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

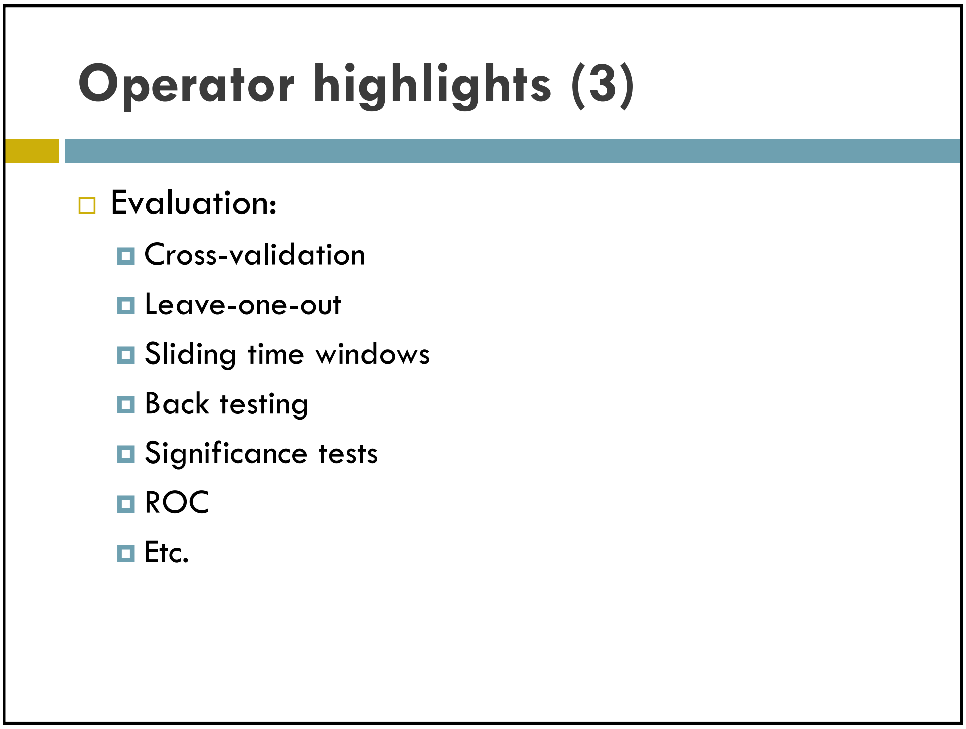
- RapidMiner can be used in several ways:
  - As a standalone tool by means of the simple GUI, connecting the requested operators to build your process, executing it and getting its result directly in the RapidMiner environment
  - As a batch process one can build the workflow by means of the GUI and 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

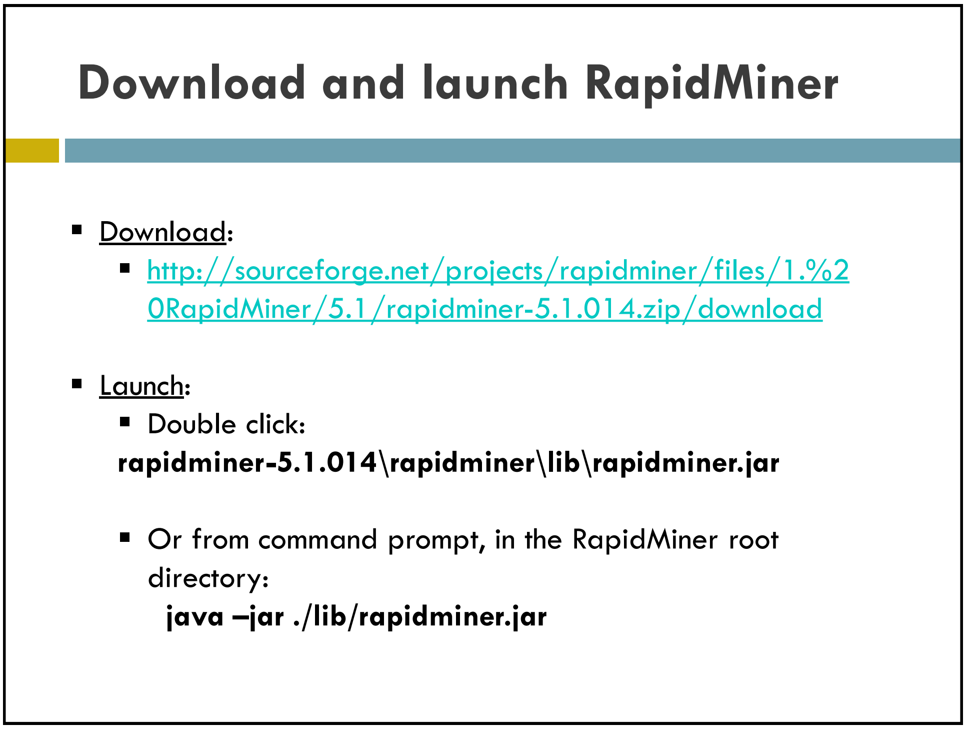
Operator highlights (2)

