Exercise 1: Data Mining II – Data Preprocessing

In this exercise we will mainly focus on the data preprocessing. In addition we will repeat some parts of classification within Rapidminer (A short introduction on RapidMiner can be found here). If you are unfamiliar with the basics in one of those topics, please rework through the lectures and exercises of the last semesters Data Mining I course, especially the lecture about pre-processing and classification (1,3,4,5) (http://dws.informatik.uni-mannheim.de/en/teaching/courses-for-master-candidates/archive/hws2014/ie-500-data-mining/).

1) In the following we will work with the Data Mining Cup Data Set of 2010. Please download the dataset (http://www.data-mining-cup.de/en/review/goto/article/dmc-2010.html) and import the training data into your RapidMiner. Please make sure, you set the right attribute type for each of the attributes. (**HINT**: Read the data description before importing)
   What are the numbers of binominal, poly-nominal, date and continues attributes in your dataset?

2) Inspect your data (data visualization). Which attributes highly correlate and by this are potential candidates for removal? What other characteristics can be found in the data set for different attributes? (**HINT**: Keep the last lecture in mind.)

3) Build up a standard classification process using 10-fold cross-validation with a Decision Tree and Naïve Bayes classifier. Inspect the results of the classification process and try to improve your results by including and excluding different attribute sets and balancing the data.

4) Generate new attributes from your date attributes by thinking about useful attributes for the case of the data set. (**HINT**: Ask yourself, e.g. which days/month/times you normally shop). To do so, have a look at *Generate Attribute*, as well as *Date to Numerical*. Do all/some of the newly generated attributes improve your results.

5) In a next step, try to predict values for examples with missing values based on the different options you learned in the lecture (default, average, max, min ...). Does the accuracy/precision/recall of the learned classifier improve?

6) Try out PCA (Principal Component Analysis) within RapidMiner on your new set of attributes and observe the results of your classifier.

7) Finally, remember that you can do everything automatically. RapidMiner offers a set of operators which test the best Attribute Set and the best Parameters for your dataset. Especially when using SVM those are needed. If you cannot remember how to do it you should have a look at the second classification exercise of last semester (see link above).