Chapter 31
Bayesian Random Forest for the Classification of High-Dimensional mRNA Cancer Samples

Oyebayo Ridwan Olaniran and Mohd Asrul Affendi Bin Abdullah

Abstract The goal of many machine learning algorithms is to adequately identify the informative biomarkers in the biological samples useful for predicting disease outcome. Several algorithms have been proposed to perform this task using high-dimensional genomic messenger Ribonucleic Acid (mRNA) data. High-dimensionality poses serious problem in statistical analysis in terms of parameter estimation and inference. To address this problem, a powerful method has been developed called Random forest. Random forest was able to tackle high-dimensionality problem but it fails because it's more of computer program than a statistical learning method thus uncertainty in prediction cannot be quantified. In this paper, we develop Bayesian Random Forest (BRF) model for the classification of high-dimensional mRNA data. Bayesian procedures are the emerging solution to most applications of statistics in the recent time and in fact it has the least error rate in theory. In addition, they give appealing results in terms of parameter uncertainty, model uncertainty and data uncertainty. BRF model fitting and inference were achieved via Metropolis-Hasting (MH) MCMC algorithm. The model strength was illustrated using bake-off of 10 different mRNA cancer datasets. Results from data calibration established appreciable supremacy over competing methods.

Keywords Bayesian · Random forest · mRNA · High-dimensional · Classification

O. R. Olaniran (✉) · M. A. A. B. Abdullah
Faculty of Applied Sciences and Technology,
Department of Mathematics and Statistics, Universiti Tun Hussein Onn Malaysia,
Pagoh Educational Hub, 84600 Pagoh, Johor, Malaysia
e-mail: rid4stat@yahoo.com

M. A. A. B. Abdullah
e-mail: afendi@uthm.edu.my

L.-K. Kor et al. (eds.), *Proceedings of the Third International Conference on Computing, Mathematics and Statistics (iCMS2017)*,
https://doi.org/10.1007/978-981-13-7279-7_31

1 Introduction

The growth in computer applications have enhanced collection and analysis of big datasets. Big datasets are often referred to as high-dimensional data in statistical parlance. The difficulties faced while analyzing big datasets has led to development of many statistical or machine learning procedures in the recent time [1]. In most areas of research especially bioinformatics, it is usual to have relatively small sample sized datasets collected on large number of features.

Random forests (RF), a tree based non-parametric method originally proposed by [2] is one of the popular methods for handling high-dimensional data, mainly because of its computational speed and high accuracy. Bayesian procedures are the emerging solution to most applications of statistics in the recent time in fact it has the least error rate in theory [3–6]. Chipman et al. [7] proposed Bayesian Additive Regression Trees (BART) which is a probabilistic approach to sum of trees model. However, BART is more of Bayesian approach to sum of trees model than to call it a Bayesian random forest. Specifically, BART did not incorporate bootstrapping of trees as in RF but a posterior distribution of trees. In addition, BART controls tree depth by imposing restrictive priors on tree with large daughter nodes. BART uses prior distribution specification as a pruning tool to avoid large trees [8]. Taddy et al. [9] proposed Bayesian forest (BF) as a nonparametric Bayesian approach to RF. They used posterior of trees instead of bootstrap of trees based on a nonparametric Bayesian model using multinomial draws. BF tried to mimic RF by replacing the bootstrapping procedure by [10] by its Bayesian counterpart (Bayesian Bootstrap, [11]). This implies BF focuses on the data generating process of RF but not its impurity measures.

Based on the aforementioned features of the Bayesian variation of Random Forest (RF), we observed that none of the existing methods fully captures the complete framework of RF and this affects their eventual results. Thus the goal of this research is to develop a complete Bayesian approach to Random Forest (RF). The method updates every aspect of RF using Bayesian reasoning. By way of example, we considered the case of binary classification with high-dimensional cancer datasets.

2 Decision Trees and Random Forests

Decisions trees is a class of methods under the broad Classification and Regression Trees (CART). The response variable of interest determines the type of model, such as decision trees if the response is categorical and regression trees if the response is continuous. CART do not have any statistical model but a set of steps called algorithm. CART modelling involves partitioning the feature space into M regions.

Formally, given training dataset $\left[y_i, x_{i1}, x_{i2}, \ldots, x_{ip}, i = 1, 2, \ldots, n\right]$, where y_i is a categorical outcome that assumes $k = 1, 2, \ldots, K$ values and x_i is the vector of

features. CART algorithm automatically decides on the splitting variables and splitting point. After, successful partitioning of the response to $R_1, R_2, \ldots, R_M$ regions, the closest form of model that CART assumes is;

$$y = \sum_{m=1}^{M} \beta_m I(x \in R_m) \tag{1}$$

where β_m is a constant in region m. Estimating β_m requires the computation of an impurity function. For classification case, the commonly used impurity functions are Misclassification Error Rate (MER), Gini Index, and deviance [12].