- 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

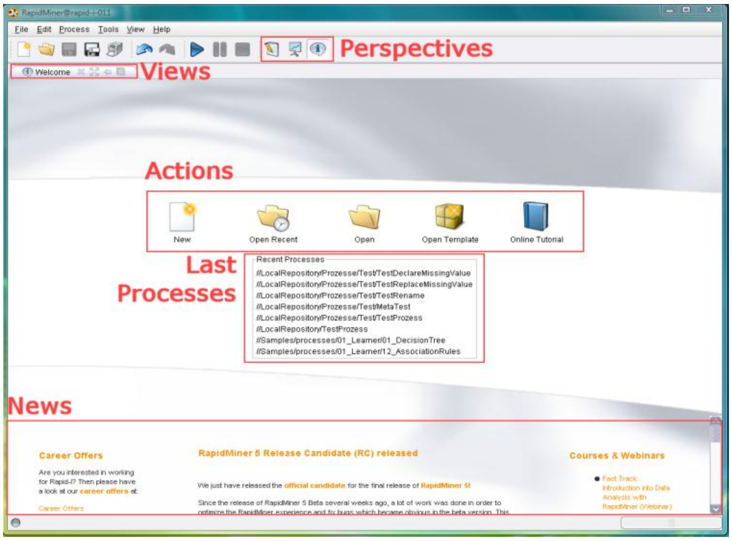
- Download:
  - http://sourceforge.net/projects/rapidminer/files/1.%20RapidMiner/5.1/rapidminer-5.1.014.zip/download
- Launch:
  - Double click:

rapidminer-5.1.014\rapidminer\lib\rapidminer.jar

  - Or from command prompt, in the RapidMiner root directory:

```
java –jar ./lib/rapidminer.jar
```

Welcome Perspective



Recent Processes

//LocalRepository/Prozesse/Test/TestDeclareMissingValue
//LocalRepository/Prozesse/Test/TestReplaceMissingValue
//LocalRepository/Prozesse/Test/TestRename
//LocalRepository/Prozesse/Test/MetaTest
//LocalRepository/Prozesse/Test/TestProzess
//LocalRepository/TestProzess
//Samples/processes/01_Learner/01_DecisionTree
//Samples/processes/01_Learner/12_AssociationRules

Career Offers

Are you interested in working for Rapid-I? Then please have a look at our **career offers** at:

Career Offers

RapidMiner 5 Release Candidate (RC) released

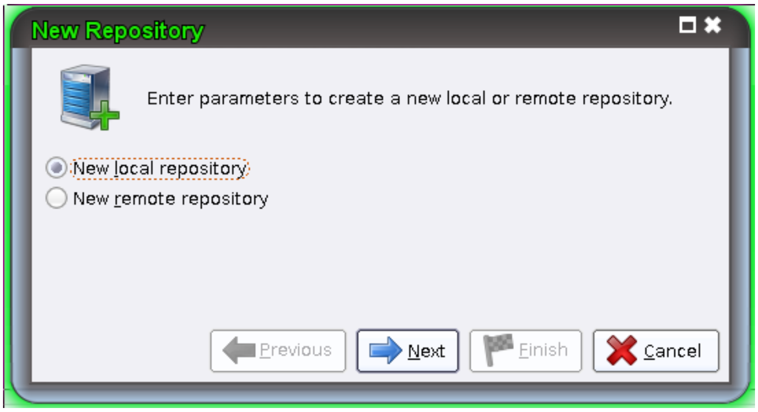
We just have released the **official candidate** for the final release of **RapidMiner 5!**

Since the release of RapidMiner 5 Beta several weeks ago, a lot of work was done in order to optimize the RapidMiner experience and fix bugs which became obvious in the beta version. This

Courses & Webinars

- Fast Track: Introduction into Data Analysis with RapidMiner (Webinar)

Repository (1)

