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.

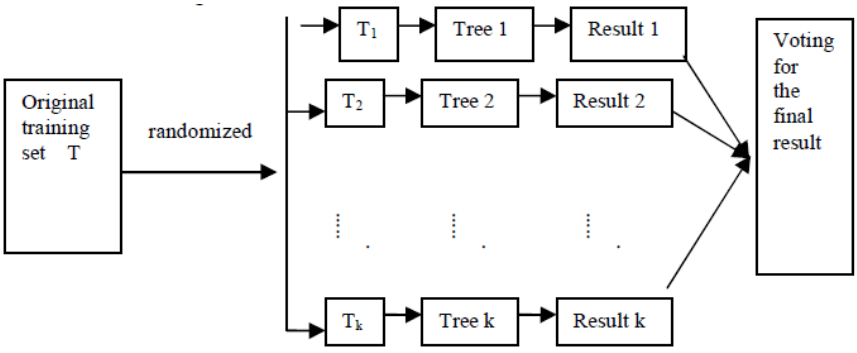

Fig. 1. Random forest schematic

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:

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

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

Define the generalization error of this classifier as:

$$PE^* = P_{X,Y}(mg(X,Y) < 0) \tag{3}$$

when the number of decision tree is big enough, $h_k(X) = h(X, \Theta_k)$ obey the Strong Law of Large Number. Leo Breiman has proved two conclusions. One is RF do not over-fitting but really produce a limiting value of the generalization error. The reason is with the number of the decision trees increases, for almost surely all sequences $\Theta_1, \cdots$ PE* converges to

$$P_{X,Y}(P_\theta(h_k(X,\theta) = Y) - \max_{j \neq Y} P_\theta(h(X,\theta) = j) < 0) \tag{4}$$

Another is the upper bound of the generalization error is exist, and

$$PE^* \leq \bar{\rho}(1 - s^2) / s^2 \tag{5}$$

s is the strength of the set of classifiers $\{h(\mathbf{x}, \theta)\}$, $\bar{\rho}$ is the mean value of the correlation [6]. It shows that the generalization error of RF depends on two aspects: one is the strength of the individual trees in the forest, another is the correlation between these trees. Obviously, the smaller this value, the better the results of random forest.

2.3 Out-of-Bag Estimation

In the process of constructing RF, the tree is planted on the new training set by using random features selection, the new training set is drawn from the original training set by bagging methods. There are two reasons for using bagging. The first is that the use of bagging seems to enhance accuracy when random features are used. The second is using bagging will bring out-of bag data, which can be used to give ongoing estimates of the PE*of RF, as well as estimates for the strength and correlation.

Given an original training set T with N samples, the kth training set is drawn from T with replacement by bagging, every T_k contains N samples. Then the probability of each sample can not be contain is $(1-1/N)^N$, when N large enough, $(1-1/N)^N$ is converges to e^{-1}. In other words, 36.8% samples of the T is not contained in T_k. This samples is called out-of –bag data. The algorithm of using these data to estimate the performances of classification is called OOB estimation. For each tree, there is an OOB estimate for its error. The estimation of generalization error of RF is the average of estimations of all tree error for every tree contained in the RF. Compared with cross-validation the OOB estimate is unbiased and runs faster. The accuracy of OOB estimate is favorable to cross-validation. Tibshirani, Wolpert and Macready Proposed using OOB estimate as an ingredient in estimates of generalization error [7-8]. Breiman has proofed the out-of-bag estimate is as accurate as using a test set of the same size as the training set .Therefore, using the OOB error estimate removes the need for a test set aside[9].Strength and correlation can also be estimated using out-of -bag methods. This gives an internal estimate what is helpful in understanding classification accuracy and how to improve it.

3 The Methods of Random Forests Construction

There are many methods to construct RF, for example bagging method, using random input selection, the effects of output noise, etc.

3.1 Using Input Variables to Construct R.F.

There are three methods to construct R.F.by using input variables, Forests-RI,Forests-RC and Categorical Variables. Because the mechanism of Categorical Variables is complex and the strength is not much better than other RF, so we won't explain this method here, and only introduce the other two methods.

