5.1. Working with SVM

The Sonar data set from the sample area of RapidMiner includes a set of 208 examples for sonar signals bounced of metal cylinder or cylindrical rocks and is used to train classifiers to divide between rock and metal.

Setup a standard X-Validation process to learn a classifier using a Support Vector Machine (LibSVM Operator). What performance do you reach with the operator?

Now try to find a more optimal configuration for the used SVM Operator. Wrap the X-Validation operator into an Optimize Parameter Operator. Than use the “Edit Parameter Settings” option of this operator too let RapidMiner test different combinations of parameters.

Try the following parameters of the SVM:
- SVM_TYPE: C-SVC, nu-SVC
- KERNEL_TYPE: all
- EPSILON: 0.001 to 0.01 (10 steps)

You should come up with 110 (2 x 5 x 11) combinations.

What is the most optimal configuration for the data set and the support vector machine?

In case you need help with the optimization, have a look in the slide set “Optimization with RapidMiner” or task 5.3 of this exercise.

5.2. Open Competition: Finding rich Americans

The Adult data set from the UCI data set library (http://archive.ics.uci.edu/ml/datasets/Adult) describes 48842 persons from the 1994 US Census. The data set is provided as adult.arff file in ILIAS.

Your task is to find a good classifier for determining whether a person earns over 50.000 $ a year. Beside of being accurate, your classifier should also have balanced precision and recall.

To evaluate your classifiers use split validation (split ratio=0.8, linear sampling).

In order to find the best classifier, you may experiment with:
1. different algorithms
2. different parameter settings
3. the balance of the two classes in the data set
4. the set of attributes that are used or not used
5. other preprocessing techniques

People are described by the following 14 attributes:
age: continuous.


fnlwgt: continuous.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num: continuous.


race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands.
In order to increase your understanding of the data set, you might want to visualize different attributes or attribute combinations.

5.3. Have RapidMiner help you!

RapidMiner offers operators to automate the process of

- feature selection (automatically try different attribute combinations in order to find the attribute combination that works best for learning) as well as
- to automatically try out different parameter settings of the learning algorithms in order to find the parameter setting that results in the best performance of the learned model.

The parameter optimization operators are found under “Process Control/Parameter”.

A video tutorial on how to use the *parameter optimization* operators is available at:

http://www.youtube.com/watch?v=R5vPrTLMzng

The feature selection operators are found under “Data Transformation/Attribute Set Reduction/Selection/Optimization”. Start with the *brute force* optimizer first, maybe also try the other ones.

A video tutorial on how to use the *feature selection* operators is available at:

http://www.youtube.com/watch?v=7IC3IQEdWxA  and  
http://www.youtube.com/watch?v=j5vhwbLIZWg

Experiment with the parameter optimization operators as well as the features selection operators.