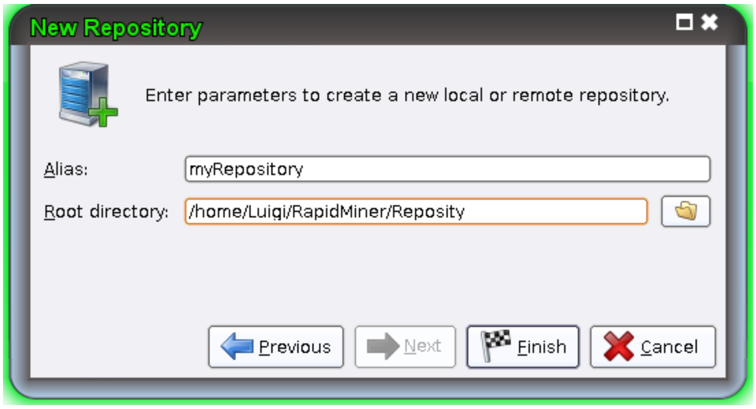


Enter parameters to create a new local or remote repository.

● New local repository
○ New remote repository

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

Repository (2)



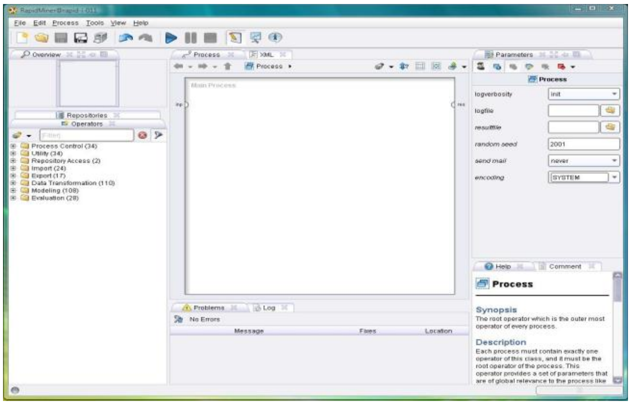
Enter parameters to create a new local or remote repository.

Alias: myRepository

Root directory: /home/Luigi/RapidMiner/Reposity

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

RapidMiner GUI



No Errors

Process

logverbosity: init

logfile:

resultfile:

random seed: 2001

send mail: never

encoding: SYSTEM

Process

Synopsis

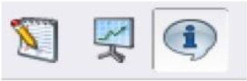
The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like

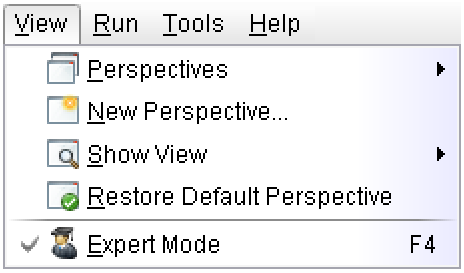
Creating a new process, the GUI generates an **XML file** that defines the analytical processes the user wishes to apply to the data. This file is then read by the RapidMiner engine to run the analyses automatically. While these are running, the GUI can also be used to interactively control and inspect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

RapidMiner GUI – Perspectives



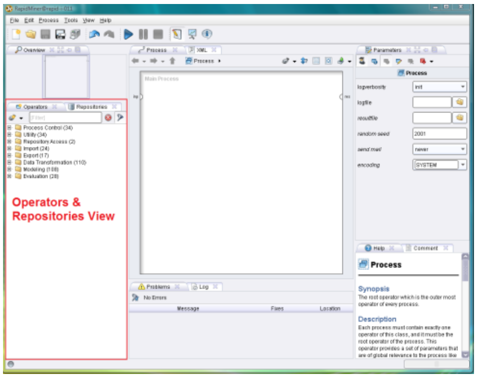
- **Design Perspective:** is the central RapidMiner perspective where all analysis processes are created and managed
- **Result Perspective:** If a process supplies results then RapidMiner takes you to this Result Perspective
- **Welcome Perspective:** first perspective when RapidMiner is lunched, where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



Operators & Repositories View

Process

logverbosity: init

logfile

resultfile

random seed: 2001

send mail: never

encoding: SYSTEM

Process

Synopsis

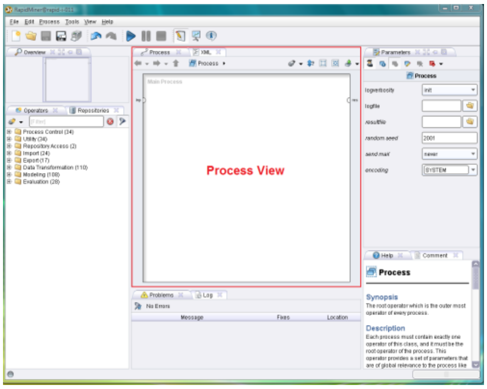
The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like

