Lessons Learned from Feature Selection Competition

Nitesh V. Chawla NITESH.CHAWLA@CIBC.CA
Grigoris Karakoulas GRIGORIS.KARAKOULAS@CIBC.CA
Danny Roobaert DANNY.ROOBAERT@CIBC.CA
Customer Behavior Analytics
Canadian Imperial Bank of Commerce
Toronto, Ontario M6S 5A6
Canada

Abstract

The purpose of this paper is to provide insight on the performance of different feature selection techniques and learning algorithms that we used on the five datasets of the competition. As part of our participation (CBAgroup) we considered filtering and wrapper feature selection techniques, combined with different learning algorithms. In terms of feature selection techniques, we used information gain, Relief-f, linear SVM together with forward selection and a genetic algorithm. In terms of inductive learning techniques we used a proprietary Bayesian learning algorithm as well as different types of hyperparameter-tuning algorithms for (standard and Bayesian) SVMs with linear and RBF kernel. We provide an evaluation of these techniques on the datasets.

By examining the properties of the data, i.e. feature and class distributions, and the models learned we try to answer: (i) why certain feature selection techniques performed better than others; (ii) why in a couple of datasets SVM performed better when using a selected feature subset than using the entire feature set; (iii) why in the case of Dorothea, the only significantly-imbalanced dataset amongst the five, we improved performance when we applied SMOTE, an ensemble technique for handling imbalanced data.

As is apparent from the above, we make a distinction between feature selection and inductive learning, in the sense that in most real-world applications – e.g. medical diagnosis, mechanical and engineering diagnosis, robotics, marketing, credit scoring, etc. – there is a cost associated with observing the value of a feature. Hence in all those applications the goal should be to come up with a small subset of features that gives the best performance. Therefore, using only classification error as a performance measure, is in practice often not optimal, as the models built with the full feature set may have a lower error, but an overall higher cost. As part of our evaluation we propose a measure that takes into account this trade-off between the number of features and the classification error.