COMBINING BAGGING, BOOSTING AND RANDOM SUBSPACE ENSEMBLES FOR REGRESSION PROBLEMS

SOTIRIS KOTSIAKTIS AND DIMITRIS KANELLOPOULOS

Educational Software Development Laboratory
Department of Mathematics
University of Patras
Patras 26504, Greece
sotos@math.upatras.gr; dkanellop@teipat.gr

Received February 2011; revised August 2011

ABSTRACT. Bagging, boosting and random subspace methods are well known re-sampling ensemble methods that generate and combine a diversity of learners using the same learning algorithm for the base-regressor. In this work, we built an ensemble of bagging, boosting and random subspace methods ensembles with 8 sub-regressors in each one and then an averaging methodology is used for the final prediction. We performed a comparison with simple bagging, boosting and random subspace methods ensembles with 25 sub-regressors, as well as other well known combining methods, on standard benchmark datasets and the proposed technique had better correlation coefficient in most cases.

Keywords: Machine learning, Data mining, Regression

1. Introduction. Many regression problems involve an investigation of relationships between attributes in heterogeneous databases, where different prediction models can be more appropriate for different regions [5,9]. As a consequence multiple learner systems (an ensemble of regressors) try to exploit the local different behavior of the base learners to improve the correlation coefficient and the reliability of the overall inductive learning system [10].

Three of the most popular ensemble algorithms are bagging [3], boosting [1] and random subspace method [21]. In bagging [3], the training set is randomly sampled \( k \) times with replacement, producing \( k \) training sets with sizes equal to the original training set. Theoretical results show that the expected error of bagging has the same bias component as a single bootstrap replicate, while the variance component is reduced [6]. Boosting, on the other hand, induces the ensemble of learners by adaptively changing the distribution of the training set based on the performance of the previously created regressors. There are two main differences between bagging and boosting. First, boosting changes adaptively the distribution of the data set based on the performance of previously created learners while bagging changes the distribution of the data set stochastically [33]; Second, boosting uses a function of the performance of a learner as a weight for averaging, while bagging utilizes equal weight averaging. On the other hand, in random subspace method [21] the regressor consists of multiple learners constructed systematically by pseudo-randomly selecting subsets of components of the feature vector, that is, learners constructed in randomly chosen subspaces.

Boosting algorithms are considered stronger than bagging and random subspace method on noise-free data; however, bagging and random subspace methods are much more robust than boosting in noisy settings. For this reason, in this work, we built an ensemble combining bagging, boosting and random subspace version of the same learning algorithm using