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



Process View

Process

logverbosity: init

logfile

resultfile

random seed: 2001

send mail: never

encoding: SYSTEM

Process

Synopsis

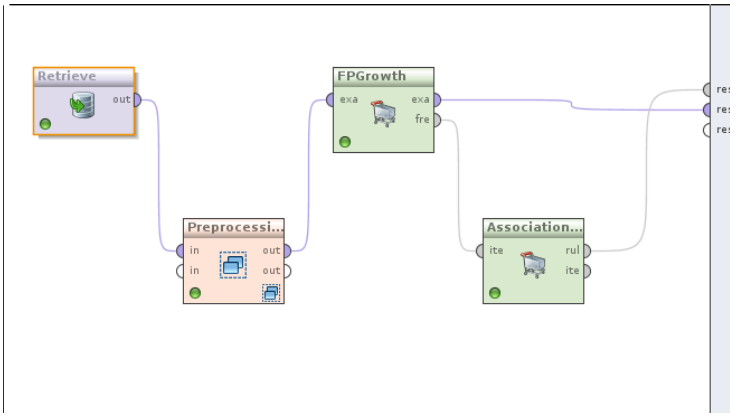
The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like

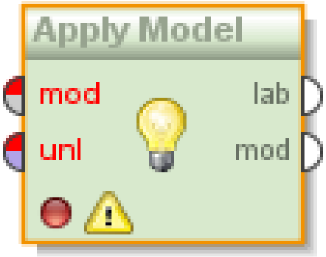
The process view shows the individual steps within the analysis process as well as their connections. New steps can be added to the current process. Connections between them can de defined and detached.

Operators (1)



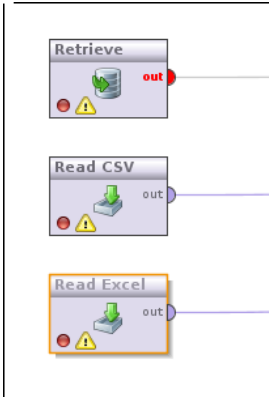
Working with RapidMiner fundamentally consists in defining analysis process by indicating a **succession of operators**.

Operators (2)



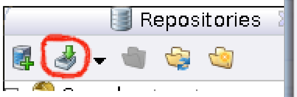
The inputs and outputs of operators are generated and consumed by **ports**. Every operator is defined by its **inputs**, **outputs**, **action performed** and **parameters**.

Operators – Load data

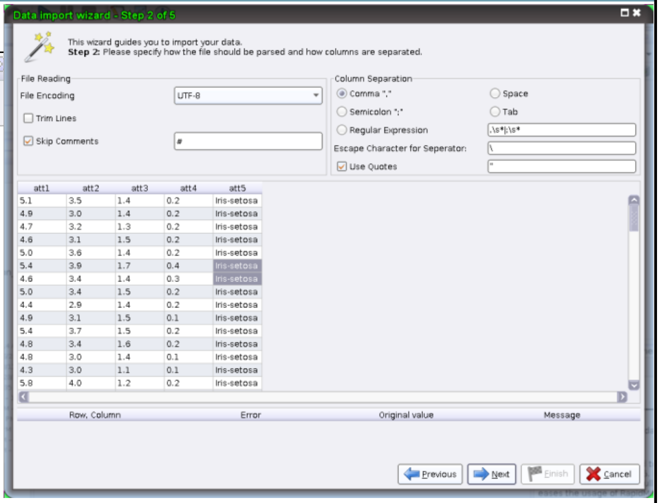


You can load data in different ways: from **repository**, **csv**, **excel**, etc..

Store data into Repository (1)



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



This wizard guides you to import your data.
Step 2: Please specify how the file should be parsed and how columns are separated.

File Reading

File Encoding: UTF-8

□ Trim Lines

☑ Skip Comments: #

Column Separation

● Comma ","
○ Semicolon ";"
○ Regular Expression: ,\s*|;\s*
○ Space
○ Tab

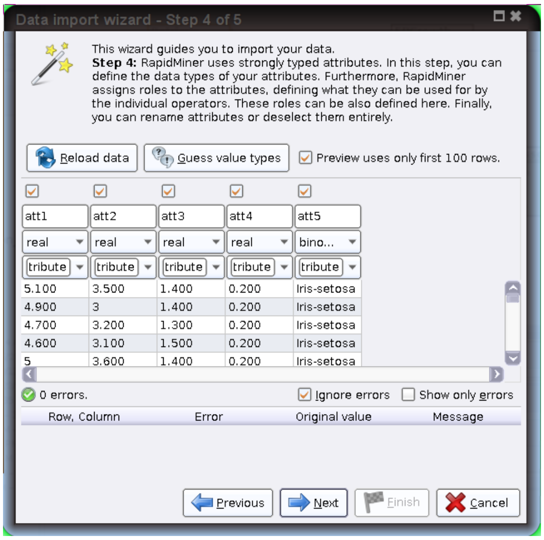
Escape Character for Seperator: \

☑ Use Quotes: "

