Data Mining I

Optimization
with Rapidminer

Outline

1. Introduction
2. Parameter Optimization
3. Attribute Selection Optimization

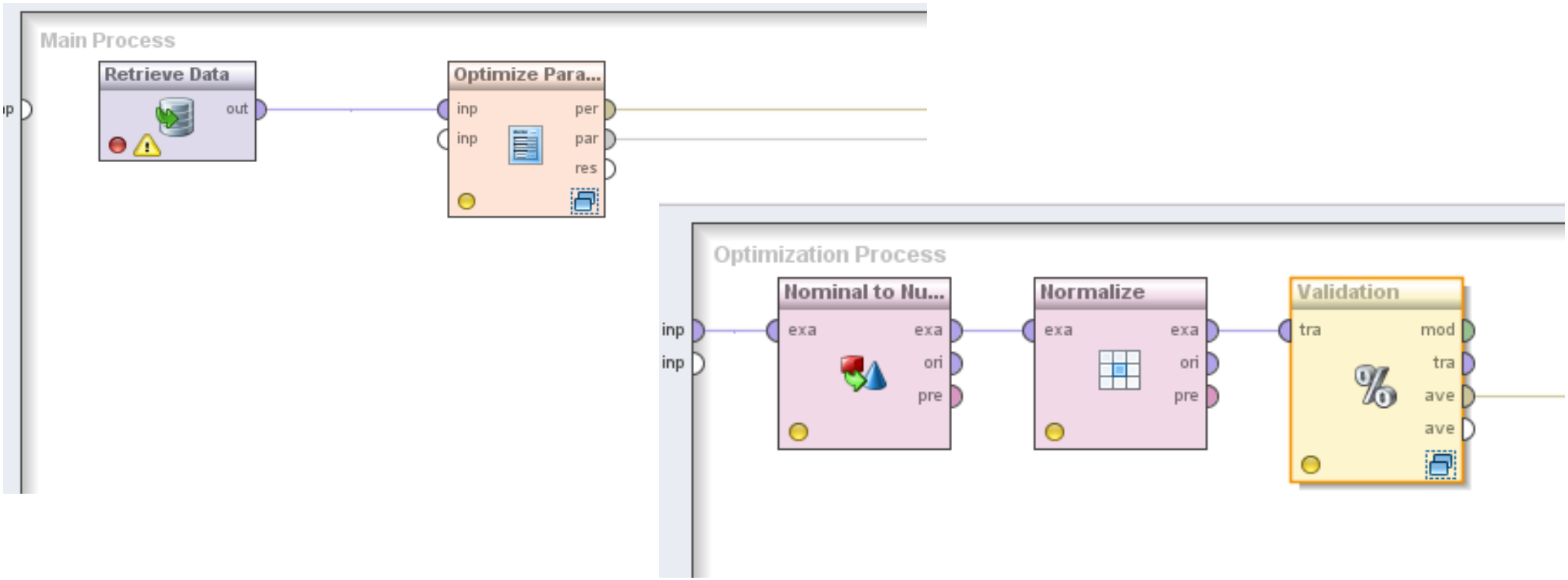
Introduction

- Why should we optimize?
  - Default configuration does not work always best
  - Operators come with a (large) set of different parameters
  - Dataset attributes might be useful or useless
- Why not manual?
  - Testing all combinations of kernel type and svm type (of LibSVM) means rerunning the process 25 times.
  - Playing around with epsilon in addition (5 different values) leads to 125 possible set-ups
  - But, yes you can do it manually, if you are patient
- What can be optimized?
  - Parameters of Operators
  - Selection of Attributes from the Dataset

