Indian Journal of Science and Technology, *Vol 8(S8), 614–619, April 2015*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
DOI: 10.17485/ijst/2015/v8iS8/71714

Improving Classification Accuracy based on Random Forest Model through Weighted Sampling for Noisy Data with Linear Decision Boundary

S. Sasikala[1], S. Bharathidason[2*] and C. Jothi Venkateswaran[3]

[1]Department of Computer Science, IDE, University of Madras, Chennai - 600005, India; sasikalarams@gmail.com
[2]Department of Computer Science, Loyola College, Chennai, India; bharathidasan82@gmail.com
[3]Department of Computer Science, Presidency College, Chennai, India; jothivenkateswaran@yahoo.co.in

Abstract

Background: Random forest algorithms tend to use a simple random sampling of observations in building their decision trees. The random selection has the chance for noisy, outlier and non informative data to take place during the construction of trees. This leads to inappropriate and poor ensemble classification decision. This paper aims to optimize, the sample selection through probability proportional to size sampling (weighted sampling) in which the noisy, outlier and non informative data points are down weighted to improve the classification accuracy of the model. **Methods:** The weights of each data point is determined in two aspects, finding each data point influence on the model through Leave-One-Out method using a single classification tree and measuring the deviance residual of each data point using logistic regression model, these are combined as the final weight. **Results:** The proposed Finest Random Forest (FRF) performs consistently better than the conventional Random Forest (RF) in terms of classification accuracy. **Conclusion:** The classification accuracy is improved when random forest is composed with probability proportional to size sampling (weighted sampling) for noisy data with linear decision boundary.

Keywords: Classification Accuracy, Decision Trees, Noisy Data, Outlier, Random Forest, Weighted Sampling

1. Introduction

Random Forest (RF) builds a classification ensemble with a set of decision trees that grow using randomly selected subspaces of data[1–4]. There are many studies showing that RFs have impressive predictive performance in regression and classification problems in various fields, including financial forecasting, remote sensing, and genetic and biomedical analysis[5–13].

It is common that noise and outliers exist in real world datasets due to errors such as, typographical errors or measurement errors. When the data is modeled using machine learning algorithms, the presence of noise and outliers can affect the model that is generated. Improving how learning algorithms handle noise and outliers can produce better models[14]. Outlier problem could be traced to its origin in the middle of the eighteenth century, when the main discussion is about justification to reject or retain an observation. "It is rather because of the loss in the accuracy of the experiment caused by throwing away a couple of good values is small compared to the loss caused by keeping even one bad value"[15].

Handling noise and outliers has been addressed in a number of different ways, beginning with preventing overfit. A common approach to prevent overfit is adhering to Occam's razor which states that the simplest hypothesis that fits the data tends to be the best one. Using Occam's razor, a trade off is made between accuracy on the training set and the complexity of the model, preferring a simpler model that will not overfit the training set.

*Author for correspondence

Another technique to prevent overfit is to use a validation set during training to ensure that noise and outliers are not learned[16].

In a dataset all the observations are not equally informative to build a model. Some observations are highly informative and some are not. The performance of the model is improved while neglecting or giving less importance for the non informative observations during the model construction[17]. In random forest, random selection has the chance for noisy, outlier and non informative data to take place during the construction of trees. This will decrease the classification accuracy of the individual tree in the forest. This paper aims to optimize, the sample selection through probability proportional to size sampling (weighted sampling) in which the noisy, outlier and non informative data points are down weighted, to improve the classification accuracy of the model.

2. Random Forest Algorithm

Random forest is an ensemble prediction method by aggregating the result of individual decision trees. In the past decade, various methods have been proposed to grow a random forest[1–3,18]. Among these methods, Breiman's method[1] has gained increasing popularity because it has higher performance against other methods[19].

Let D be a training dataset in an *M*-dimensional space X, and let Y be the class feature with total number of c distinct classes. The method for building a random forest[1] follows the process including three steps[18]:

Step 1: Training data sampling: use the bagging method to generate *K* subsets of training data $\{D_1, D_2,..., D_K\}$ by randomly sampling D with replacement;

Step 2: Feature subspace sampling and tree classifier building: for each training dataset D_i ($1 \leq i \leq K$), use a decision tree algorithm to grow a tree. At each node, randomly sample a subspace X_i of F features ($F << M$), compute all splits in subspace X_i, and select the best split as the splitting feature to generate a child node. Repeat this process until the stopping criteria is met, and a tree $h_i(D_i, X_i)$ built by training data D_i under subspace X_i is thus obtained.

