



RAPIDMINER
FREE SOFTWARE FOR DATA
MINING, ANALYTICS AND
BUSINESS INTELLIGENCE

Luigi Grimaudo (luigi.grimaudo@polito.it)
Data Base 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

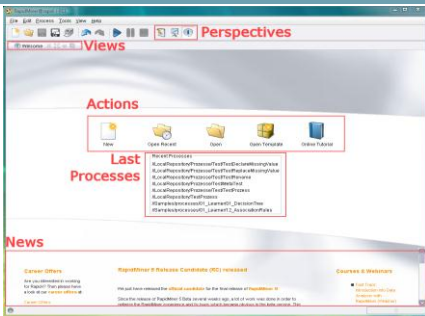
Operator highlights (3)

- 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/1.%20ORapidMiner/5.1/rapidminer-5.1.014.zip/download>
- **Launch:**
 - Double click:
rapidminer-5.1.014\rapidminer\lib\rapidminer.jar
 - Or from command prompt, in the RapidMiner root directory:
java -jar ./lib/rapidminer.jar

Welcome Perspective

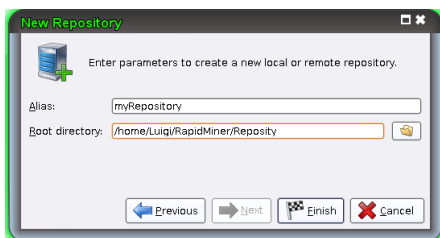


Repository (1)



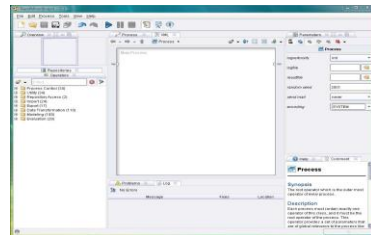
When you launch RapidMiner for the first time, it asks you to create a new **Repository** to store/load processes and data.

Repository (2)



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

RapidMiner GUI



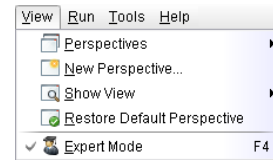
Creating a new process, the GUI generates an **XML file** that defines the analytical processes the user wishes to apply to the data. This file is then read by the RapidMiner engine to run the analyses automatically. While these are running, the GUI can also be used to interactively control and inspect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

RapidMiner GUI – Perspectives



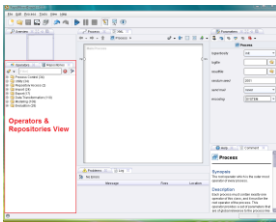
- **Design Perspective:** is the central RapidMiner perspective where all analysis processes are created and managed
- **Result Perspective:** If a process supplies results then RapidMiner takes you to this Result Perspective
- **Welcome Perspective:** first perspective when RapidMiner is launched, where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

Expert mode



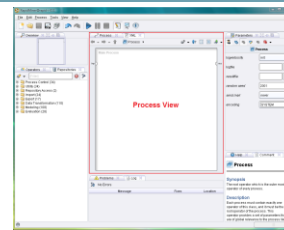
To set your custom parameters for the models you need to enable the **expert mode**.

Design Perspective



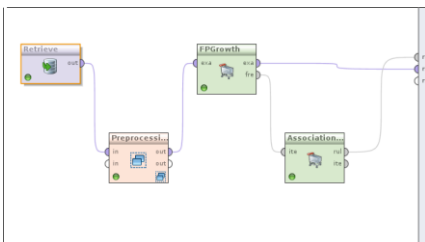
In this view all the work steps (called **operators**) available in RapidMiner are presented and they are used as building block for every process. The **repository** section serves for the management and structuring of your analysis processes into projects and at the same time as both a source of data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

Operators (1)



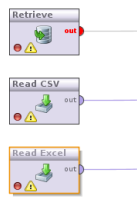
Working with RapidMiner fundamentally consists in defining analysis process by indicating a **succession of operators**.

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)

The screenshot shows the 'Data import wizard' dialog box. It has a 'File Reading' section with 'File name' set to 'com-0'. There are checkboxes for 'Use cases', 'Skip comments', 'Column separation', 'Delimiter', 'Regular Expression', and 'Escape Character for Separator'. A preview window shows a table with 5 columns (att1 to att5) and 10 rows of data. At the bottom, there are 'Previous', 'Next', 'Cancel', and 'OK' buttons.