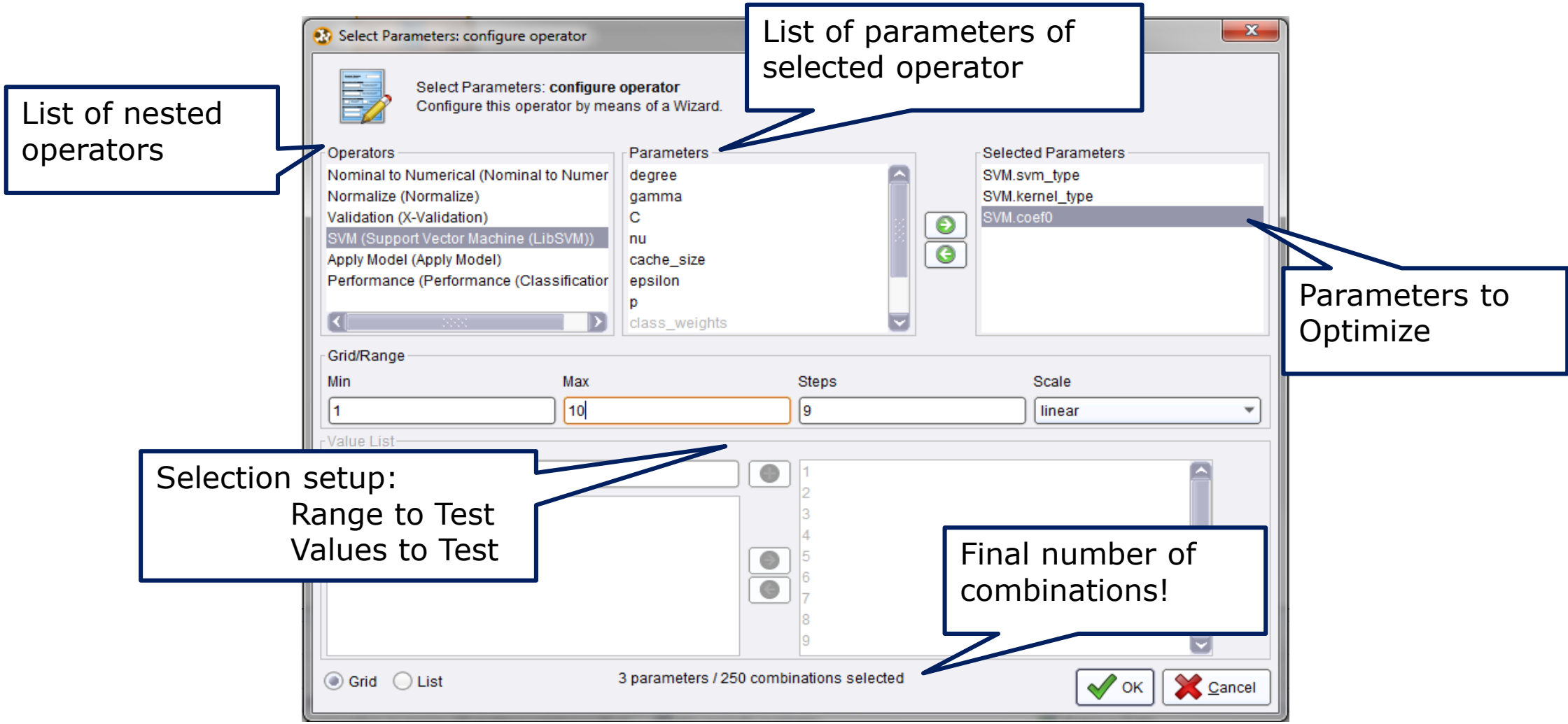
Parameter Optimization

Main Idea:

Let Rapidminer test possible operator parameter combinations for you.



