Adaptation of Error Adjusted Bagging Method for Prediction

Selen Yılmaz Isikhan, Hacettepe University, Ankara, Turkey
Erdem Karabulut, Hacettepe University, Ankara, Turkey
Afshin Samadi, Hacettepe University, Ankara, Turkey
Saadettin Kılıçkap, Hacettepe University, Ankara, Turkey

ABSTRACT

In this study, the error-adjusted bagging technique is adapted to support vector regression (SVR) and regression tree (RT) methods to obtain more accurate predictions, and then the method performances are evaluated with real data sets and a simulation study. For this purpose, the prediction performances of single models, classical bagging models, and error-adjusted bagging models obtained via complementary versions of the above-mentioned methods are constructed. The comparison is mainly based on a real dataset of 295 patients with Hodgkin’s lymphoma (HL). The effect of several parameters such as training set ratio, the number of influential predictors on model performances, is examined with 500 repetitions of simulation data. The results reveal that error-adjusted bagging method provides the best performance compared to both single and classical bagging performances of the methods. Furthermore, the bias variance analysis confirms the success of this technique in reducing both bias and variance.

KEYWORDS

Complementary Neural Network, Error-Adjusted Bagging, Regression Tree, Support Vector Regression

1. INTRODUCTION

In recent years, complementary neural networks (CMTNNs) have been used to solve classification and regression problems. Additionally, a novel error-adjusted averaging technique was initially proposed in a study (Kraipeerapun, Nakkrasae, Fung, & Amornsamankul, 2010) to increase the accuracy of complementary multilayer neural networks (NNs). A study by (Kraipeerapun, Nakkrasae, Amornsamankul, & Fung, 2009) first applied CMTNNs to solve the regression problem and then evaluated the error-adjusted bagging method by applying it to the complementary version of an artificial neural network (ANN). Similar to the ANN method, the support vector regression (SVR) and regression tree (RT) methods are also the preferred data-mining methods, as indicated by extant research in cases where the distribution of the dataset is not known, there is insufficient information with respect to the relationship beten dependent and independent variables, or when the sample size is insufficient (Yu, Liu, Valdez, Gwinn, & Khoury, 2010). Furthermore, some studies evaluated the performance of these methods together with bagging (Friedman, Hastie, & Tibshirani, 2001; Kraipeerapun et al., 2010). Therefore, the present study investigated the use of ANN, SVR, and RT methods to adapt the error-adjusted bagging method to the aforementioned regression methods. The goal of this study was twofold. First, the objective included adapting the average-error-adjusted bagging
method with a complementary version of the SVR and RT methods. Second, the study objective included comparing single performances in addition to the performances of the classical bagging method and error-adjusted bagging method for the three prediction techniques. The performance evaluations of the average-error-adjusted bagging method were implemented on a real oncological dataset and two other synthetic datasets. First, related algorithms were applied to real-world regression problem to predict the extent of lymph node metastasis. Extant studies on lymph node metastases typically examined classification or logistic regression methods (Joo et al., 2012). A previous study included an example of linear regression in which the rate of central nodal metastases was predicted by factors such as age, gender, tumor size, and multifocality (Joo et al., 2012). However, in the present study, since the model between nine clinical characteristics and the number of lymph nodes involved was not known, strong prediction methods of data mining, such as the SVR, ANN, and RT methods were preferred in this study. The adjusted bagging technique based on Boston housing and prostate cancer datasets from the UCI (Center for Machine Learning and Intelligent Systems) machine learning database was also tested. Second, a simulation study was conducted to explore method comparisons related to a few important parameter changes.

Section 2 presents the main concepts of CMTNN, SVM and decision tree for the regression problem, discusses the modeling process and parameter settings, and describes an adjusted average method to increase the accuracy of the classical bagging performances of the methods. Section 2 also describes the real HL patients and UCI datasets used in the experiments. The results of our experiment are presented in section 3, and discussed in section 4. Conclusions and details regarding future work are presented in Section 5.

2. MATERIALS AND METHODS

In this section, we briefly describe the data mining methods we use for our problem.

2.1. Multilayer Feedforward Neural Network

An ANN is a non-parametric statistical model to extract nonlinear relations in data. ANNs have been successfully used in several complex and diverse tasks in clinical medicine, e.g., clinical outcome predictions of head injury or bone mineral density (Rughani et al., 2010; Shi, Wang, & Wang, 2013). Instability is a drawback of ANNs. Generally, ANNs tend to possess a low bias and high variance; thus, they may benefit from aggregation methods such as bagging. Breiman (1996) indicated that combining the outputs of a number of models reduces variance and results in more precise predictions for unstable predictors (Shafiee & Jazayeri-Rad, 2012). In practice, bagging has been applied to a large number of model structures, including RTs (Büchmann & Yu, 2002), regression with variable selection (Triadaphillou et al., 2007) and NNs (Chen & Ren, 2009).

In ANN’s structure, the input, hidden, and output variables correspond to three layers of nodes (Figure 1). Each input node corresponds to a predictor variable, whereas an output node corresponds to a response variable. Each node receives multiple inputs from other nodes in proportion to their connection weights and generates a single output (Mashaly & Alazba, 2016). We employed multilayer perceptron network (MLPNNs) which is the most popular type of NN and termed it as ANNs for application (Ha, Cho, & MacLachlan, 2005; Menke, 2012).

A cost function is given by the following expression:

\[ J = (y - y_{obs})^2 \]  

It measures the mean square error between the model output y and the observed values \( y_{obs} \).
Contextualized Text OLAP Based on Information Retrieval
[www.igi-global.com/article/contextualized-text-olap-based-on-information-retrieval/125648?camid=4v1a](www.igi-global.com/article/contextualized-text-olap-based-on-information-retrieval/125648?camid=4v1a)

Construct a Bipartite Signed Network in YouTube
[www.igi-global.com/chapter/construct-a-bipartite-signed-network-in-youtube/150175?camid=4v1a](www.igi-global.com/chapter/construct-a-bipartite-signed-network-in-youtube/150175?camid=4v1a)