att1	att2	att3	att4	att5
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5.0	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3.0	1.4	0.1	Iris-setosa
4.3	3.0	1.1	0.1	Iris-setosa
5.8	4.0	1.2	0.2	Iris-setosa

Row, Column	Error	Original value	Message

Store data into Repository (2)



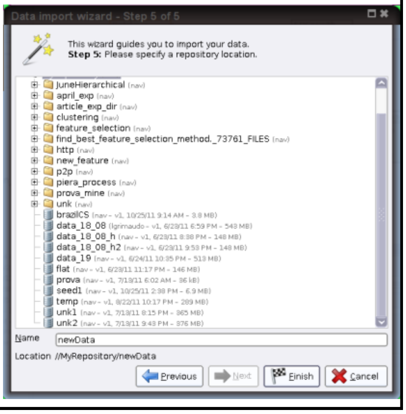
This wizard guides you to import your data.
Step 4: RapidMiner uses strongly typed attributes. In this step, you can define the data types of your attributes. Furthermore, RapidMiner assigns roles to the attributes, defining what they can be used for by the individual operators. These roles can be also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

☑ Preview uses only first 100 rows.

☑	☑	☑	☑	☑
att1	att2	att3	att4	att5
real	real	real	real	bino...
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

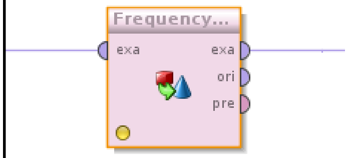


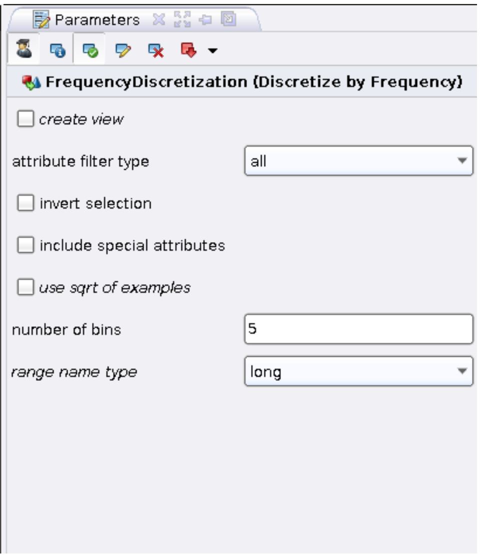
This wizard guides you to import your data.
Step 5: Please specify a repository location.

```
⊞ JuneHierarchical (nav)
⊞ april_exp (nav)
⊞ article_exp_dir (nav)
⊞ clustering (nav)
⊞ feature_selection (nav)
⊞ find_best_feature_selection_method._73761_FILES (nav)
⊞ http (nav)
⊞ new_feature (nav)
⊞ p2p (nav)
⊞ piera_process (nav)
⊞ prova_mine (nav)
⊞ unk (nav)
  brazilCS (nav – v1, 10/25/11 9:14 AM – 3.8 MB)
  data_18_08 (lgrimaudo – v1, 6/23/11 6:59 PM – 543 MB)
  data_18_08_h (nav – v1, 6/23/11 8:38 PM – 148 MB)
  data_18_08_h2 (nav – v1, 6/23/11 9:53 PM – 148 MB)
  data_19 (nav – v1, 6/24/11 10:35 PM – 513 MB)
  flat (nav – v1, 6/23/11 11:17 PM – 146 MB)
  prova (nav – v1, 7/13/11 6:02 AM – 36 kB)
  seed1 (nav – v1, 10/25/11 2:38 PM – 6.9 MB)
  temp (nav – v1, 8/22/11 10:17 PM – 289 MB)
  unk1 (nav – v1, 7/13/11 8:15 PM – 365 MB)
  unk2 (nav – v1, 7/13/11 9:43 PM – 376 MB)
```

Name: newData
Location: //MyRepository/newData

Process – Discretization





FrequencyDiscretization (Discretize by Frequency)