Random Forest (RF) update built CART trees in two steps; (i) bootstrapping the training dataset J times to obtain a total of J trees (ii) Subsampling $l < p$ features without replacement at each split step in each j tree. Thus given a CART model $\Im\left(\hat{\beta}_m : x \in R_m\right)$, RF model is;

$$\hat{y} = \sum_{j=1}^{J} \Im_j\left(\hat{\beta}_m : x \in R_m\right) \tag{2}$$

RF has two tuning parameters, the number of trees J and number of subsampled features l. Breiman [2] suggested using at least $J = 200$ and $l = \sqrt{p}$ for classification task.

3 Bayesian Random Forests

Section 2 established the weakness of RF as the necessary tuning parameters are not chosen by any probabilistic law. The approach is nothing but a trial and error hence often referred to as black box method. A quick solution to avert trial and error is to select the tuning parameters by cross validation but at the expense of computation time. Therefore, the focus of this research is to modify RF forest by updating the two steps in (2) via Bayesian approach. For the bootstrapping step, we propose the Bayesian Simple Random Sampling With Replacement (BSRSWR) described by the posterior distribution in (3);

$$P(\pi|a,b) = \frac{\Gamma(n+a+b)}{\Gamma(a+1)\Gamma(b+n-1)}\pi^a(1-\pi)^{b+n}, 0 \le \pi \le 1 \tag{3}$$

where π is the probability of selecting any $i \in n$ in each j step, $\Gamma(\mathrm{d})$ is the gamma function evaluated at d, a is the prior expected number of times any $i \in n$ could be selected and b is its complement. It's clear that the density function in (3) is a resemblance of $Beta(a+1, b+n-1)$. A weighted CART tree $\Im\left(\widehat{\beta}_m : x \in R_m\right)$

can be obtained using $\omega = Beta(a+1, b+n-1) \forall i \in n$. Similarly, for the subsampling of $l < p$ steps, we propose Bayesian Simple Random Sampling Without Replacement (BSRSWOR) with posterior density given in (4);

$$P(V|h,l,p,S,T) = \frac{\binom{S+V}{S+1}\binom{T+p-V}{T+l-h}}{\binom{S+T+p+1}{S+T+l+1}}, h \leq v \leq p-l+h \tag{4}$$

where V is the number of relevant features whose posterior is sought, h is the sample realization of relevant features, p is the total number of features, l is the number of subsampled features as in RF, S is the prior number of relevant features and T is the prior number of irrelevant features. If we denote the posterior density in (4) as δ, we can use δ to obtain a weighted splitting procedure where each impurity used at every splitting stage would be weighed by δ. For a Gini index impurity, we propose a weighted Gini index ϑ;

$$\vartheta = \sum_{k=1}^{K} (1-\delta)\widehat{p}_{mk}(1-\widehat{p}_{mk}) \tag{5}$$

where $\widehat{p}_{mk}$ is the estimated class probability at each node m. The variable with weight $\delta \to 1$, will correspond to variable with minimal unweighted Gini index and therefore useful for further splitting step. If on the other hand $\delta \to 0$, implies the variable is not useful and therefore expected to yield a maximal unweighted Gini index. In this case, the proposed weighted Gini index returns the unweighted Gini index so that the variable is dropped at the splitting stage.

4 Application to Cancer Datasets

In this section we illustrate the application of Bayesian Random Forest (BRF) on published real data. We use the "bake-off," approach of [7] to study the predictive performance comparison of BRF with competing methods on 10 different real cancer data sets. Table 1 presents the data set which is a subset of 22 datasets from package "`datamicroarray`" in R [13]. For each of the 10 data sets, we created 10 independent train/test splits by randomly selecting $9/10$ of the data as a training set and the remaining $1/10$ as a test set. Thus, $10 \times 10 = 100$ test/train splits were created (Fig. 1).

Based on each training set, each method was then used to predict the corresponding test set and evaluated on the basis of its predictive misclassification error rate and accuracy. The competing methods used alongside with BRF include Random Forest (RF), Bayesian Forest (BF), Gradient Boosting Machine

Table 1 The 10 datasets used in the bake-off and their associated dimensions

Cancer type	*n*	*p*
Colon Cancer	62	2000
Breast Cancer 1	168	2905
Lung Cancer	181	12533
Prostate Cancer	102	12600
Breast Cancer 2	49	7129
Leukemia Cancer 1	111	12625
Lymphoma Cancer	58	6817
CNS Tumor	60	7128
Myeloma Cancer	173	12625
Leukemia Cancer 2	50	10100