Step 3: Decision aggregation: ensemble the K trees $\{h_1(D_1, X_1), h_2(D_2, X_2), ..., h_K(D_K, X_K)\}$ to form a random forest and use the majority vote of these trees to make an ensemble classification decision.

The algorithm has two key parameters, *i.e.*, the number of *K* trees to form a random forest and the number of *F* randomly sampled features for building a decision tree. According to Breiman[1], parameter *K* is set to 100 and parameter *F* is computed by $F = [\log_2 M + 1]$. For large and high dimensional data, a large *K* and *F* should be used.

3. Weight Calculation of Training samples Based on the Influence and the Deviance Residual

In the proposed approach, before constructing a random forest with many trees, a single classification tree is used to measure the influence, and the logistic regression model is used to measure the deviance residual of each data point, which will be used to train the Random Forest model.

The weights of each data point is determined in two aspects, which are *(i) finding each data point influence on the model through Leave-One-Out method* using a single classification tree (*ii) measuring the deviance residual of each data point* using logistic regression model. The AUC accuracy is used to measure the performance of the classification tree.

If a data point has high negative influence (degrade the performance) on the model (a classification tree) and has high deviance residual, then it will be treated as a noisy or outlier data point. These, data points will be down weighted to maximize the overall classification accuracy during the construction of Random Forest model.

3.1. Measuring the Influence of Training Samples using Leave-One-Out Method

Leave-one-out is a method where in each iteration, all the data except for a single observation are used for training the model. Using this method each observation's influence on the model can be measured. A single classification tree is used to measure the influence of each data point. The model (a tree) trained without a single observation is called Reduced Model and a model (a tree) trained with full set of training observations is called Full model. The influence of a data point is the difference between these two models performance, which is as follows

$$Influence_i = \eta_{Full} - \eta_{Reduced}$$

Where, η_{Full} is the AUC accuracy of the full model and $\eta_{Reduced}$ is the AUC accuracy of the reduced model[17].

Likewise, each data point's influence on the model is estimated. The estimated influence of each data point is normalized using min-max normalization and it is used as a part of weight calculation to perform the probability proportional to size sampling (weighted sampling) in random forest construction.

3.2 Measuring the deviance residual of Training Samples

The logistic regression is a linear model, works well for the dataset with linear decision boundary. A decision boundary is the region of a problem space in which the output label of a classifier is ambiguous. If the decision boundary is a hyper plane, then the classification problem is linear, and the classes are linearly separable.

The logistic regression model is used to measure the standardized deviance residual of each data point. The raw residual for observation i is the difference between the observed and predicted value, i.e.

$$r_i = y_i - \hat{y}_i$$

These are difficult to interpret, as they will have different levels of natural variation, if we spot what seems to be a large residual, compared with the rest, it may simply be caused by natural variation rather than a problem with the model. This issue can be resolved by calculating a standardized deviance residual as follows,

$$r_{D_i} = sign(r_i)\sqrt{\frac{D_i}{1-h_i}}$$

where sign(r_i) is the sign of r_i, and *Di* is the contribution to the deviance made by the *i*th observation.

Similarly, each data point's standardized deviance residual is estimated. The absolute deviance residual of each data point is normalized and used as a part of weight calculation to perform the probability proportional to size sampling (weighted sampling) in building the random forest.

3.3 Combining the Weights

The measured Influence and the deviance residual are combined as a weight for each data point in the training sample and these are used to carry out the probability proportional to size sampling for building a random forest.

$Weight_i = Influence_i * (1 - Deviance_i)^2$, i = 1, 2, 3, ... , n

Thus, the combined weight of each data point in the training sample is calculated and the same is used for weighted sampling to train the Random Forest.

Based on the range of Influence and deviance residual the weights may vary for each data point. If a data point has high negative Influence and also has high deviance residual, then it is highly down weighted to optimize the Random Forest through weighted sampling.

4. Proposed Finest Random Forest Algorithm

Let D be a training dataset in an *M*-dimensional space X, and let Y be the class feature with total number of c distinct classes. The method to build a Finest Random Forest from *X* with *probability proportional to size sampling* (weighted sampling) based on the weight calculated for each data point mentioned in section3 follows the following steps.