□ *create view*

attribute filter type: all

□ invert selection

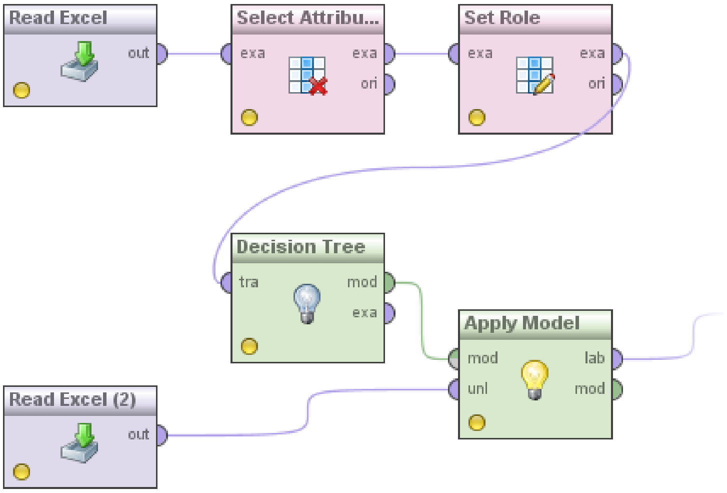
□ include special attributes

□ *use sqrt of examples*

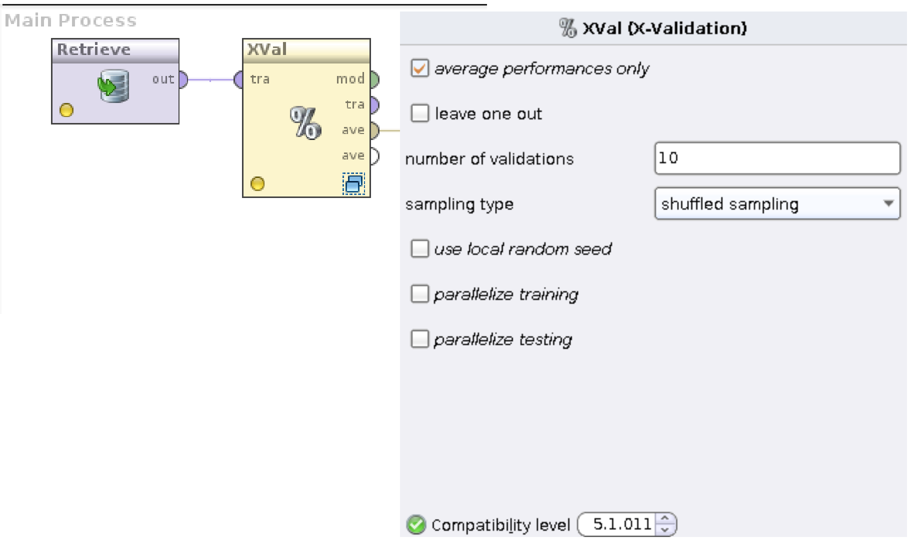
number of bins: 5

range name type: long

Process – Classification



Process – Validation (1)



XVal (X-Validation)

