Data & Knowledge Engineering 87 (2013) 41–65

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Editorial

Modifications of the construction and voting mechanisms of the Random Forests Algorithm

Evanthia E. Tripoliti [a], Dimitrios I. Fotiadis [a,*], George Manis [b]

[a] Unit of Medical Technology and Intelligent Information Systems, Department of Materials, Science and Engineering, University of Ioannina, GR 45110, Greece
[b] Department of Computer Science, University of Ioannina, GR 45110, Greece

ARTICLE INFO

Article history:
Received 19 May 2011
Received in revised form 12 July 2013
Accepted 12 July 2013
Available online 6 August 2013

Keywords:
Classification
Random Forests
Ensemble methods
Weighted voting
Decision tree

ABSTRACT

The aim of this work is to propose modifications of the Random Forests algorithm which improve its prediction performance. The suggested modifications intend to increase the strength and decrease the correlation of individual trees of the forest and to improve the function which determines how the outputs of the base classifiers are combined. This is achieved by modifying the node splitting and the voting procedure. Different approaches concerning the number of the predictors and the evaluation measure which determines the impurity of the node are examined. Regarding the voting procedure, modifications based on feature selection, clustering, nearest neighbors and optimization techniques are proposed. The novel feature of the current work is that it proposes modifications, not only for the improvement of the construction or the voting mechanisms but also, for the first time, it examines the overall improvement of the Random Forests algorithm (a combination of construction and voting). We evaluate the proposed modifications using 24 datasets. The evaluation demonstrates that the proposed modifications have positive effect on the performance of the Random Forests algorithm and they provide comparable, and, in most cases, better results than the existing approaches.

1. Introduction

An active area of research in machine learning is the combination of classifiers, the commonly called "ensemble methods". Ensemble methods are learning algorithms that construct a set of classifiers and then classify unknown data by taking a (weighted) vote of their predictions [1]. Several authors have demonstrated the advantages of the ensemble methods over the individual classifier models and have noted that they can significantly improve the performance of learning [2,3]. A necessary and sufficient condition for an ensemble classifier to be more accurate than any of its individual base classifiers is the individual classifiers to be accurate and diverse.

Several methods have been developed concerning the construction of an ensemble classifier [1]. These methods are grouped into those that cope with: a) training examples to generate multiple hypotheses, b) the set of input features available to the learning algorithm, c) the output targets that are given to the learning algorithm, d) those that inject randomness into the learning algorithm and e) those that use Bayesian voting. The most popular variants of ensemble methods are Bagging [4], Boosting [5], Random Subspace methods [6] and, the last decade, Random Forests [7].

In Bagging [4] each base classifier is trained using a set generated by randomly drawing with replacement of N examples, where N is the number of instances in the original training set. The combination of base classifiers is made by majority voting. Boosting [5] produces a series of classifiers and the training set used for each classifier is chosen based on the performance of the earlier classifier(s) in the series. The classifiers are added one at a time and are trained on the data which have been "hard" for the previous ensemble members. This is achieved by assigning a weight to each training example and may adaptively change the weight at the end

* Corresponding author at: Unit of Medical Technology & Intelligent Information Systems, Department of Materials, Science and Engineering, University of Ioannina, Stavros Niarchos Avenue, GR 45110, Ioannina, Greece. Tel.: +30 2651008803; fax: +30 2651008889.
E-mail address: fotiadis@cc.uoi.gr (D.I. Fotiadis).

0169-023X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.datak.2013.07.002

of each boosting round. In the Random Subspace methods the training data are also modified. However, this modification is performed in the feature space. More specifically, each base classifier is built on a different subset of features randomly chosen from the original feature set. The outputs of the models are then combined, usually by a simple majority voting process.

Random Forests algorithm [7] constructs a set of tree-based learners. Each base learner is grown on a bootstrap sample of the dataset. As the tree is constructed, a random sample of predictors is drawn before each node is splitted. The number of the selected predictors remains constant throughout the construction of the forest. The split of the node is based on the best of the randomly chosen predictors. This procedure is repeated for each node of the tree which is fully grown and not pruned. Then, each tree of the forest casts a vote for the instance being classified and the predicted class is determined by a majority voting procedure.

Random Forests (*RF*) presents a variety of advantages over other ensemble methods. It provides estimates about the importance of the input variables and it detects the interactions between them. In Random Forests, there is no need for a separate test set to get an unbiased estimate of the generalization error. It is estimated internally through the use of the out-of-bag instances (one third of the training instances not taking part in the construction of the tree). It incorporates methods for handling missing values and it can balance the error in class population of unbalanced datasets. Finally, Random Forests is less sensitive to noise data compared to other ensemble methods and according to Breiman [7] it does not overfit. The first two advantages are attributed to the fact the algorithms has its root to Bagging [2].

