Exercise 3: Data Mining II – Anomaly Detection

In this exercise we will focus on Anomaly Detection. RapidMiner – out of the box – offers four different operators, which can handle outliers. In addition, we will use the “Anomaly Detection” extension, which is freely available in the Extension section of RapidMiner.

1) Get to know your operators:
   Import the artificial.txt dataset into your RapidMiner. This dataset (2 attributes and one id) was created manually and has three different dense points with a higher data point density. In addition 27 randomly chosen data points were added. Visualize the data and spot the dense points. Now try to adapt one of the Anomaly Detection Algorithms you learned in the lecture to identify these randomly added points. Especially have a closer look at the difference between local and global outlier detection approaches. You can visualize the results using the Bubbles Chart coloring the points based on the outlier decision.

2) Improve Classification
   Apply outlier detection to improve the results of classification. Try to classify the Sonar (RapidMiner Dataset) using X-Validation with Decision Trees. Then apply outlier detection (your choice) to the data and remove outlier from the data. Make sure you are filtering at the right position in your process? As you have learned in the lecture, it might also be a good idea to filter some attributes before trying to detect the outliers. Apply a forward/backward selection for the outlier detection and see if you can improve the results of your classifier.

3) Let us get real:
   Load the NBA dataset (http://databasebasketball.com/stats_download.htm) and try to find the most outstanding players within the dataset (player_regular_season.csv). The data is aggregated already by season and player which makes sense, as we want to find outstanding player for each season. Information about the abbreviations can be found here (http://databasebasketball.com/about/aboutstats.htm). Think about which characteristics of a NBA player are semantically important for this task and filter the corresponding attributes before applying an outlier detection algorithm. Also think about, if you should rather use a global or local algorithm to find a good solution.

4) Semi-supervised Anomaly Detection
   We will now try to semi-automatically find outliers in the shuttle dataset of the UCI which is available as pre-processed download on the website. The original dataset can be found here (http://archive.ics.uci.edu/ml/datasets/Statlog+(Shuttle)). The ZIP contains test and training dataset which needs to be imported into RapidMiner with the labels (anomaly) being nominal (not binominal) attributes types. Within the test set are both (normal and outlier data), where in the training only normal data exists (anomaly). Use a one-class LIBSVM Anomaly Score to train on a sample of 5000 instances of the training
data. Then apply the model to the test data. Make sure the naming of the labels is the same (HINT: Use the mapping operator to map strings to others) and measure the performance of the detection.