☑ average performances only

□ leave one out

number of validations 10

sampling type shuffled sampling

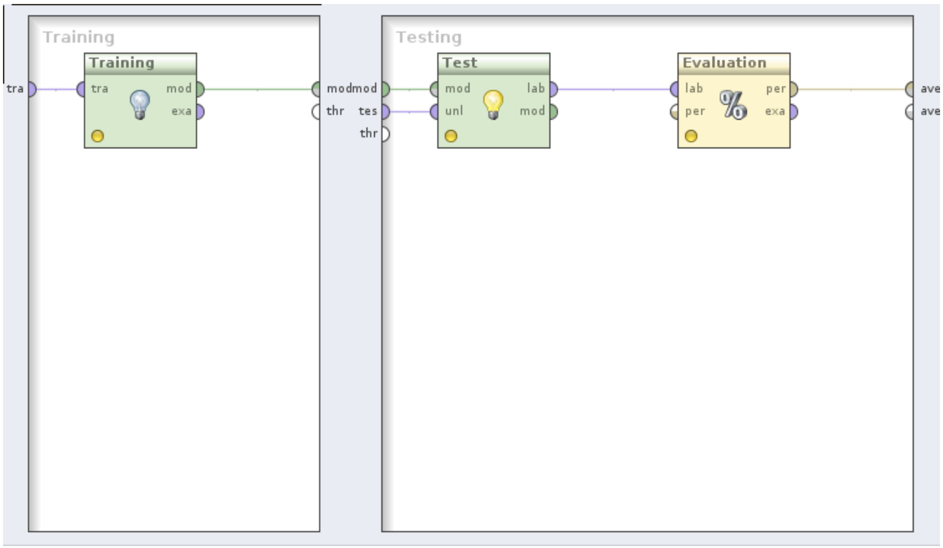
□ use local random seed

□ parallelize training

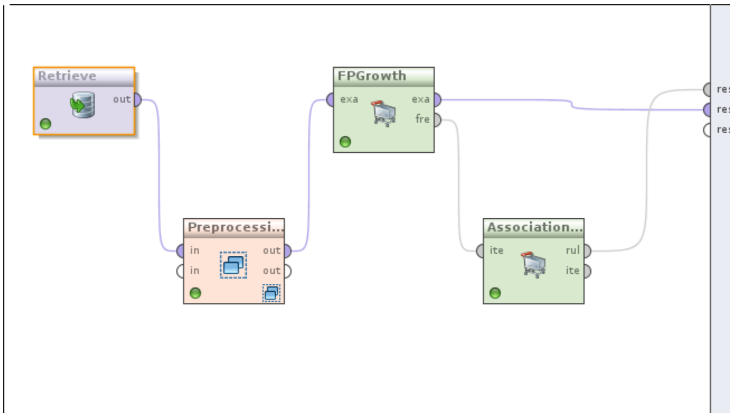
□ parallelize testing

Compatibility level 5.1.011

Process – Validation (2)

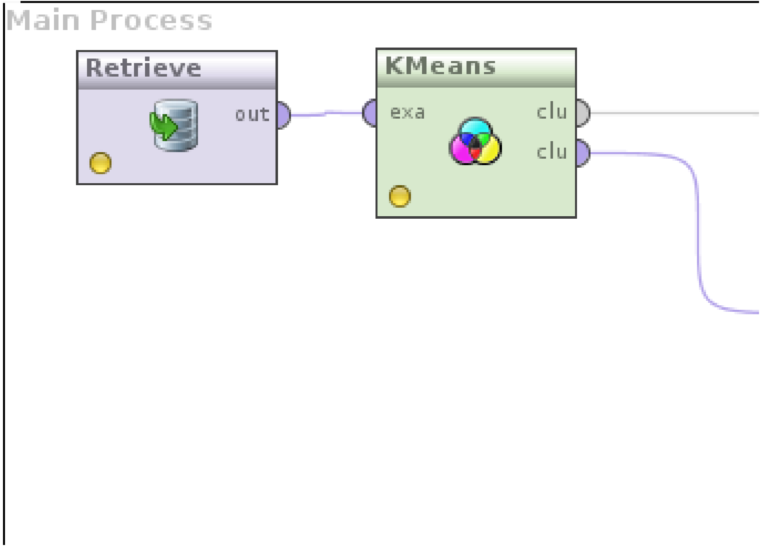


Process – Association rule extraction

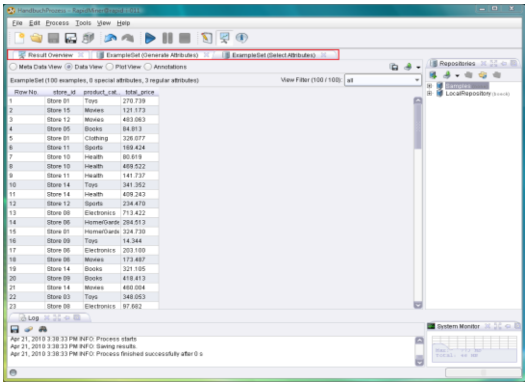


Remember: transactional data in RapidMiner are treated as **binominal values** (Item presents: true/false). You can use a pre-processing operator to do this conversion.

Process - Clustering



Result Perspective


HandbuchProzess – RapidMiner@rapid-i-011
File Edit Process Tools View Help
Result Overview
ExampleSet (Generate Attributes)
ExampleSet (Select Attributes)
○ Meta Data View ● Data View ○ Plot View ○ Annotations
Repositories
Samples
LocalRepository
Log
System Monitor

ExampleSet (100 examples, 0 special attributes, 3 regular attributes)

