

GfKI Data Mining Competition 2005: Predicting Liquidity Crises of Companies Part II: Training and Classification

A. Mauser, D. Keyzers[†], A. Hegerath, T. Deselaers, and I. Bezrukov

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{mauser, keyzers, hegerath, deselaers, bezrukov}@i6.informatik.rwth-aachen.de

Abstract. Data preprocessing and a careful selection of the training and classification method are key steps for building a predictive model with high performance. Here, we present the approach used for training and classification in our solutions submitted to the 2005 GfKI Data Mining Competition. The initial step of data preprocessing is described in the first part of this work.

The task to be solved for the competition was the prediction (binary classification) of a possible liquidity crisis of a company. The prediction was to be based on a set of 26 variables describing attributes of the companies with unknown semantics. The size of the training data set was 20,000 samples, with a class distribution of 10% positive cases, i.e. where an liquidity crisis occurred, and 90% negative cases. The test data consisted of 10,000 unlabeled samples. The models entering the competition were ranked according to the number of true positive cases within the 2,000 samples regarded most likely to enter a liquidity crisis.

From our experience with data mining tasks, we know that it is crucial to avoid overfitting. Furthermore we observed that — given a suitable feature transformation — the actual choice of the classifier has less impact and depends on the task. First, we separated 20% of the training data into a hold-out set to be used for validation of the internal results. Then, a variety of classifiers were examined on the remaining 80% of the training data using five-fold cross-validation. We employed a variety of standard off-the-shelf classifiers as available e.g. in Netlab and Weka (neural networks, nearest-neighbor techniques, decision trees, support vector machines) as well as some in-house classifiers for maximum entropy training and naive Bayes estimation. For each of the classifiers we assessed suitable parameter choices and followed those approaches that gave the best results on the cross-validation data. Finally, we chose a small set of well-performing classifiers, evaluated these on the hold-out set and submitted the predictions of the classifiers with the best performance, now trained on the complete training data. In addition, we also included a combination of three of the classifiers, because often classifier combination can reduce the likeliness of overfitting. The final result showed that the submitted classifier combination (naive Bayes & maximum entropy, alternating decision tree, and logistic regression) and the logistic model tree were both able to correctly detect 894 true positive companies. Only the winning approach that detected 896 companies was better than these results, and the difference is statistically not significant.

Keywords

GfKI Data Mining Competition, classification, predictive modeling

[†] : corresponding author