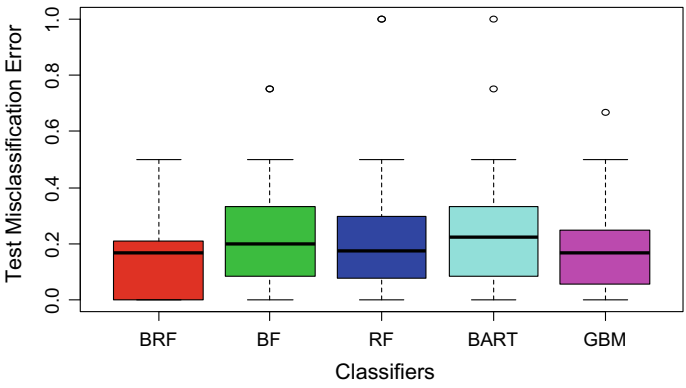

Fig. 1 Boxplot of test misclassification error rate (MER) for the five methods over 100 train/test partitions. BRF has the least MER with 25% of the MER equal zero. Also, the absence of outlying point(s) in BRF indicate that it is more stable than its competitors

(GBM) and Bayesian Additive Regression Trees (BART). Of all the five methods compared, GBM is the only frequentist method and also a major competitor of RF within the same classifier class [12].

5 Discussion and Conclusion

In this paper, we have established the weakness of RF and possible way to improve by formulating a probabilistic approach to tree sampling and split selection. We demonstrated the applicability of the method using 10 real cancer data sets. The individual and overall results in Table 2 show that in almost all the data sets used BRF accuracy is relative higher than its competitors. The result further shows that in any datasets used, BRF accuracy is bounded below at RF accuracy. This implies

Table 2 Accuracy of the methods in each and all of the 10 datasets

Cancer type	BRF	BF	RF	BART	GBM
Colon Cancer	87.14	74.76	82.38	72.62	80.71
Breast Cancer 1	79.19	73.27	75.63	76.25	76.80
Lung Cancer	99.44	98.89	99.44	99.44	97.78
Prostate Cancer	90.18	89.18	90.18	88.18	90.18
Breast Cancer 2	80.67	63.67	56.50	59.00	88.00
Leukemia Cancer 1	95.26	92.18	92.18	92.18	94.55
Lymphoma Cancer	90.36	82.86	86.61	77.86	88.21
CNS Tumor	74.33	67.05	64.81	66.81	61.48
Myeloma Cancer	80.99	79.26	79.26	79.26	78.01
Leukemia Cancer 2	79.67	66.67	70.33	68.33	69.00
Mean	85.72	78.78	79.73	77.99	82.47
SEM	11.98	15.55	17.39	16.47	14.37
25%	100.00	91.67	92.31	91.67	94.44
75%	79.44	66.67	70.44	66.67	75.00

that BRF accuracy will in most cases be higher than RF accuracy and at least RF accuracy. Therefore, it can be concluded that the Bayesian weighing scheme developed indeed correct the RF weakness.

Funding This work was supported by Universiti Tun Hussein Onn, Malaysia [grant numbers Vot, U607].

References

1. Lynch, C.: Big data: how do your data grow? Nature **455**(7209), 28–29 (2008)
2. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
3. Olaniran, O.R., Yahya, W.B.: Bayesian hypothesis testing of two normal samples using bootstrap prior technique. J. Mod. Appl. Stat. Methods **16**(2), 618–638 (2017). https://doi.org/10.22237/jmasm/1509496440
4. Olaniran, O.R., Olaniran, S.F., Yahya, W.B., Banjoko, A.W., Garba, M.K., Amusa, L.B., Gatta, N.F.: Improved Bayesian feature selection and classification methods using bootstrap prior techniques. Anale. Seria Informatică **14**(2), 46–52 (2016)
5. Olaniran, O.R., Affendi, M.A.: Bayesian analysis of extended cox model with time-varying covariates using bootstrap prior. J. Mod. Appl. Stat. Methods (2017) (in press)
6. Yahya, W.B., Olaniran, O.R., Ige, S.O.: On Bayesian conjugate normal linear regression and ordinary least square regression methods: a monte carlo study. Ilorin J. Sci. **1**(1), 216–227 (2014)
7. Chipman, H.A., George, E.I., McCulloch, R.E.: BART: Bayesian additive regression trees. Ann. Appl. Stat. **4**, 266–298 (2010)
8. Pratola, M.T.: Efficient metropolis-hastings proposal mechanisms for Bayesian regression tree models. Bayesian Anal. **11**(3), 885–911 (2016)

9. Taddy, M., Chen, C.S., Yu, J., Wyle, M.: Bayesian and empirical Bayesian forests (2015). arXiv:1502.02312
10. Efron, B.: Bootstrap methods: another look at the jackknife. In: Breakthroughs in Statistics, pp. 569–593. Springer, New York (1992)
11. Rubin, D.: The Bayesian bootstrap. Ann. Stat. **9**, 130–134 (1981)
12. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Prediction, Inference and Data Mining, 2nd edn. Springer, New York (2011)
13. Ramey, J.A.: Datamicroarray: collection of data sets for classification. https://github.com/ramhiser/datamicroarray, http://ramhiser.com (2016)