

Getting Started with RapidMiner

Download tool and datasets:

<http://rapidminer.com/>

*Agathe Merceron
Berliner University of Applied Sciences
Germany*



Outline

Design and result perspectives in RapidMiner.

The Point Dataset Point1.csv:

- Reading the data

- Visual exploration to understand the data and guide the work

- Clustering and the Loop operator: searching for the right k for k -means.





Design and Result Perspectives

Organize your space to store processes, files and results.

Point1:

8 points defined by their x-, y-coordinates.



Reading the Point1-Dataset

Warning Read Wizard: RapidMiner guesses the type of attributes looking at the first 100 by default.

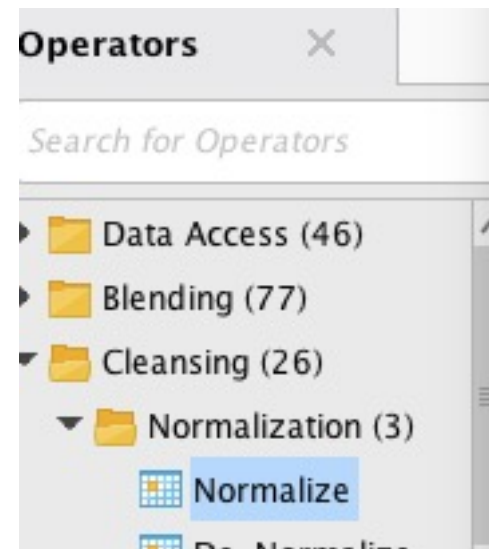
Metadata:

- Check the order of magnitudes of attributes
- Handy for classical transformations such as:

$$\frac{x - \text{average}}{\text{std deviation}}$$

$$\frac{x - \text{min}}{\text{range}}$$

- Transformation can be done inside RM:



Exploring the Point1-Dataset

Plots:

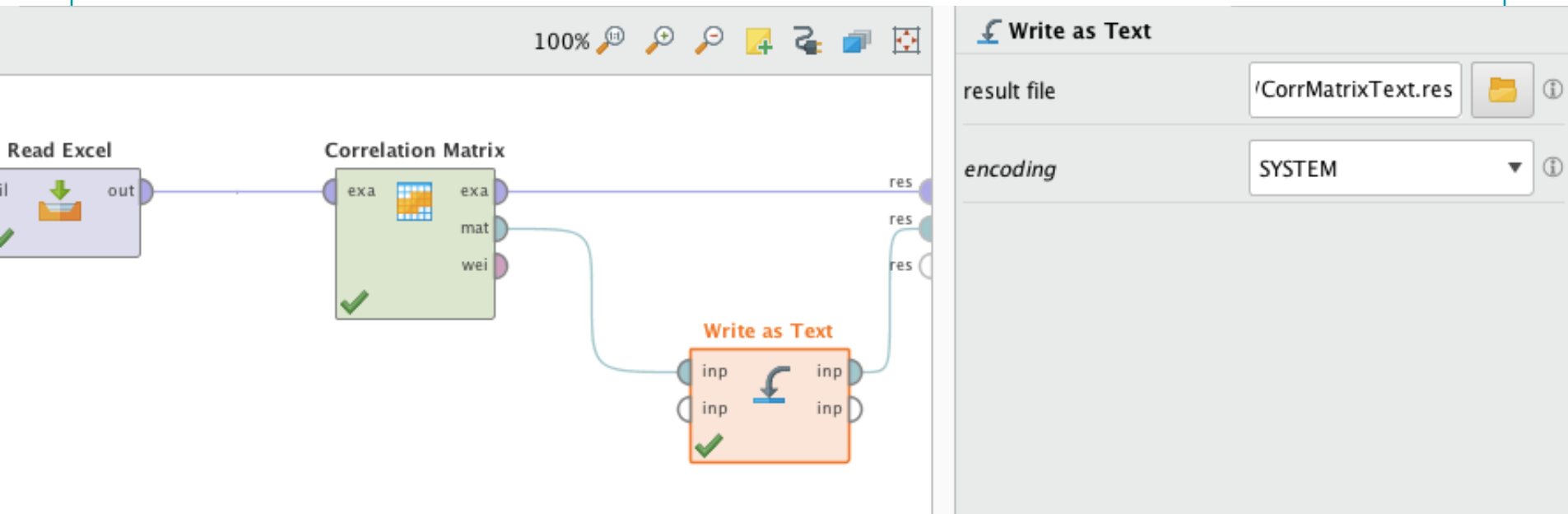
- Histogram / color
- Box plots (quartile)
- Scatter
- and much more...

