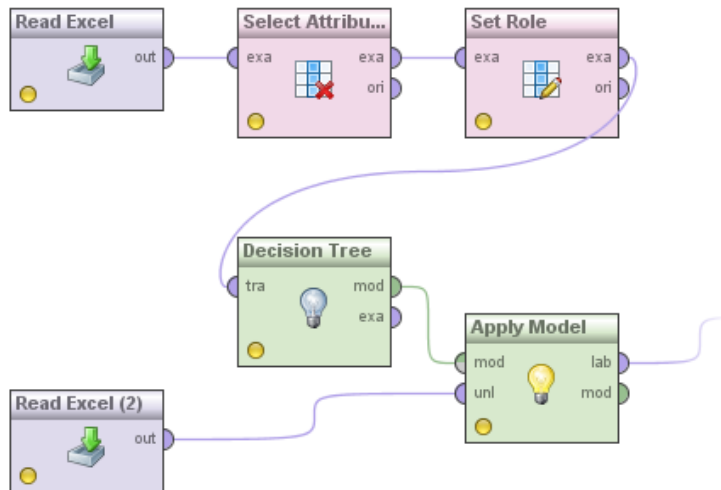
include special attributes

use sqrt of examples

number of bins: 5

range name type: long

Process – Classification



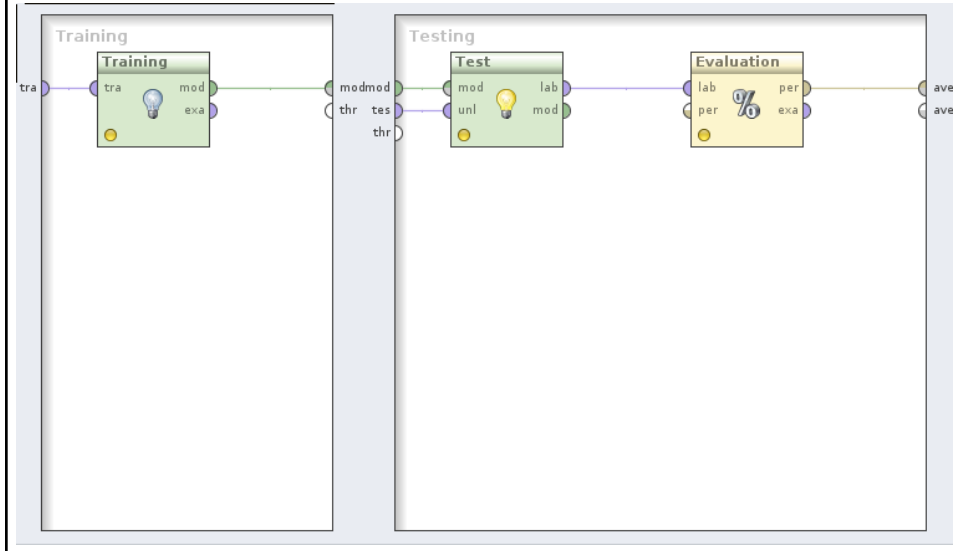
Process – Validation (1)

Main Process

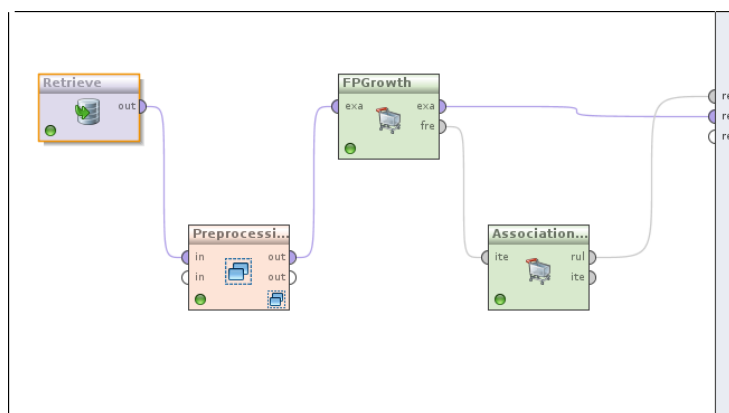
The screenshot shows a 'Main Process' window. On the left, a 'Retrieve' node is connected to an 'XVal' node. The 'XVal' node is expanded to show its configuration options. The options are as follows:

- average performances only*
- leave one out*
- number of validations: 10
- sampling type: shuffled sampling
- use local random seed*
- parallelize training*
- parallelize testing*
- Compatibility level: 5.1.011

Process – Validation (2)



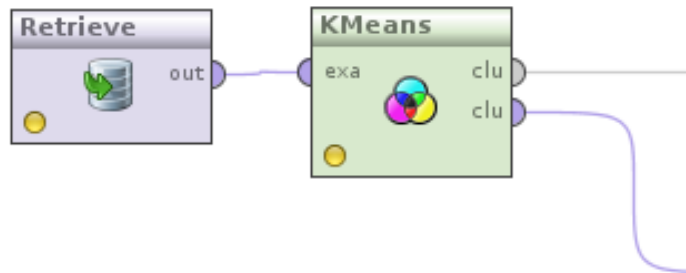
Process – Association rule extraction



Remember: transactional data in RapidMiner are treated as **binominal values** (Item presents: true/false). You can use a pre-processing operator to do this conversion.