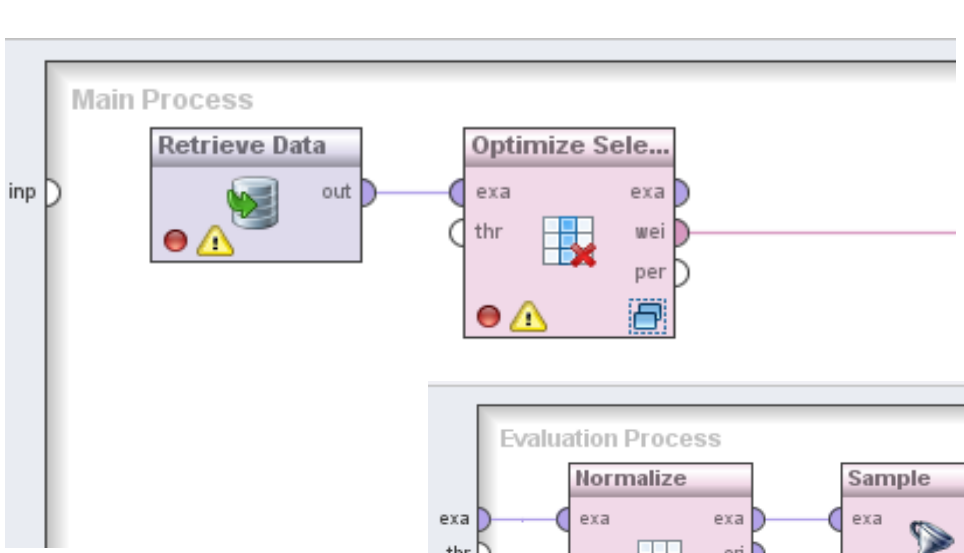
Parameter Optimization

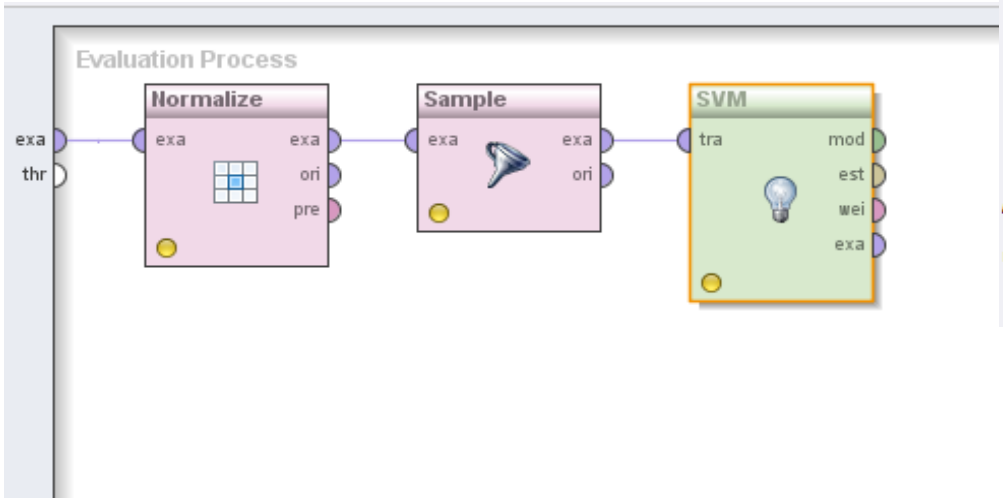


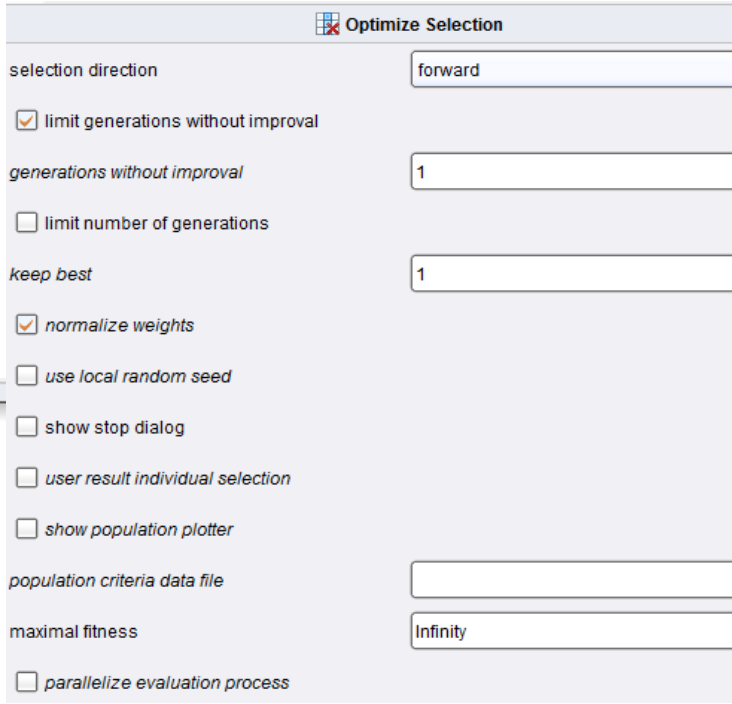
Attribute Selection Optimization

Main Idea:

Let Rapidminer select optimal configuration of attributes for the process and the selected classifier/clustering algorithm