Step 0: Weight Initialization: Assign the weight for each Training sample based on the Influence and deviance residual of the sample;

Step 1: Training data sampling: use the bagging method to generate *K* subsets of training data $\{D_1, D_2, ..., D_K\}$ by Probability Proportional to size sampling (weighted sampling) D with replacement;

Step 2: Feature subspace sampling and tree classifier building: for each training dataset D_i ($1 \le i \le K$), use a decision tree algorithm to grow a tree. At each node, randomly sample a subspace X_i of F features (F << M), compute all splits in subspace X_i, and select the best split as the splitting feature to generate a child node. Repeat this process until the stopping criteria is met, and a tree $h_i(D_i, X_i)$ built by training data D_i under subspace X_i is thus obtained.

Step 3: Decision aggregation: ensemble the K trees $\{h_1(D_1, X_1), h_2(D_2, X_2), ... , h_K(D_K, X_K)\}$ to form a random forest and use the majority vote of these trees to make an ensemble classification decision.

The algorithm has two key parameters, *i.e.*, the number of *K* trees to form a random forest and the number of *F* randomly sampled features for building a decision tree. For large and high dimensional data, a large *K* and *F* should be used.

5. Data Source

Detailed information of the Wisconsin Prognostic Breast Cancer (WPBC) dataset is available in the UCI Machine Learning Repository[20]. The Blood Transfusion Service center dataset[21] and Mammographic Mass dataset[22] information is obtained from UCI Machine Learning Repository. The SPECTF heart dataset is also obtained from UCI Machine Learning Repository[23]. In all the dataset 70% of the data used as a training sample, remaining 30% of the sample used for testing the model.

6. Results and Discussion

A series of experiments were conducted on four datasets such as, Breast cancer, Blood transfusion, Mammographic and Heart datasets. In each dataset, it is concluded that the proposed Finest Random Forest (FRF) performs consistently better than the conventional Random Forest (RF). The AUC accuracy is used as a metric to evaluate the performance of the algorithms.

6.1 Performance Analysis

The proposed finest random forest method is compared with Breiman's method, the average accuracy of 10 results were computed by performing 10 rounds of experiments on each dataset. The weight of each data point of the training sample is calculated based on the influence and deviance residual of the same. In each round, probability proportional to size sampling (weighted sampling) is performed to construct the Finest Random Forest. The random forest also builds by Breiman's method by selecting the training samples randomly. The average AUC accuracy of different random forest consisting different number of trees (ranging from 10 to 100 trees with increments 10) generated by the finest random forest method (corresponding to column FRF) and Breiman's method (corresponding to column RF) from four datasets are shown in Table 1. The proposed method achieves high classification accuracy by improving the AUC accuracy on the four datasets.

6.2 Comparison of Classification Accuracy

The preceding section has shown that the Finest Random Forest (FRF) outperforms the conventional random forest. The AUC accuracy of the random forest is increased by performing probability proportional to size sampling (weighted sampling) based on the weights calculated for each data point in the training samples. In the above mentioned four datasets, increasing the AUC accuracy ranging from 2% to 11% has achieved with the finest random forest than the original random forest.

Based on the complexity pattern of the dataset in terms of noise and outlier, the percentage of improvement in AUC accuracy of the random forest may vary. The proposed finest random forest method increase the classification accuracy on the four dataset is shown in Figure 1. The dotted blue curves represent the AUC accuracy obtained with random forest and the red curves represent the AUC accuracy obtained with Finest Random Forest.

7. Conclusion

This paper presents an evaluation method to identify the noisy, outlier and non informative data points in the training sample, and proposed a finest random forest

Table 1. Comparison of Classification accuracy between Random Forest (RF) and Finest Random Forest (FRF)

Dataset / Trees	Mammographic		Blood Transfusion		Breast Cancer		Heart Disease	
	RF	FRF	RF	FRF	RF	FRF	RF	FRF
10	0.8701	0.90056	0.637559	0.672575	0.550725	0.603261	0.76801	0.774948
20	0.891015	0.903158	0.640306	0.702403	0.549215	0.602959	0.771899	0.784238
30	0.89144	0.903571	0.636469	0.700353	0.551932	0.60087	0.78876	0.813372
40	0.884917	0.905077	0.657444	0.707352	0.583937	0.618357	0.79509	0.815439
50	0.891116	0.914256	0.640917	0.711909	0.547403	0.632778	0.80155	0.814664
60	0.896576	0.916913	0.639884	0.71023	0.558877	0.632246	0.80084	0.813954
70	0.897078	0.913518	0.658992	0.713043	0.574879	0.626739	0.800388	0.821641
80	0.894303	0.910811	0.646956	0.713261	0.566703	0.621413	0.805039	0.824339
90	0.899174	0.913932	0.650889	0.720105	0.557669	0.639191	0.810659	0.821895
100	0.90245	0.914315	0.643947	0.724267	0.582428	0.644626	0.816085	0.830911

