A Review on Random Forest: An Ensemble Classifier

Aakash Parmar(✉), Rakesh Katariya, and Vatsal Patel

SS Agrawal Institue of Engineering and Technology, Navsari, Gujarat, India
parmar.akashll@gmail.com, raahir80@gmail.com,
vatsalpatel4u@gmail.com

Abstract. Ensemble classification is an information mining approach which utilizes various classifiers that cooperate for distinguishing the class label for new unlabeled thing from accumulation. Arbitrary Forest approach joins a few randomized choice trees and totals their forecasts by averaging. It has grabbed well-known attention from the community of research because of its high accuracy and superiority which additionally increase the performance. Now in this paper, we take a gander at improvements of Random Forest from history to till date. Our approach is to take a recorded view on the improvement of this prominently effective classification procedure. To begin with history of Random Forest to main technique proposed by Breiman then successful applications that utilized Random Forest and finally some comparison with other classifiers. This paper is proposed to give non specialists simple access to the principle thoughts of random forest.

1 Introduction

One of the vital and intriguing field of Computer Science is data mining that has gotten a considerable measure of attention by the exploration group especially finished in the recent time. It is vital in light of the fact that it has some expertise in dissecting the information from alternate points of view and compressing it into valuable data - data which can be able to be used to expand incomes, cut expenses, or discretely both. It is intriguing on the grounds that it applies strategies at the convergence of numerous orders including manmade brainpower, machine learning, measurements, and database frameworks [1].

In this paper, the applicable task to us is grouping which sorts out information into the classes by utilizing the foreordained class names. The characterization calculation picks up from the planning set and manufactures a model, furthermore named as classifier. After that the model is connected to anticipate the class marks for the unclassified questions in testing the information.

Ensemble classification techniques prepare a few classifiers then consolidate their outcomes over a voting procedure. The most generally utilized collective strategies are bagging and boosting [2]. Bootstrap amassing depends on preparing numerous classifiers taking place on samples of bootstrapped from the preparation set. All of these have been appeared to diminish the fluctuation of the classification. Conversely, boosting utilizes iterative re-training, wherever the mistakenly characterized samples

J. Hemanth et al. (Eds.): ICICI 2018, LNDECT 26, pp. 758–763, 2019.
https://doi.org/10.1007/978-3-030-03146-6_86

are specified with extra weight in each progressive emphases. This makes the calculation moderate (considerably steadier then bagging) while as a rule it is significantly more precise than bagging. Improving, for the maximum part lessens the variance in the bias of the order and has been appeared to be an exceptionally exact classification strategy. Be that as it may, it has different disadvantages: it is moderate, it can overstrain and it is delicate to noise [2]. Hence, there is much enthusiasm for exploring strategies, for example, Random Forests, however Random Forests have been appeared to be practically identical to boosting as far as exactnesses, yet without the drawbacks of boosting. Likewise, the Random Forests remains computationally considerably less raised than boosting.

The random forest is a hot spot of this domain in recent years, as a combined classifier, the random forest can increase forecasting accuracy by combining the outcomes from each single classifier. It is also effective to solve the problem of overfitting and has broad applications in many fields, including text classification and image classification and so on [3].

2 Background

To improve the precision of classification, ensemble learning is used which is one of the application of ensemble classification where various models are utilized to tackle a similar issue [4]. In ensemble classification, numerous classifiers are utilized and are much precise than the different classifiers that are available in the ensemble. A simple voting plan is used at that point to decide the class name for not labeled occurrences. An effective and very simple voting pattern is majority voting [5]. In this type of voting, individual classifier within the ensemble is requested to forecast the class label and its occurrence being reflected. After the interrogation of all the classifiers, the class that gets the maximum amount of votes is given back as an ultimate conclusion of ensemble. One alternative is veto voting scheme in which one particular classifier prohibits the choice of every other classifiers [6].

Three generally utilized ensemble methodologies are, in particular, bagging, stacking and boosting. It is an incremental procedure of building a classification of classifiers, wherever every classifier takes a shot at the mistakenly classified examples of the past one in the arrangement. The added class of ensemble methods is the Bootstrap Aggregating (Bagging) [8]. Bagging includes constructing every classifier in the gathering utilizing an arbitrarily drawn data sample, having every classifier assigned an equivalent vote while marking unlabeled occurrences. Bagging is identified as the most robust than boosting in contradiction of model overfitting. The main representative of bagging is RF [9].

Experiments conducted by Dietterich [10] to associate the three approaches and its performance for building ensembles of classifiers with C4.5, namely, bagging, boosting and randomization. If small or no amount of noise is present in the information, experiments indicated that boosting has been shown superior to randomization and bagging. Bagging and randomization proved alike performance, still, with little noise present within the data, randomization has better performance. Boosting execution appeared to disintegrate by noise. Furthermore bagging execution appeared to enhance, for which it could use the ailment to deliver more assorted classifiers.

