Feature selection with RapidMiner

João Mendes Moreira

LIAAD - INESC TEC, University of Porto, Portugal

Motivation

- **Some algorithms perform worst with irrelevant features**
  - For instance, the k-nearest neighbors classifier and the simple linear classifier perform worst in the presence of irrelevant features.
- **Question: which is the feature subset with best performance?**
- Two different approaches:
  - Filter approaches
  - Wrapper approaches

Agenda

- Filter approaches
- Wrapper approaches
- Exercises

Feature selection with RapidMiner

FILTER APPROACHES

Filter approaches

Features are removed based on their characteristics and independently of the predictive algorithm to be used

Correlated features: a feature is removed when it has a correlation with other feature larger than a given threshold (an input parameter):

- In RapidMiner this is done using the *remove correlated attributes* operator
- See also the *correlation matrix* operator
- An usual threshold is 0.75 using the absolute correlation
- The elimination of correlated features is particularly important for some algorithms such as the Naïve Bayes classifier or multiple linear regression among others

Filter approaches

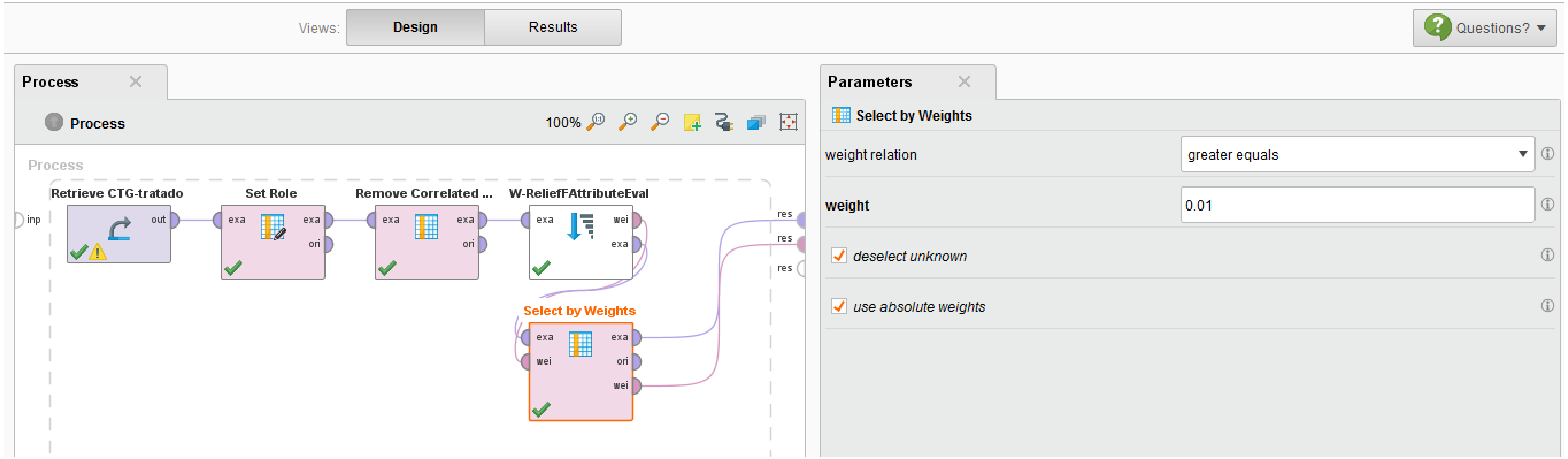
Features are removed based on their characteristics and independently of the predictive algorithm to be used

ReliefF algorithm: it calculates the relevance of each feature:

- The key idea of Relief is to estimate the quality of features according to how well their values distinguish between the instances of the same and different classes that are near each other
- The ReliefF is for classification
- There is a version for regression called RReleifF but it is not available in RapidMiner
- A common threshold to eliminate features is when their ReliefF weights are ≤ 0.01
- This filter is complementary to the correlated features filter, i.e., the correlated features filter should be used first and then, the ReliefF filter

Filter approaches

Features are removed based on their characteristics and independently of the predictive algorithm to be used



Select by Weights

weight relation: greater equals

weight: 0.01

☑ *deselect unknown*

☑ *use absolute weights*

Feature selection with RapidMiner

WRAPPER APPROACHES

Wrapper approaches

Features subsets are tested with a given predictive algorithm. The feature subset with best performance is the selected one.

Exhautive search

- Test all possible feature subsets
- If the dataset has *m* features, the number of possible feature subsets is 2^m*-1*
- For *m>10*, this approach is too expansive computationally

Wrapper approaches

Features subsets are tested with a given predictive algorithm. The feature subset with best performance is the selected one.