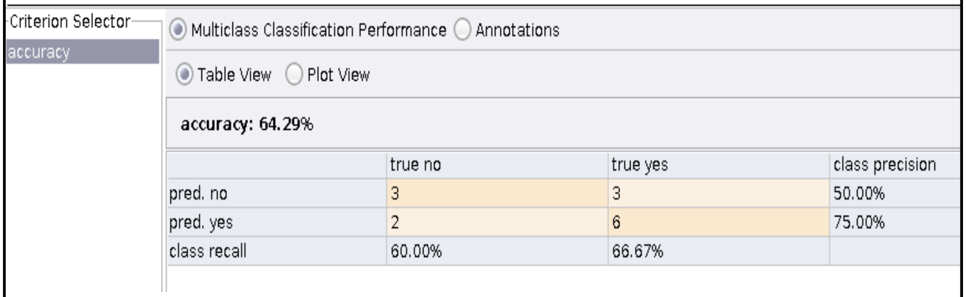
View Filter (100 / 100): all

Row No.	store_id	product_cat...	total_price
1	Store 01	Toys	270.739
2	Store 15	Movies	121.173
3	Store 12	Movies	483.063
4	Store 05	Books	84.813
5	Store 01	Clothing	326.077
6	Store 11	Sports	169.424
7	Store 10	Health	80.619
8	Store 10	Health	469.522
9	Store 11	Health	141.737
10	Store 14	Toys	341.352
11	Store 14	Health	409.243
12	Store 12	Sports	234.470
13	Store 08	Electronics	713.422
14	Store 06	Home/Garde	284.513
15	Store 01	Home/Garde	324.730
16	Store 09	Toys	14.344
17	Store 06	Electronics	203.100
18	Store 08	Movies	173.487
19	Store 14	Books	321.105
20	Store 09	Books	418.413
21	Store 14	Movies	460.004
22	Store 03	Toys	348.053
23	Store 08	Electronics	97.682

```
Apr 21, 2010 3:38:33 PM INFO: Process starts
Apr 21, 2010 3:38:33 PM INFO: Saving results.
Apr 21, 2010 3:38:33 PM INFO: Process finished successfully after 0 s
```

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

accuracy

● Multiclass Classification Performance ○ Annotations

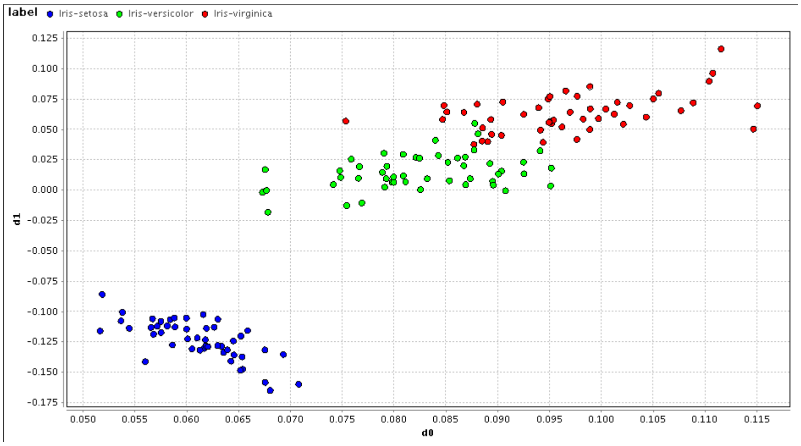
● Table View ○ Plot View

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%	

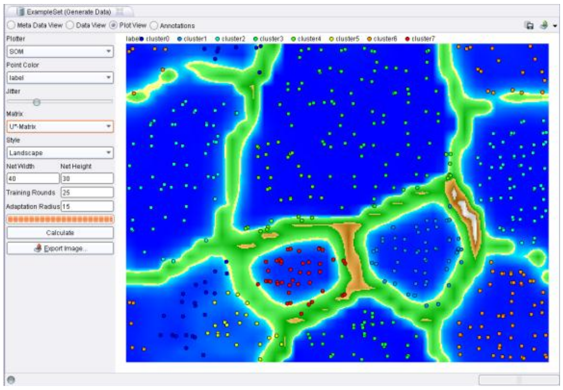
One of the strongest features of RapidMiner are the numerous visualisation methods for data, other tables, modells and results offered in the Plot View.

Plot View – Clustering result



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

Plot View – Complex plot


ExampleSet (Generate Data)
○ Meta Data View ○ Data View ● Plot View ○ Annotations
label ● cluster0 ● cluster1 ● cluster2 ● cluster3 ● cluster4 ● cluster5 ● cluster6 ● cluster7
Calculate
Export Image...

Plotter: SOM
Point Color: label
Jitter
Matrix: U*-Matrix
Style: Landscape
Net Width: 40
Net Height: 30
Training Rounds: 25
Adaptation Radius: 15

One of the strongest features of RapidMiner are the numerous visualisation methods for data, other tables, modells 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