



Institute of Computer Science
Academy of Sciences of the Czech Republic

Trends in random forests parameters for classification of imbalanced data

Marko Robnik-Šikonja¹, Petr Savicky²

Technical report No. 1153

February 2012

Abstract:

The classification with imbalanced class proportions is a particularly difficult problem. We investigate the classification of imbalanced data sets with random forests method, which is one of the state-of-the-art classifiers. To this aim we created a semi-artificial data generation engine which for a supplied real world data set estimates its joint probability distribution with an RBF network and generates new data from this distribution at any proportion of imbalance required. This engine allowed us to systematically and consistently vary level of imbalance for approximations of several real world data sets and to study the parameters of the learning algorithms which influence the classification performance. The results show that consistently across different data sets and imbalance levels, there are notable trends in settings of the stopping criterion and the level of smoothing for an improved performance. These findings are confirmed on large UCI data sets, where imbalance can be observed naturally without modeling.

Keywords:

Machine learning, imbalanced data, classification trees, random forests

¹University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, 1001 Ljubljana, Slovenia, Marko.Robnik@fri.uni-lj.si

²Institute of Computer Science, Academy of Sciences of Czech Republic Pod Vodarenskou Vezi 2, 182 07 Praha 8, Czech Republic, savicky@cs.cas.cz