Correlation Operator:

- **Handy:** store the matrix with the Write as text Operator



Exploring the Point1-Dataset



Looking for groups of points with Clustering

K-means / Scatter Plot with Color Column cluster

The screenshot displays the Orange3 data mining software interface. On the left, a workflow is visible with two nodes: 'Read Excel' and 'Clustering'. The 'Clustering' node is highlighted with an orange border and a green checkmark. The 'Clustering' node's settings panel is open on the right, showing the following configuration:

- Clustering (k-Means)**
- add cluster attribute
- add as label
- remove unlabeled
- k: 3
- max runs: 10
- determine good start values
- measure types: NumericalMeasures
- numerical measure: EuclideanDistance

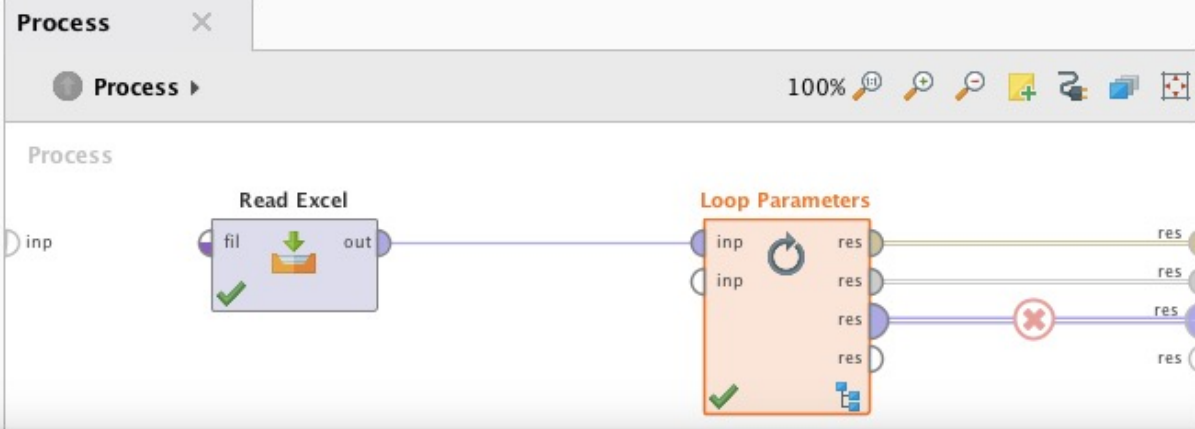
The workflow shows data flowing from 'Read Excel' to 'Clustering'. The 'Clustering' node has three output ports labeled 'res', 'res', and 'res'. A red rectangle is visible in the bottom right corner of the interface.

Looking for groups of points with Clustering

Loop Operator to discover the right k:

- **Warning:** skip the first port per inside the loop operator to get the results
- **Handy:** Log Operator to plot average within distance cluster or Davies Bouldin against k. Edit the parameters you want to plot, k and avg_within_distance for instance.





Parameters

Loop Parameters

Edit Parameter Settings...

error handling fail on error

synchronize

Select Parameters: configure operator

Select Parameters: **configure operator**
Configure this operator by means of a Wizard.

Operators

- Clustering (k-Means)
- Performance (Cluster Distance Performa
- Log (Log)
- Multiply (Multiply)

Parameters

- add_cluster_attribute
- add_as_label
- remove_unlabeled
- max_runs
- determine_good_start_values
- measure_types
- mixed_measure
- nominal_measure

Selected Parameters

- Clustering.k

Grid/Range

Min	Max	Steps	Scale
2	5	6	linear

Value List

- 2
- 3
- 4
- 5

[advanced parameters](#)

[age compatibility \(7.2.002\)](#)