Process - Clustering

Main Process

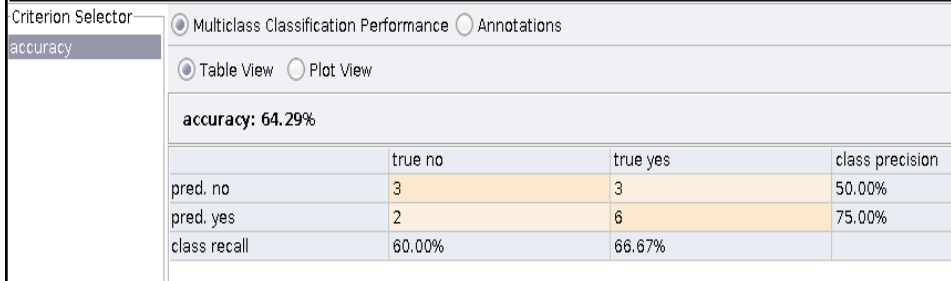


Result Perspective

Store No.	store_id	product_name	item_price
1	Store 01	Toys	270.739
2	Store 02	Books	121.172
3	Store 12	Books	483.063
4	Store 06	Books	84.493
5	Store 01	Clothing	320.077
6	Store 11	Sports	169.434
7	Store 10	Health	80.919
8	Store 10	Health	486.522
9	Store 11	Health	141.132
10	Store 14	Toys	341.362
11	Store 14	Health	409.243
12	Store 12	Sports	234.410
13	Store 08	Electronics	713.422
14	Store 06	HomeGoods	264.913
15	Store 01	HomeGoods	234.730
16	Store 08	Toys	14.344
17	Store 06	Electronics	203.100
18	Store 06	Books	173.497
19	Store 14	Books	321.195
20	Store 09	Books	416.413
21	Store 14	Books	402.084
22	Store 02	Toys	340.063
23	Store 06	Electronics	97.492

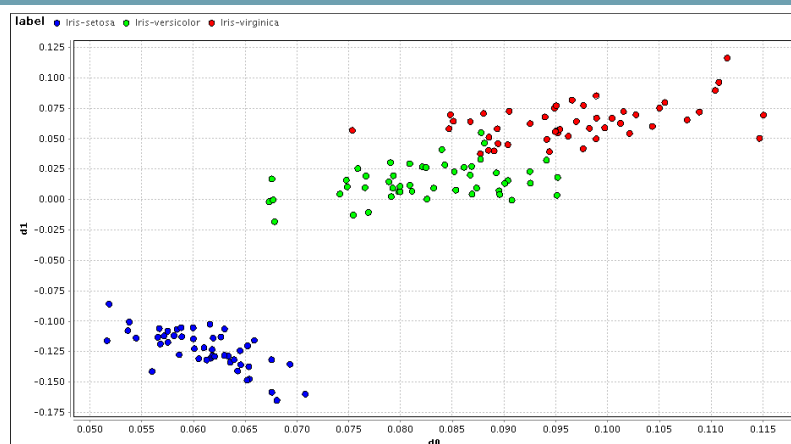
Objects which are placed at the result ports at the right-hand side of a process are automatically displayed in the Result Perspective after the process is completed. Each currently opened and indicated result is displayed as an additional tab in this area.

Plot View – Confusion Matrix



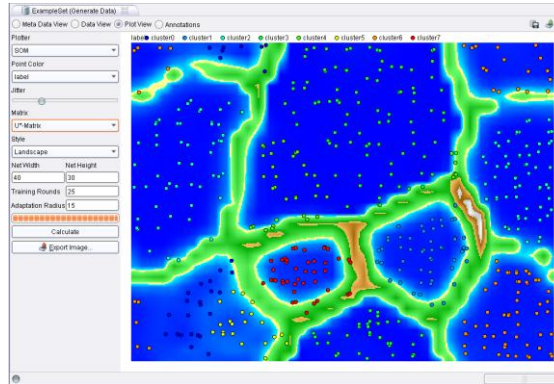
One of the strongest features of RapidMiner are the numerous visualisation methods for data, other tables, models and results offered in the Plot View.

Plot View – Clustering result



One of the strongest features of RapidMiner are the numerous visualisation methods for data, other tables, models and results offered in the Plot View.

Plot View – Complex plot



One of the strongest features of RapidMiner are the numerous visualisation methods for data, other tables, models and results offered in the Plot View.

References

- Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M. and Euler, T., Yale (now: *RapidMiner*): Rapid Prototyping for Complex Data Mining Tasks. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*
- <http://rapid-i.com>