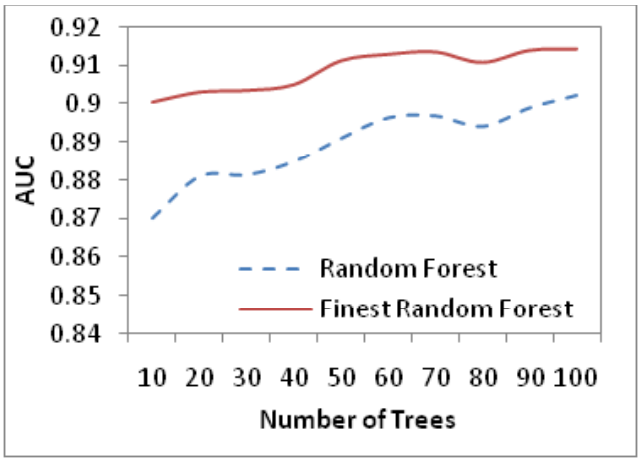

Figure 1a. Mammographic.

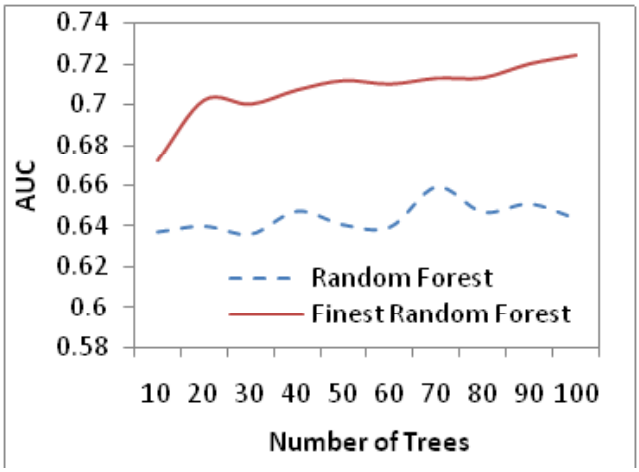

Figure 1b. Blood Transfusion.

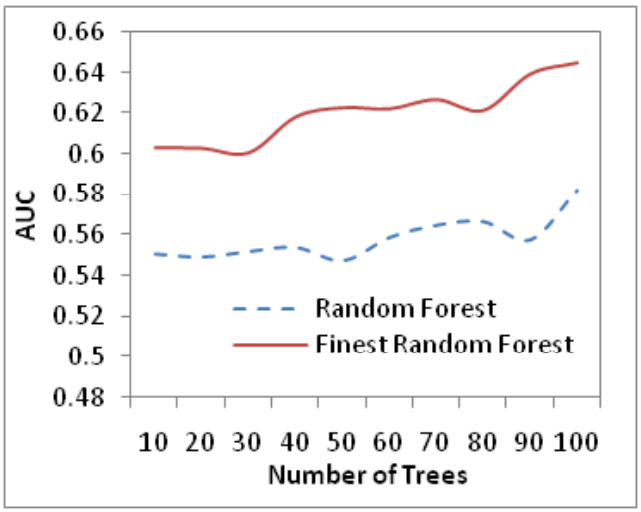

Figure 1c. Breast Cancer.

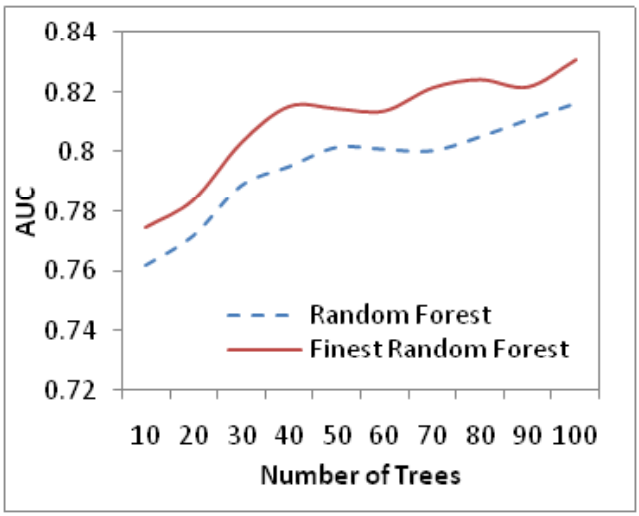

Figure 1d. Heart Diseases.