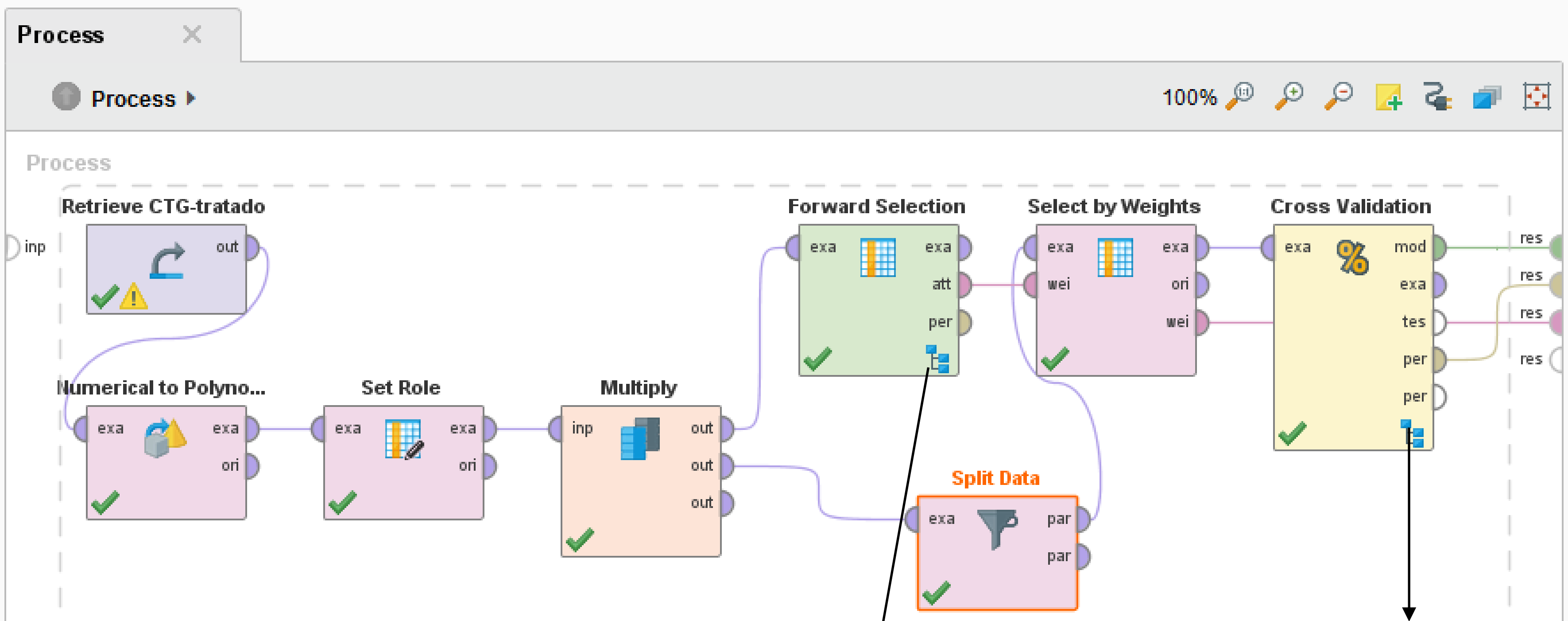
Forward selection

1. It trains a model using separately each of the features
2. It selects the feature whose model is the most accurate
3. It repeats 1 using each of the remaining features plus the selected one
4. If the best model has a better performance than the best model with -1 feature, it adds the feature to the selected ones
5. This process is repeated until the addition of a new feature does not increase performance or when the increase is shorter than a given threshold

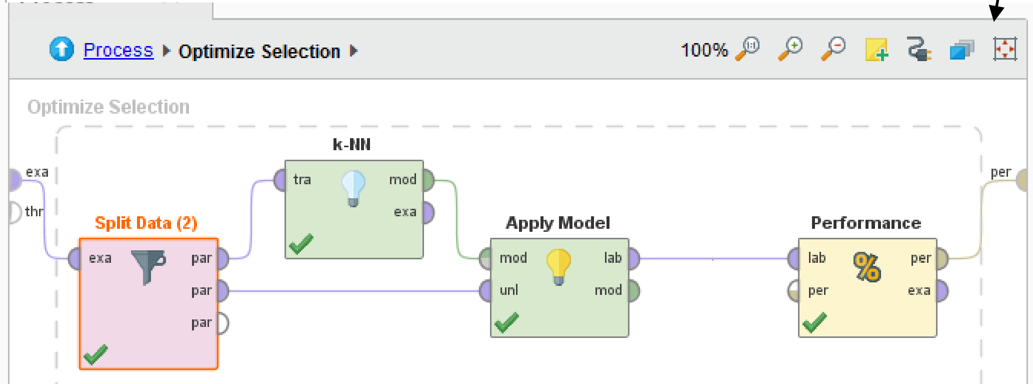
- RapidMiner: the *forward selection* operator