Although, the mechanisms that explain the high performance of the Random Forests have been detected, they are not fully exploited to utilize the potential of this method. A possible extension of Random Forests concerns the increase of the strength, the decrease of the correlation or the improvement of the combination of tree-based classifiers.

For this purpose several studies have been reported in the literature each one addressing, separately, one of the above issues. More specifically, Robnik-Sikonja [8], Bernard et al. [9], Rodriguez et al. [10], Lemmond et al. [11] focused on the construction of the base classifiers either affecting the number of features selected at each node or the evaluation measures which determine the best split of the node. On the other hand, Robnik-Sikonja [8], Tsymbal et al. [12], Hu et al. [13] and Gunter et al. [14] focused on finding the best combination function of base classifiers, since each one of the base classifiers has a different impact on the processing of different instances.

Bernard et al. [9] (*RK − RF*) focused on the setting of the hyperparameter *m* (a parameter that is not automatically tuned by the algorithm and in *RF* expresses the number of features which are used to determine the decision at a node of the tree [7]). Instead of fixing the value of *m*, in order to be identical for all the decision trees, a new value of *m* is randomly chosen for each node of the trees, and used only for the splitting of this node. Rodriguez et al. [10] proposed a modification which is based on the utilization of a linear combination of features in each splitting node (*Rotation Forest*). A similar approach is the employment of Linear Discriminant Analysis (*LDA*) to create a linear combination of features [11]. Robnik-Sikonja [8] proposed the replacement of the *Gini* index by *ReliefF* (*RF with ReliefF*). *ReliefF* evaluates partitioning power of attributes according to how well their values distinguish similar instances. However, the results indicated that *ReliefF*, on average, increased the correlation between the trees and resulted in decreased performance. Thus, moving one step forward, Robnik-Sikonja [8] replaced the *Gini* index as the sole attribute evaluation measure with several others, decreasing in this way the correlation but retaining the strength of the tree classifiers (RF with multiple estimators — *RF with me*).

Robnik-Sikonja [8] uses internal estimates to identify the instances that are most similar to the one we wish to classify and then weights the votes of the trees with the strength they demonstrate on these near instances (*RF with wv-1*). The similarity measure used is given by the ratio of the number of times the unknown instance and the training instance are in the same terminal node over the number of trees, while the similarity with nearest neighbors is expressed through different distance measures in *RF with wv-2* and *RF with wv-3*. Another approach based on the performance of the base classifiers is the one proposed by Hu et al. [13]. It is based on the maximally diversified multiple decision tree algorithm (*RF with wv-4*). Tsymbal et al. [12] modifies the combination of trees by taking into account their local predictive performance (*RF with wv-5*). One such technique is the dynamic integration where the local predictive performance is first estimated for each base model based on the performance of similar instances, and then it is used to calculate the corresponding weight for combining predictions with Dynamic Selection (*DS*), Dynamic Voting (*DV*) or Dynamic Voting with Selection (*DVS*). Finally, a completely different approach is followed on the sixth weighted voting scheme (*RF with wv-6*). The weights of the trees are determined using genetic algorithms [14]. The fitness function is defined as the recognition rate of the forest when weighted voting is used.

The contribution of this work is twofold. First, it proposes modifications of the Random Forests algorithm which aim to improve either the construction of the forest or the voting mechanism as those reported previously and second, it proposes modifications that address both factors which affect the performance of the algorithm (construction and voting). Regarding the construction of the forest, a combination of a variable number of features, selected at each node, and different evaluation measures, utilized at each base classifier, are proposed for the first time. The voting mechanism consists of novel procedures that either select the best subset of base classifiers, by employing feature selection techniques, or assign weights to the votes of the trees, by adopting nearest neighbors, dynamic integration and optimization techniques. Finally, modifications of the Random Forests algorithm addressing the number of the predictors at each node, the evaluation measure that determines the best split and the voting scheme are proposed for the first time. More specifically, they address the three aforementioned factors jointly or their combination in pairs. The proposed modifications provide a strong ensemble classifier (having small generalization error) since they construct accurate, non-correlated base classifiers which are combined in an optimal way in order to classify an unknown instance. The resulted forest is characterized by robustness since the modifications incorporate mechanisms that address the weaknesses related to either the procedure of node splitting or the integration of the trees or both. The robustness is also ensured by the existence of procedures which determine the optimal configuration (parameterization) of the forest.