3 Random Forest

The random forest was first proposed by Leo Breiman from the University of California in 2001 [9]. It is composed of many basic classifiers (decision tree) which are completely independent from each other. Input a test sample to the new classifier and the class label of this sample can be decided based on the voting results from each single classification. The entire process of classification based on the random forest is shown in the Fig. 1.

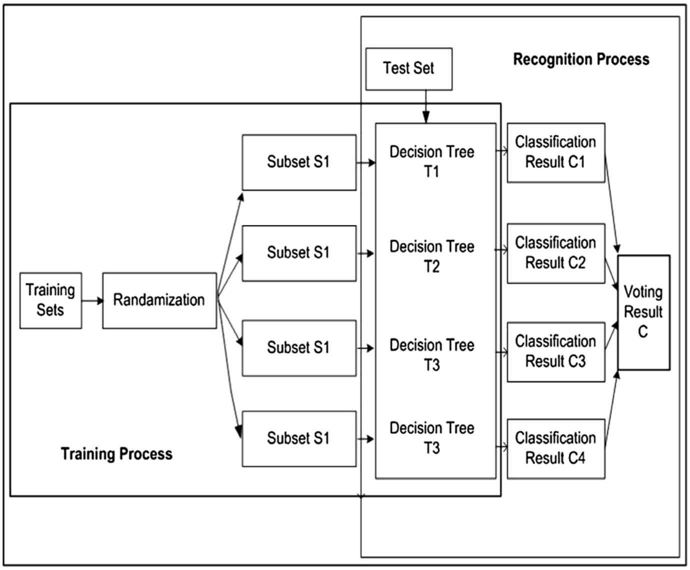

Fig. 1. Conceptual framework of random forest classifier [3].

The following are the main steps in building up the random forest classifier [9]:

(a) Set a proper value for the variable called “M” which is the number of elements of each feature subset.
(b) Select a new feature subset θ_k from the whole feature set at a random depending on the value of M. θ_k is independent from the other subset in the sequence of $\theta_1, \ldots, \theta_k$.
(c) Training the data set with the feature subset to create decision tree for each group of training set. Each single classier can be expressed as h(X,θ_k) (where X indicates the inputs).

(d) Choose a new θ_k and repeat the process above until travel all the feature subsets. A random forest classifier is completed.

(e) Input the test set. Decide the class label of this sample based on the voting results from each single classification.

The random forest is made up of large amount of decision trees. The random operation is introduced in the build process, including the selection of samples subset and feature subset, to guarantee the independence of each decision tree, improve classification accuracy and gain better generalization ability [9].

Using random operation in selecting sample subset is to pulling the subsets as training sets from the original samples with the method of bagging. This operation ensures the freedom of each preparation subset. In selecting feature subsets, the method of bagging is used again to choose subsets based on the value of variable M from the entire feature set. These feature subsets is used to training datasets. Ranking all the features according to its importance based on the contribution of each training results to final decision is also can be implemented. Variable M has a big influence on the strength and correlation of the random forest. Both of the strength and correlation can be improved with increase of M with in certain range and vice versa. Breiman conducted tests to show that the random forest will have the best performance when the variable M is close to the value of D [9]. (where D indicates the number feature values).

The random operation in the random forest significantly improved the performance of classifier. Because of the build process of each single decision tree is very fast, the parallelization in the creation of the random forest can be realized, which improve the speed of classification greatly.

4 Literature Review

Aung et al. [11] proposed RFC based on random forest method for multiple category for classification of web page. The objective of website page classification is to characterize the data on Internet into a specific number of predefined categories. Website pages classification can likewise help enhance the nature of web inquiry. Experiment performed on data collected from Yahoo web site using Decision tree classifier and RFC. They first consider the pattern of the attributes for complete categories of the downloaded web pages. Analyzing the web pages afterwards, they have found that few of the terms are normally present in the similar web pages. The classification accuracy of RFC is compared with classification accuracy of decision tree. Results clears that Random Forest Classifier obtained almost higher than 90% classification accuracy in all categories which is also better than Decision Tree Classifier in all categories of web pages.

Azar et al. [12] experimented on A Random Forest Classifier for Lymph Diseases. Exactness of classification algorithms utilized as a part of sickness diagnosing is positively an imperative issue to be considered. High dimensional information, all in all, requires the extraction of most engaging or discriminative components to be chosen and henceforth the measurement of dataset is lessened. Dimension lessening methodology is helpful to diminish dataset unpredictability with the conceivable favorable position of expanded classification execution. Expelling the quantity of superfluous elements for model usage makes screening tests speedier, more helpful and less expensive and nature of the produced include subset solutions. The lymphatic framework helps the immune framework in expelling and wrecking waste, trash, dead platelets, pathogens, poisons, and disease cells. Additionally, it expels overabundance liquid, and waste items from the interstitial spaces among the cells. Objectives of the experiment are to increase the performance of classification correctness and obtain the significant features as an indicator to the existence of lymph disease class. Genetic Algorithm is used to find an optimal feature set thus it reduces the dimension of lymph dieses dataset. RFC is used for classification task. Proposed GA-RFC approach obtained 92.2% classification accuracy.