Loop Parameters

pidMiner Studio Core

ate, Settings, Grid, Search, Tune, Optimal, L

ator iterates over its subprocess for all t

parameter combinations. The parameter

tions can be set by the wizard provided

Looking for groups of points with Clustering

Process ×

Process ▶ Loop Parameters ▶ 100%

Loop Parameters

Parameters ×

Log

filename Mining/SeminarTask 📁 ⓘ

log 📄 Edit List (2)... ⓘ

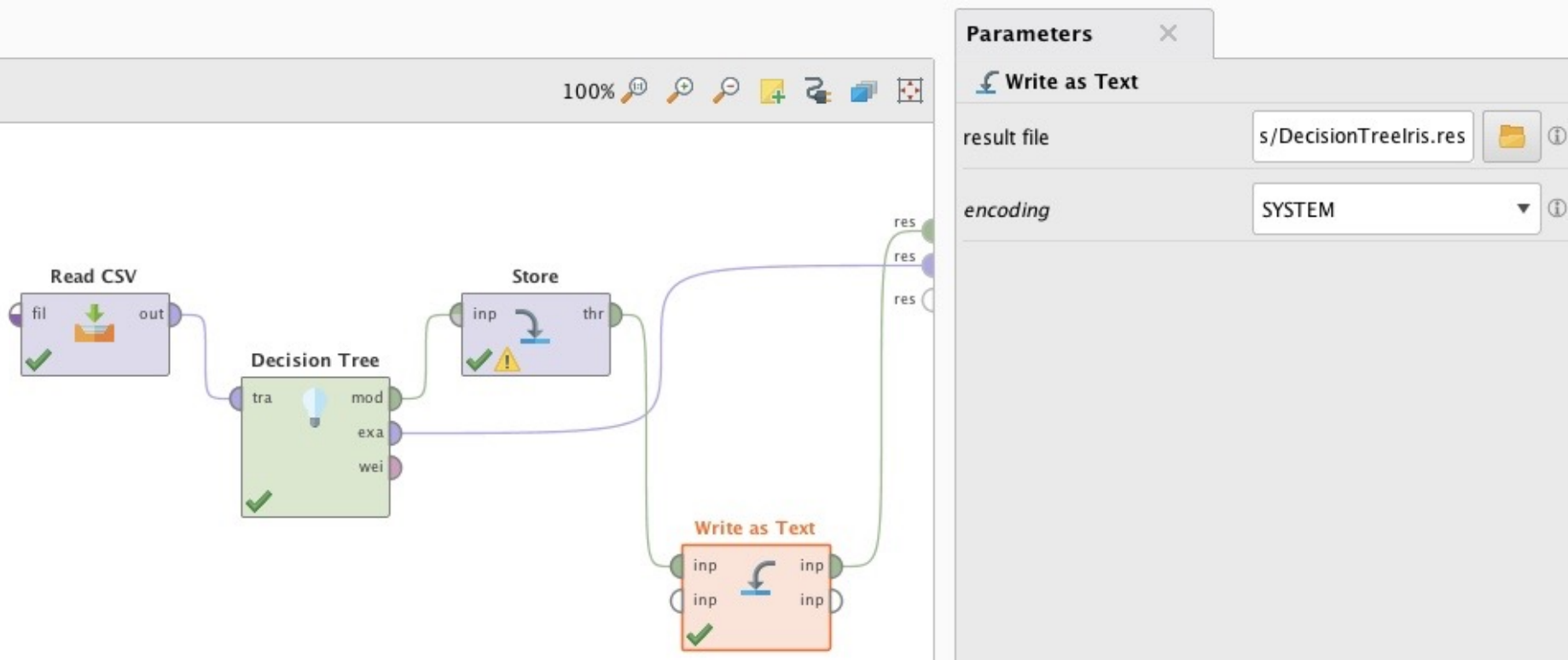
sorting type none ▼ ⓘ

persistent ⓘ



Classification with the iris dataset

Class should be of type Label



Classification

Decision tree:

- **label** to set the attribute to predict

Validation:

- X Validation Operator
- **Handy:** Write as Text Operator the output of Performance Operator to check the result step by step.



Classification

The screenshot displays the Orange3 data mining software interface. On the left, a workflow is visible with two nodes: 'Read CSV' and 'Cross Validation'. The 'Read CSV' node has a green checkmark and a green arrow icon. The 'Cross Validation' node has a yellow background, a percentage sign icon, and a green checkmark. The 'Cross Validation' node is connected to four 'res' output ports. On the right, the 'Parameters' panel for the 'Cross Validation' node is open, showing the following settings:

- split on batch attribute
- leave one out
- number of folds 10
- sampling type stratified sampling
- use local random seed
- enable parallel execution



Classification

The screenshot displays the Orange3 software interface for a classification task. The workflow is divided into two main sections: **Training** and **Testing**.

- Training Section:** A **Decision Tree** widget is used for model training. It has a 'tra' (train) input and 'mod' (model) and 'exa' (examples) outputs.
- Testing Section:** The trained model is used by the **Apply Model** widget, which takes 'mod' and 'exa' as input and produces 'mod' and 'thr' (threshold) outputs.
- Performance Evaluation:** A **Performance** widget calculates various metrics. It receives 'mod' and 'thr' as input and outputs 'lab' (label), 'per' (precision), and 'exa' (examples).
- Output:** A **Write as Text** widget is connected to the Performance widget to save the results.

The **Parameters** panel on the right is titled **Performance (Performance (Classification))**. It shows the following settings:

- main criterion: accuracy
- accuracy
- classification error
- kappa
- weighted mean recall
- weighted mean precision
- spearman rho
- kendall tau



References

Rapid Miner website: rapidminer.com with good tutorials.

- <https://community.rapidminer.com/>
- <https://docs.rapidminer.com/>

Many videos on youtube, see for example:

<https://www.youtube.com/watch?v=C8Ko3-2f-pA&list=PLssWC2d9JhOZLbQNZ80uOxLypglgWqbjA&index=16>



References

Thank your for your attention!



Schloß Charlottenburg, Berlin

