GfKl Data Mining Competition 2005: Predicting Liquidity Crises of Companies Part I: Data Preprocessing

I. Bezrukov, T. Deselaers†, A. Hegerath, D. Keysers, and A. Mauser

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{bezrukov, deselaers, hegerath, keysers, mauser}@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 preprocessing in our solutions submitted to the 2005 GfKl Data Mining Competition. The subsequent steps of training and classification are described in the second part of this work.

The task to be solved for the competition was the prediction of a possible liquidity crisis of a company. The prediction (binary classification) was to be based on a set of 26 variables describing attributes of the companies. The semantics of the attributes were not disclosed.

Real-world data like the data used here usually suffer from deficiencies that make classification tasks difficult: missing values, outliers, and noisy distributions affect the performance of classification algorithms. Many classifiers perform better if the feature values are adjusted to a common interval or if they are generalized using histograms.

We transformed the data using binary features and two variants of histograms: One histogram type had as many bins as there were different feature values. Each feature value was replaced with the normalized index of its bin. This transformation normalizes the distances between individual feature values, thus also the values of outliers are moved towards the mean value. The other histogram was a 10-bin "equi-depth" histogram. That is, bin borders were adjusted such that each bin contained approximately the same number of elements. Each individual feature value was then replaced by the center of the bin it belonged to. This procedure approximately conserves the original distance proportions.

Missing values were not included in the calculation of the histograms. During the transformation they were replaced by zeros. For features containing a significant amount of unknown values or zeros, we created additional binary features indicating whether the values was missing or not.

The final result showed that our five submissions were within the top ten ranks with two submissions being ranked equally on the second place. As our submissions all use different classifiers (combination of classifiers, logistic model tree, alternating decision tree, maximum entropy & naive Bayes, and neural nets) it can be seen that using appropriate preprocessing techniques it is possible to create an accurate predictive model without knowledge of the content of the data.

Keywords

GfKl Data Mining Competition, data preprocessing, predictive modeling

† : corresponding author