Figure 1. Comparison of Classification accuracy between Random Forest (RF) and Finest Random Forest (FRF)

algorithm which replaces the existing random sampling with probability proportional to size sampling (weighted sampling) in the construction of random forest model. The classification accuracy is improved when a random forest is composed with probability proportional to size sampling (weighted sampling) in which the data points has high deviance residual and also negatively influence the model are down weighted. As a result, the prediction accuracy of the random forest is improved in classification problems.

8. Acknowledgement

We are grateful to Prof. Syluvai Antony, Assistant Professor, Dept. of Statistics, Loyola College and Dr. M. Raja, Assistant Professor, Dept. of Advanced Zoology and Biotechnology, Loyola College, Chennai for their constant support and valuable suggestions to complete this research work.

9. References

1. Breiman L. Random forests. Mach Learn. 2001; 45:5–32.
2. Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn. 2000; 40(2):139–57.
3. Ho TK. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998; 20(8):832–44.

4. Amit Y, Geman D. Shape quantization and recognition with randomized trees. Neural Comput. 1997; 9(7):1545–88.
5. Goldstein B, Polley E, Briggs F. Random forests for genetic association studies. Stat Appl Genet Mol Biol. 2011; 10(1):1–34.
6. Siroky D. Navigating random forests and related advances in algorithmic modeling. Stat Surv. 2009; 3:147–63.
7. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. Mipred: classification of real and pseudo microrna precursors using random forest prediction model with combined features. Nucleic Acids Res. 2007; 35(2):339–44.
8. Palmer D, O'Boyle N, Glen R, Mitchell J. Random forest models to predict aqueous solubility. J Chem Inf Model. 2007; 47(1):150–8.
9. Kumar M, Thenmozhi M. Forecasting stock index movement: A comparison of support vector machines and random forest. Indian Institute of Capital Markets 9th Capital Markets Conference, 2006.
10. Diaz-Uriarte R, de AndrsSs SA. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. 2006; 7:3–15.
11. Ward M, Pajevic S, Dreyfuss J, Malley J. Short-term prediction of mortality in patients with systemic lupus erythematosus: Classification of outcomes using random forests. Arthritis Rheum. 2006; 55:74–80.
12. Shi T, Seligson D, Belldegrun A, Palotie A, Horvath S. Tumor classification by tissue microarray profiling: Random forest clustering applied to renal cell carcinoma. Mod Pathol.2005; 18(4):547–57.
13. Pal M. Random forest classifier for remote sensing classification. Int J Rem Sens. 2003; 26(1):217–22.
14. Smith MR, Martinez T. Improving classification accuracy by identifying and removing instances that should be misclassified. Proceedings of The 2011 International Joint Conference on neural networks, IEEE. 2011; p.2690–7.
15. Barnett V, Lewis T. Outliers in statistical data. John Wiley and Sons. 1978; p. 1.
16. Quinlan JR. C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, CA, USA. 1993.
17. Bharathidason S, Jothi Venkataeswaran C. Improving prediction accuracy based on optimized random forest model with weighted sampling for regression trees, Int J Comput Trends Tech. 2015; 21(1):23–8.
18. Baoxun Xu, Junjie Li, Qiang Wang, Xiaojun Chen. A tree selection model for improved random forest. Bulletin of Advanced Technology Research. 2012; 6(2).
19. Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP. A comparison of decision tree ensemble creation techniques. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2007; 29(1):173–80.
20. Wolberg WH, Street WN, Mangasarian OL. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Anal Quant Cytol Histol. 1995; 17(2):77–87.
21. Yeh I-C, Yang K-J, Ting T-M. Knowledge discovery on RFM model using Bernoulli sequence, Expert Systems with Applications. 2008; doi:10.1016/j.eswa.2008.07.018.
22. Elter M, Schulz-Wendtland R, Wittenberg T. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. Med Phys. 2007; 34(11):4164–72.
23. Cios KJ, Kurgan L. Hybrid inductive machine learning: An overview of CLIP Algorithms. In: Jain LC, Kacprzyk J, editors. New learning paradigms in soft computing, physica-verlag. Springer; 2001.