Forest-RI is the simplest RF with random features. Forest-RI is formed by randomly selecting a small group of input variables at each node to split on. F the size of the group is fixed. Using CART methodology to plant tree, maximum size and do not prune. In Breiman's experiment two values of F were tried. One is F=1, another is the first integer less than $\log_2^{M+1}$, M is the number of inputs .The accuracy of Forest-RI is favorable with Adaboost . Forest-RI can be much faster than both Adaboost and Bagging. And the procedure is not overly sensitive to the value of F. It is surprising that when F=1, the procedure has good accuracy too.

When there are a few inputs, M is not big, taking F inputs from all as random selection might lead an increase in strength but higher correlation too. Defining new feature by random linear combination of specifying L input variables. Then there are much enough features. At each given node, L variables are randomly selected and added together with coefficients $k_i, v = \sum_{i=1}^{L} k_i v_i$. $k_i \in [-1,1]$ .F linear combinations are generated, and the best split can be found over these. We call this procedure Forest-RC. Breiman's study show Forest-RC has merits: 1) Forest-RC can deal with data set contain incommensurable input variables; 2) On the synthetic data sets Forest-RC does exceptionally well; 3) Compared with Forest-RI, the accuracy of Forest-RC is more favorably to Adaboost.

3.2 Using Output Construct Random Forest

There are two methods to construct RF using output. One is output smerring, putting Gauss noise in the procedure of output. Another is output flipping, changing one or several classifying labels of the output [10]. In this procedure, the variable remained relatively the same in the classification section is very important. The most obvious virtue of this idea is the RF process the ability of estimating the importance of each feature. The RF constructed by this method can be used to regression well as classification, and better than Bagging in strength. But output flipping depend on the selection of flip rate.

Using updated the weight to built RF [11-12], the merits of this idea is easy and run faster, and easy to realize by program. But this method obviously relay to the data itself and week learning, and can be easy influenced by noises. SRF is built by

randomly selected feature subspace [13-14]. In the given sample space, using this idea you can built as many tree as you want. The strength is much better than tree. With the complexity of the construct, the overall accuracy is almost monotonically increasing. SRF accuracy for multi-tree is optimal.

4 Random Forests for Regression

Random forest can be used to regression too. Specific construction method of RF regression model can be found in [15]. RF regression model can be briefly summarized as: given sample space $\mathbf{x}$ and classification labels Y ,random forests for regression are formed by planting trees depending on the random variable Θ , relative to each category label, tree predictor $h(x,\Theta)$ can give a numerical result. The random forest predictor is formed by taking the average over k of the trees $h(x,\Theta_k)$ .Similarly to the classification case, the following holds:

Lemma 4.1 As the number of trees in the forest goes to infinity, almost surely,

$$E_{X,Y}(Y - av_k h(X,\theta_k))^2 \to E_{X,Y}(Y - E_\theta(X,\theta_k))^2 \tag{6}$$

Random forests regression function is $Y = E_\theta(X,\theta_k)$. In practice, when k big enough $Y = av_k h(X,\theta_k)$ is usually used to instead of the regression function. Breiman has proofed the conclusion that assume for all Θ , $E_Y = E_X h(X,\theta)$ then

$$PE^*(forest) \le \bar{\rho} PE^*(tree) \tag{7}$$

This pinpoints that low correlation between residuals and low error trees can give high accurate regression forest. To test effect of the RF regression, compare this regression with SVR [16] and linear regression. Do regression on the data set CPU.arff (Weka's data set) , the resulting parameters are shown in Table 1.

Table 1. The results of three regression models

Regression model name Parameter	RFR	SVR	Linear R
Correlation coefficient	0.9613	0.9398	0.9544
Mean absolute error	13.0878	19.7969	32.1855
Root mean squared error	50.3600	62.0144	46.0993
Relative absolute error	14.9775%	22.6554%	36.8327%
Root relative squared error	32.6195%	40.1683%	29.8597%

The results show that random forest regression better than the other two regression models. RF can deal with numerical data and data, but the other two only can deal with numerical variables and continuous variables but the other two only can deal with numerical data. So RFR can be more widely applied. The results show that random forest regression better than the other two regression models.