Song et al. [3] applied Random Forest Classifier in Droplet Fingerprint Recognition. Identification methods of liquid include minimum distance algorithm, Bayes algorithm, support vector machines, neural networks and fuzzy recognition method. Experiment performed in paper is for effective classification of liquid from fingerprint droplet which contains liquid of nine different classes. RFC is implemented in python. Out of 650 test samples of different classes of liquid, 648 test samples are correctly recognized which tells that Random Forest Classifier achieves very high recognition rate for the experiment. Class 3 and 5 remain out from 100% recognition rate.

Pal [13] has constructed one classifier for Remote Sensing Classification named as Random Forest Classifier. The above mentioned classifier is a tree-based classifier, which supports vector machines, created for increasing the margin amongst two dissimilar classes, that are used in this experiment. The random forest classifier utilizes the Gini Index as a trait choice measure, that measures the polluting influence of an ascribe regarding the classes. The reviews recommend that the decision of the pruning techniques, and not the attribute determination measures, influence the execution of tree based classifiers. Data of an agriculture zone nearby Littleport, Cambridgeshire, UK that was used for remote sensing classification. Support Vector Machine and Random Forest Classifier are used for land classification task. Classification accuracy and training time of RFC is compared with SVM. In results, Random Forest Classifier achieves higher classification accuracy and even training time is less than Support Vector Machine.

Garcia et al. [14] proposed Random Forest Based Ensemble Classifiers for Predicting Healthcare-Associated Infections in Intensive Care Units. Observation and counteractive action of diseases procured in the hospital facility condition is a vital test in the present health systems specified the great impact of these kinds of contaminations on patient mortality in addition to the costs of sanitary. Implementations of surveillance and stoppage actions have decreased the incidence amount of infections in ICU in addition to their adverse effects. In this effort data mining methods are used to discover the maximum significant HAI hazards and to recognize patients that are more vulnerable to infections concerning their appearances, invasive devices, treatment and other information concerning their stay in ICU. All the dataset used in study contains information about 4616 patients hospitalized in ICU of the university hospital of Salamanca. J48 tree and Bayesian Networks are used as simple classifiers. Random Forest, Bagging and AdaBoost are used as ensemble classifiers. Obtained results proves that AdaBoost's combination with Random Forest achieves highest classification accuracy of 95.26% amongst all and bagging with Random Forest has second highest classification accuracy of 95.10%

5 Conclusion

Random Forest Classifier has higher classification rate than single classifiers and also combination of Random Forest with AdaBoost gives better classification accuracy. Random Forest Classifier takes less training time than Decision Tree and Support Vector Machine hence it is fast. It is very useful in cases of large data sets and also avoids the overfitting problem by handling noises presents in data sets. Due to these capabilities, Random Forest an approach proposed by Breiman getting more popularity day by day in research community for classification task.

References

1. Fayyad, U.P, Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Mag. **17**(3), 37 (1996)
2. Briem, G.J., Benediktsson, J.A., Sveinsson, J.R.: Multiple classifiers applied to multisource remote sensing data. IEEE Trans. Geosci. **40**(10) (2002)
3. Song, Q., Liu, X., Yang, L.: The random forest classifier applied in droplet fingerprint recognition. In: 2015 12th International Conference on FSKD, pp. 722–726. IEEE, August 2015
4. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach. Learn. **51**(2), 181–207 (2003)
5. Lam, L., Suen, C.Y.: Application of majority voting to pattern recognition: an analysis of its behavior and performance. IEEE **27**(5), 553–568 (1997)
6. Shahzad, R.K., Lavesson, N.: Veto-based malware detection. In: 2012 Seventh International Conference on Availability, Reliability and Security (ARES), pp. 47–54, 20 August 2012
7. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
8. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996)
9. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
10. Dietterich, T.G.: An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach. Learn. **40**(2), 139–157
11. Aung, W.T., Myanmar, Y., Hla, K.H.: Random forest classifier for multi-category classification of web pages. In: Services Computing Conference. APSCC 2009, pp. 372–376. IEEE (2009)
12. Azar, A.T., Elshazly, H.I., Hassanien, A.E., Elkorany, A.M.: A random forest classifier for lymph diseases. Comput. Methods Programs Biomed. **113**(2), 465–473 (2014)
13. Pal, M.: Random forest classifier for remote sensing classification. IJRS **26**(1), 217–222 (2005)
14. García, M.N., Herráez, J.C., Barba, M.S., Hernández, F.S.: Random Forest Based Ensemble Classifiers for Predicting Healthcare-Associated Infections in Intensive Care Units. Springer (2016)