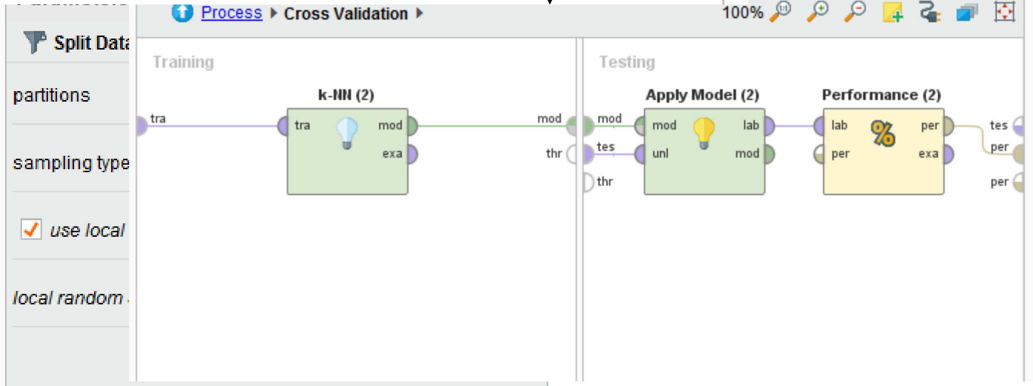
Wrapper approaches

Features subsets are tested with a given predictive algorithm. The feature subset with best performance is the selected one.

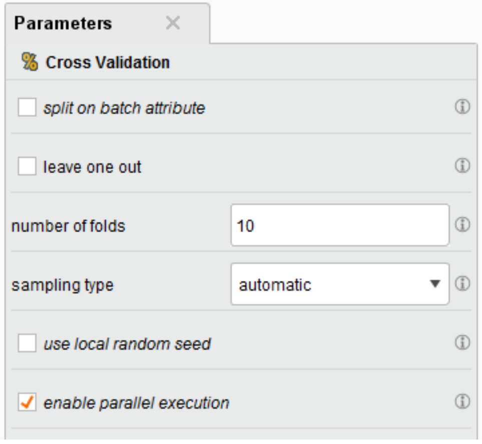


- Both splits should have the same local random seed and should have the first partition of equal size. Try to understand why ...
- The parameter *k* of *k-nn* is 12
- The *select by weight* operator uses as threshold ≥1





☑ use local



Parameters

Cross Validation

□ split on batch attribute

□ leave one out

number of folds: 10

sampling type: automatic

□ use local random seed

☑ enable parallel execution

Wrapper approaches

Features subsets are tested with a given predictive algorithm. The feature subset with best performance is the selected one.