The study about RFR is still going on. 2006, Quantile Regression Forest defined by Nicolai, is derived from random forests[17]. Nicolai has proved mathematically that Quantile Regression Forest is consistent. Quantile regression forests can be seen as one of the applicability of the nearest neighbor classification and regression process[18]. In addition Brence and Brown improve the robust RF regression algorithm based on the to booming algorithm[19].

5 The Application of Random Forest

RF can be used to deal with micro-information data, and the accuracy of RF is higher than those traditional predictions. So in recently 10 years, Random Forest has been got a rapid development, and widely used in many areas, such as bioinformatics,medicine, management science, economics. In bioinformatics, Smith et al. studied the tracking data on bacteria by RF, and compared with Discriminant Analysis method. Alonso et al. use biomarkers parasite to discriminate fish stocks [20-21]; In medicine, Using RF technology such as Lee to help lung CT images of lung nodules automatic detection, and also in the RF (CAC)[22].In China, Jia FuCang,Li Hua, Using RF to the Dhoop magnetic resonance image segmentation, and that the RF has fast speed and high accuracy, is a promising muti-channel image segmentation method[23].The main application in economic management field, is predicating the loss degrees of customers. Bart used RF in customer relationship management, found that the effect of RF is better than ordinary linear regression and Logistic model[24]. Coussement et al. compared the predictive ability of SVM, logistic model and the RF in loss of customers, found that RF is always better than the SVM[25].Burez et al. applied weighted RF in loss of customers, comparing with the RF, and found the weighted RF has better prediction[26].

Today, the range of application of RF is very broad, in addition to the above mentioned application ,the RF also used in ecology[27-28], remote sensing geography terms[29-30], customer's loyalty forecasting[31];Lessmann etc. also use Random Forest predict horse racing winning, and that the predictions of the Random Forest is superior to traditional forecasting methods can bring in huge commercial profits[32].

6 Conclusions and Outlook

In summary, the RF as a combination of the tree classifer is an effective classification predicting tool. It has the following advantages:1) the accuracy of random forests is not less than Adaboost, run faster, and does not produce over-fitting. 2) the OOB data can be used to estimate the the RF generalization error, correlation and strength, can also estimate the importance of individual variables. 3) the combination of bagging and the random selection of features to split allows the RF to better tolerate noise. 4) RF can handle continuous variables and categorical variables.

Recently, the RF theory is more mature and the application range is becoming wilder. But there many work to do to further improve RF, and use RF to much wider fields. Hopping the interested scholars can do further research.

References

1. Breiman, L.: Bagging Predictors. Machine Learning 24, 123–140 (1996)
2. Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. Neural Computation 9, 1545–1588 (1997)
3. Dietterich, T.: An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization. Machine Learning, 1–22 (1998)
4. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. on Pattern Analysis and Machine Intelligence 20(8), 832–844 (1998)
5. Breiman, L.: Using adaptive bagging to debias regressions, Technical Report 547, Statistics Dept. UCB (1999)
6. Breiman, L.: Random Forests. Machine Learning 45(1) (2001)
7. Tibshirani, R.: Bias, Variance, and Prediction Error for Classification Rules, Technical Report, Statistics Department, University of Toronto (1996)
8. Wolpert, D.H., Macready, W.G.: An Efficient Method to Estimate Bagging's Generalization Error. Machine Learning (1997) (in press)
9. Breiman, L.: Out-of-bag estimation [EB/OL] (June 30, 2010), `http://stat.berkeley.edu/pub/users/Breiman/OOBestimation.ps`
10. Breiman, L.: Predition Games and Arcing Algorithms. Neural Computation 11, 1493–1517 (1999)
11. Bauer, E., et al.: An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants. Machine Learning 36, 105–142 (1999)
12. Freund, Y., Shapire, R.: Experiments with a new boosting Algorithm. In: Machine Learing: Proceedings of the Thirteenth International Conference, pp. 148–156 (1996)
13. Ho, T.K.: Random: Decision Forests. In: Proceeding of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada, August 14-18, pp. 278–282 (1995)