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

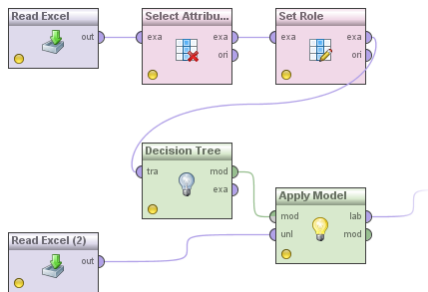
Store data into Repository (2)

The first screenshot shows the 'Data import wizard - Step 4 of 6' dialog box. It has a 'Guess value types' section with a table of data and a 'Preview' section showing the first 100 rows. The second screenshot shows the 'Data import wizard - Step 5 of 6' dialog box, which is for specifying the repository location.

Process – Discretization

The screenshot shows the 'Frequency Discretization (Discretize by Frequency)' operator. The 'Parameters' section includes: 'attribute filter type' set to 'all', 'invert selection' (unchecked), 'include special attributes' (unchecked), 'use sqrt of examples' (unchecked), 'number of bins' set to 5, and 'range name type' set to 'long'.

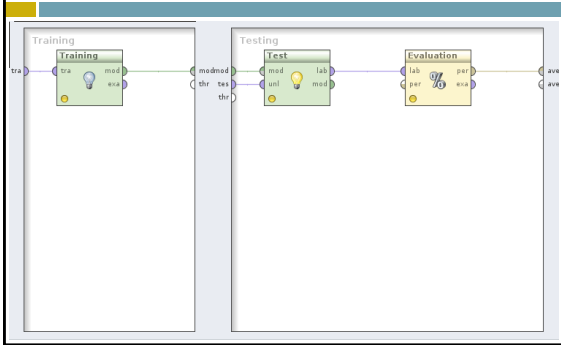
Process – Classification



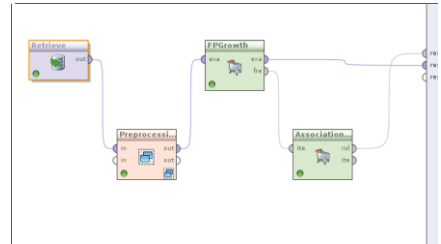
Process – Validation (1)

The screenshot shows the 'XVal (X-Validation)' operator. The 'Parameters' section includes: 'average performances only' (checked), 'leave one out' (unchecked), 'number of validations' set to 10, 'sampling type' set to 'shuffled sampling', 'use local random seed' (unchecked), 'parallelize training' (unchecked), and 'parallelize testing' (unchecked). The 'Compatibility level' is set to 5.1.011.

Process – Validation (2)

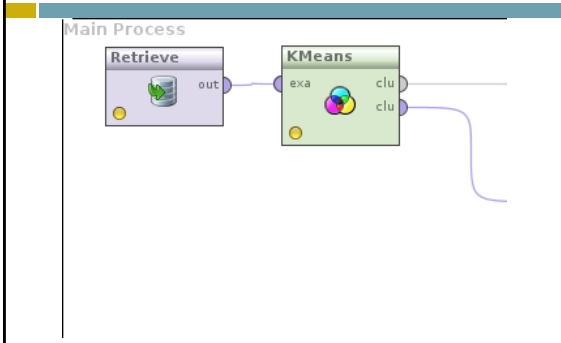


Process – Association rule extraction

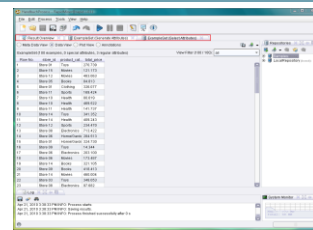


Remember: transactional data in RapidMiner are treated as **binominal values** (Item presents: true/false). You can use a pre-processing operator to do this conversion.

Process - Clustering



Result Perspective



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

Criterion Selector: Multiclass Classification Performance Annotations

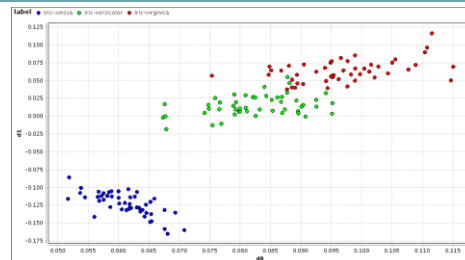
Accuracy: Table View Plot View

accuracy: 64.29%

	true no	true yes	class precision
pred. no	3	3	50.00%
pred. yes	2	6	75.00%
class recall	60.00%	66.67%	

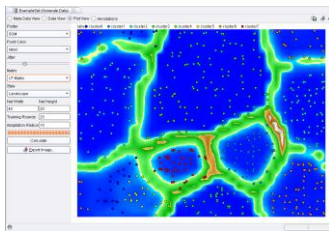
One of the strongest features of RapidMiner are the numerous visualisation methods for data, other tables, models and results offered in the Plot View.

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>