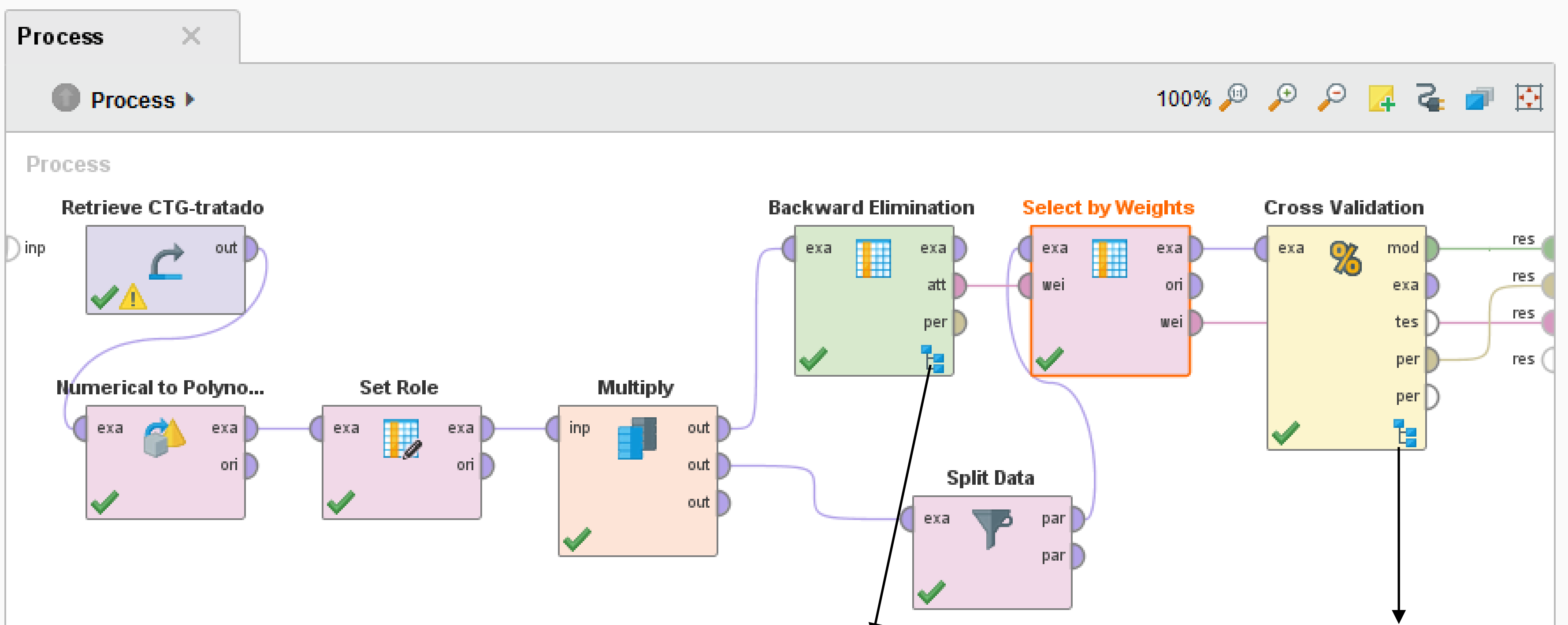
Backward selection

1. It trains a model using all *m* features
2. It trains *m* models by removing a different feature from the initial *m* features, i.e., each model is trained using *m-1* features
3. It selects the feature subset whose model is the most accurate
4. It repeats 2 by removing from the feature subset selected in 3 each feature
5. If the best model generated in 4 has a better performance than the best model using one more feature, it jumps to step 3 and does a new iteration
6. Otherwise it stops

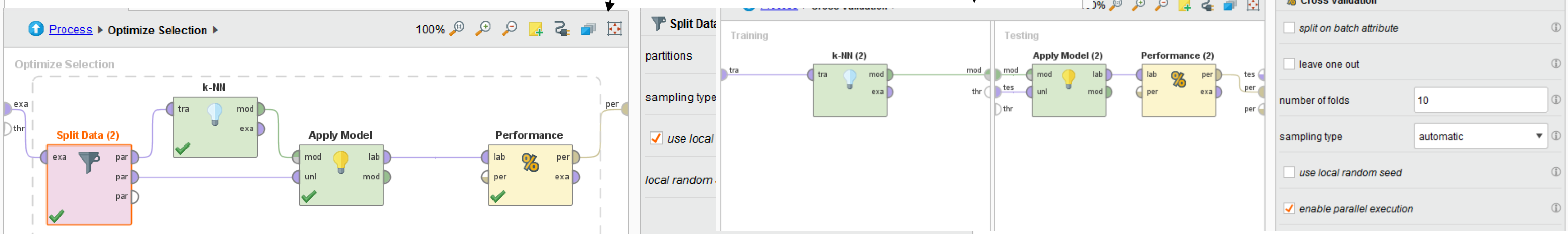
- RapidMiner: the *backward elimination* operator

Wrapper approaches

Features subsets are tested with a given predictive algorithm. The feature subset with best performance is the selected one.



- Both splits should have the same local random seed and should have the first partition of equal size. Try to understand why ...
- The parameter *k* of *k-nn* is 12
- The *select by weight* operator uses as threshold ≥1



partitions

sampling type

☑ use local

local random

□ split on batch attribute

□ leave one out

number of folds: 10

sampling type: automatic

□ use local random seed

☑ enable parallel execution

Wrapper approaches

Features subsets are tested with a given predictive algorithm. The feature subset with best performance is the selected one.

Forward-backward selection

1. Similar to forward selection but allows the addition of a new feature even when the performance does not improve. This is controled by a parameter

2. It can do backward steps

Feature selection with RapidMiner

EXERCISES

Exercises

Context:

- Use the dataset http://archive.ics.uci.edu/ml/datasets/Cardiotocography

- The target variable is *NSP*

- Questions:

1. Load data. The meaningful data is in spreadsheet “Data”. Use only the columns with numbers in the first line.

2. Use the correlation matrix to discard features.

3. Use the ReliefF algorithm to discard features.

4. Find the most promising features subset for k-nn.

5. Find the most promising features subset for ANN.