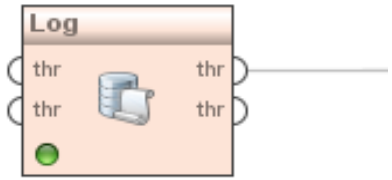


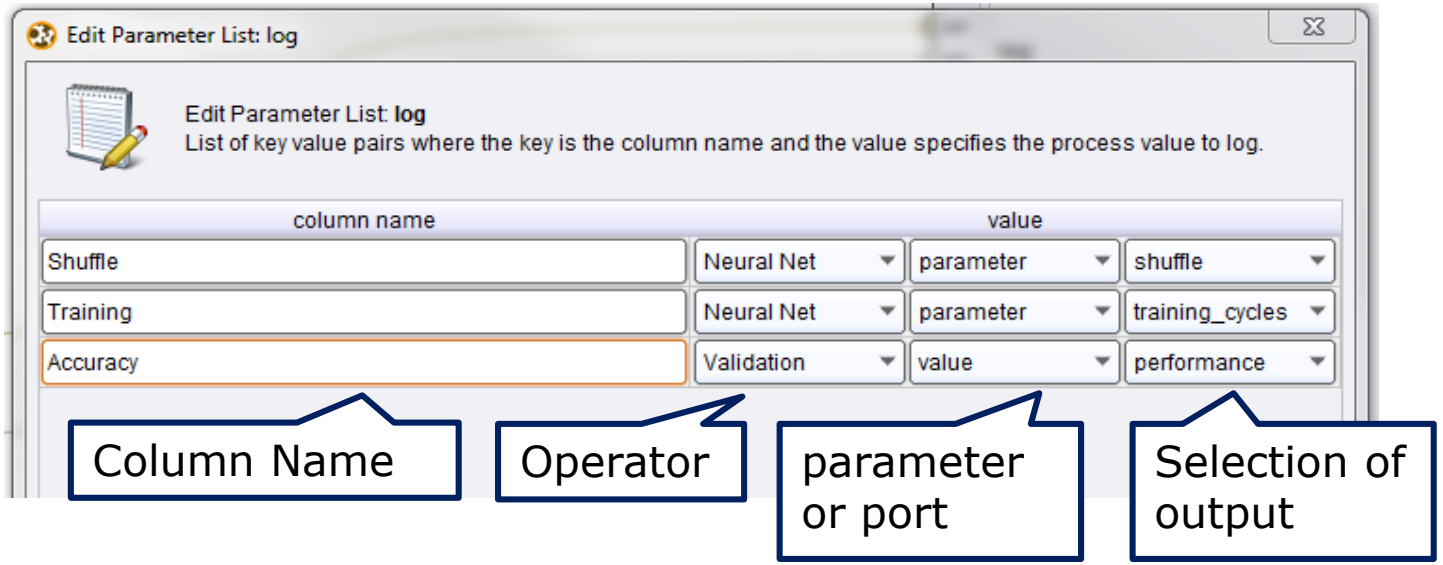


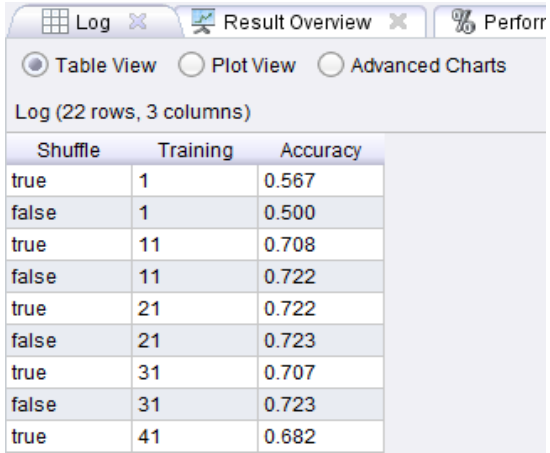


Interim Results

– Using the *Log* Operator







Shuffle	Training	Accuracy
true	1	0.567
false	1	0.500
true	11	0.708
false	11	0.722
true	21	0.722
false	21	0.723
true	31	0.707
false	31	0.723
true	41	0.682

Need more information?

- Parameter Optimization YouTube Video:
  - http://www.youtube.com/watch?v=R5vPrTLMzng
- Attribute Selection Optimization YouTube Video:
  - Part 1: http://www.youtube.com/watch?v=7IC3IQEdWxA
  - Part 2: http://www.youtube.com/watch?v=j5vhwbLlZWg

Questions?