New Machine Learning Algorithm: Random Forest

Yanli Liu, Yourong Wang, and Jian Zhang

Basic Teaching Department, Tangshan College, Tangshan Hebei 063000, China
lyl7937@126.com, yourong1214@163.com, zhjian8765@yahoo.com.cn

Abstract. This Paper gives an introduction of Random Forest. Random Forest is a new Machine Learning Algorithm and a new combination Algorithm. Random Forest is a combination of a series of tree structure classifiers. Random Forest has many good characters. Random Forest has been wildly used in classification and prediction, and used in regression too. Compared with the traditional algorithms Random Forest has many good virtues. Therefore the scope of application of Random Forest is very extensive.

Keywords: random forest, accuracy, generalization error, classifier, regression.

1 Introduction

The traditional machine learning algorithms usually give low classifier accuracy, and easy got over-fitting .To improve the accuracy, many people research on the algorithm of combining classifiers. Many scholar start the research on improve the classification accuracy by means of combining classifiers. In 1996, Leo Breiman advanced Bagging algorithm which is one of the early stage algorithm [1]. Amit and Geman define a large number of geometric features and search over a random selection on these for the best split at each note[2]. In 1998, Dietterich put forward the random split selection theory[3]. At each node the split is randomly selected from the N best splits. Ho[4] has done much study on "the random subspace" method which grows each tree by a random selection of a subset of features. Breiman [5]generate new training sets by randomizing the outputs in the original training set. Among these, the idea, in Amit and Geman's paper, influenced Breiman's thinking about random forests.

Random forests are a combination machine learning algorithm. Which are combined with a series of tree classifiers, each tree cast a unit vote for the most popular class, then combining these results get the final sort result. RF posses high classification accuracy, tolerate outliers and noise well and never got overfitting. RF has been one of the most popular research methods in data mining area and information to the biological field. In China there are little study on RF, so it is necessary to systemic summarize the down to date theory and application about RF.

2 The Principle of Operation and Characters of Random Forest

2.1 Principle of Operation

2001, Leo Breiman definite random forests as:

Definition 2.1 A random forest is a classifier consisting of a collection of tree–structured classifiers $\{h(\mathrm{x},\Theta_k), k = 1,...\}$ where the $\{\Theta_k\}$ are independent

B. Liu, M. Ma, and J. Chang (Eds.): ICICA 2012, LNCS 7473, pp. 246–252, 2012.

identically distributed random vectors and each tree casts a unit vote for the most popular class at input x.

This definition show RF is a combination of many tree-structure classifiers. In Breiman's RF model, every tree is planted on the basis of a training sample set and a random variable, the random variable corresponding to the kth tree is denoted as Θ_k, between any two of these random variables are independent and identically distributed, resulting in a classifier $h(\mathrm{x},\Theta_k)$ where x is the input vector. After k times running ,we obtain classifiers sequence $\{h_1(\mathrm{x}),h_2(\mathrm{x}),\cdots h_k(\mathrm{x})\}$,and use these to constitute more than one classification model system ,the final result of this system is drown by ordinary majority vote, the decision function is

$$H(x)=\arg\max_{Y}\sum_{i=1}^{k}I(h_i(x)=Y) \tag{1}$$

where $H(x)$ is combination of classification model, h_i is a single decision tree model, Y is the output variable, $I(\cdot)$ is the indicator function. For a given input variable, each tree has right to vote to select the best classification result. Specific process shown in Fig. 1.

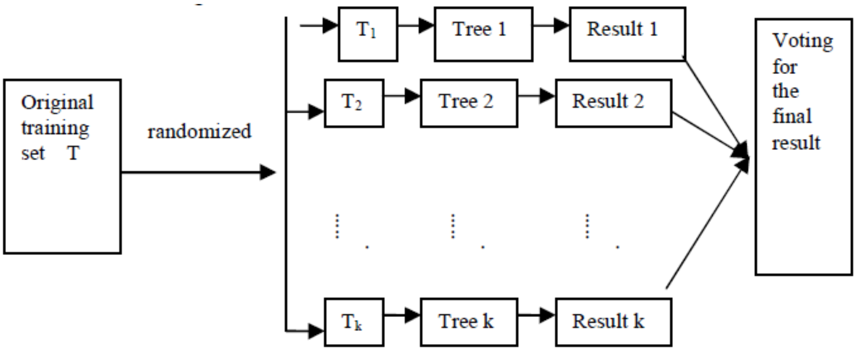

Fig. 1. Random forest schematic

2.2 Characters of Random Forest

In Random Forest, margin function is used to measure the extent to which the average number of votes at **X**,Y for the right class exceeds that for the wrong class, define the margin function as:

$$mg(X,Y)=av_kI(h_k(X)=Y)-\max_{j\neq Y}av_kI(h_k(X)=j) \tag{2}$$

The larger the margin value, the higher accuracy of the classification prediction, and the more confidence in classification.