ISSN:2229-6093

IJCTA

Mohammed Zakariah, Int.J.Computer Technology & Applications,Vol 5 (5),1663-1669

Classification of genome data using Random Forest Algorithm: Review

Mohammed Zakariah
Researcher
College of Computer and Information Sciences
King Saud University
PO.Box: 51178, Riyadh: 11543
Kingdom of Saudi Arabia
Email: mzakariah@ksu.edu.sa

Abstract: *Random Forest is a popular machine learning tool for classification of large datasets. The Dataset classified with Random Forest Algorithm (RF) are correlated and the interaction between the features leads to the study of genome interaction. The review is about RF with respect to its variable selection property which reduces the large datasets into relevant samples and predicting the accuracy for the selected variable. The variables are selected among the huge datasets and then its error rate are calculated with prediction accuracy methods, when these two properties are applied then the classification of huge data becomes easy. Various variable selection and accuracy prediction methods are discussed in this review.*

Keywords: *Random Forest Algorithm, Genome datasets, Classification, Data mining, Variable Selection, Accuracy prediction.*

1. Introduction:

Biological research has largely been influenced by high-throughput genomic technologies and genome sequencing tools

including gene expression microarray, microRNA array, Single nucleotide polymorphism (SNP) array, RNA-seq (RNA Sequencing), ChIP-seq (ChIP-sequencing) but bioinformatics data analysis and statisticians face significant challenge in processing the large scale genomic data with high dimensionality and with large genomic features which a classical regression framework can no longer handle it feasibly. Since because the genomic data is highly correlated in structure it violates the assumptions required by the standard statistical models. Gene- Gene interaction is the basic mechanism in biology and also gene network which doesn't need to specify the interaction effect when it is processed in statistical model with large dimensionality. Sophisticated methodologies are required to select the important variable for high dimensional correlated and interactive genome data. Many statistical regular learning methods have been developed in recent such as penalized regression, tree based approached apart from them a boosting methods was developed to handle high-dimensional problem. The most popular ensemble method among all the learning techniques developed in the current research is Random forest (RF) with very broad applications in data mining and machine learning **[1]**. The basic idea in Random forest is to combine adaptive nearest neighbors with the bagging to have effective adaptive inference **[2]**. Random forest can deal with correlation and interaction among the variable by splitting the node with one step at a time approach and enabling the tree to impose regularization and effective analysis of "large p and small n" and "grouping property" **[3]**. Variable importance measure is an asset which enables Random Forest to select and give rank to the variables. The above mentioned points enable Random forest to be an appropriate tool for genomic data analysis and bioinformatics research. In this article, we review applications of RF to genomic data including prediction and variable selection.

2. Literature Review:

2.1 Random Forest: Random Forest is the collection of CART where each decision tree is fully grown till the terminal node and the prediction from each tree is calculated and the average of the prediction of

ISSN:2229-6093

individual tree is calculated to form the forest **[4]**. Each individual tree in the forest is grown with dataset of N cases by generating a training set of randomly selecting N times with replacement from all the N cases this is called bootstrap sample, only 2/3 of the original data is used in this bootstrap sample the remaining cases of the dataset are used for testing purpose also called out of bag which are used to estimate the OOB error for classification. OOB error estimate plays a key role in generating the prediction accuracy of the classification technique. If the no. of features per sample is 'm' then mtry are selected in random at each node (Basically RF selects two random selections first at bootstrap aggregation and then selecting the feature at random for each node) and the node is split with the best feature among the randomly selected mtry features using gini index, info gain, and node impurity splitting criteria **[5]**. The no. of mtry features selected at random are always constant in the development of the tree and the forest. Random Forest is the collection of trees but all the trees are fully grown without pruning. Each tree in the forest plays a role as a classifier which is weak and the collection of these weak forest results in significant accuracy when it is compared to the single tree classifier, because the trees in the forest are unpruned it has low-bias and high variance and averaging these unpruned ensemble of tree would result in reduced variance while keeping bias low ensemble of trees produce useful estimation of classification accuracy as discussed above and also the OOB error estimate is used to generate the importance of the feature **[6]**.

2.1.1 The following are the steps for Construction of Random Forest:

- From the Original data draw ntree bootstrap sample.
- For each bootstrap a tree is grown, select randomly mtry variables at each node to split the node, Split the node until the tree grows to the terminal node with no fewer node size.
- Information is aggregated from ntree trees and for new data prediction is done for majority of votes for classification.
- Data not in the bootstrap sample is used to calculate the OOB error.

2.1.2 Advantages of RF:

- RF if used when there are more variables than the observations.
- Multi class or more class's problem is solved by random forest.
- Even with noisy prediction variables good predictive performance is achieved and this helps in not requiring pre-selecting the genes.
- Over fitting is avoided.
- Both continuous and categorical predictors are handled.
- Predictor variables interaction is incorporated.

2.2 Variable Importance: Ranking the variables is the important feature of Random Forest; it provides a rapid computable internal measure for each variable to calculate its rank. Genomic data which is in high dimension requires this feature of ranking the variable. There are two important measure for ranking the variable node impurity indices and permutation importance. Based on the node impurity measure gini index importance is calculated in the classification. The importance of the variable is calculated by the gini index by reduction of the variable summed over all nodes for each tree in the forest which is normalized by the no. of trees. Random

ISSN:2229-6093

Forest most frequently applies Permutation importance for variable importance measure. For a given variable to estimate its importance by variable permutation method the variable is permuted randomly in the OOB data of the tree and then the permuted OOB data are dropped down from the tree and then the estimate of OOB from the prediction error is calculated. The difference between this estimate and the OOB error without permutation are averaged over all the trees to get the variable importance. The variable is more predictive if the permutation importance of the variable is larger. Genomic data is provided with modified VIMP measures used for sub sampling without replacement in place of bootstrapping has been proposed for setting where variable vary in their scale of measurement for their no. of categories**[7]**.A conditional permutation VIMP was proposed to correct bias for correlated variables **[8]**. A maximal conditional chi-square importance measure was developed to improve power to detect SNPs with interaction effects **[9]**.

2.2.1 Selection of variables and its procedure: Random Forest are capable of achieving good predictive performance with large number of predictors but finding small no. of variables and then getting equal or better prediction ability is highly desired because it is used in practical applications and also helpful for better interpretation. Diaz-Uriarte and Alvares **[10]** Selection of genes from the microarray data using RF in the backward elimination process.

2.2.2 The following are the steps in this method to select the genes:

- All the genes are fitted by the RF are randomly given a rank based on the permutation VIMP.
- All the genes are stored in the gene importance list and the RF is iteratively fitted and at each iteration a portion of the genes is removed from the bottom of the rank importance list.
- When RF reaches the smaller OOB error rate select a group of genes.
- Using .632+ bootstrap method estimates the prediction estimate rate to mitigate selection bias **[11]**.

A 10 fold cross validation was applied and at each instance when a small set of genes were found with an accurate predictor. Two software procedures were applied to implement the method with Web based tool GeneSrF(Gene Selection in Random Forest) and R-Package varSelRF(Variable selection from random forests) . Earlier than varSelRF a similar variable elimination procedure called (GSRF) **[12]** was proposed based on Random Forest. varSelRF and GSRF differ with each other in two ways First, VIMP is recomputed by GSRF after each background gene is eliminated. Second, from an independent data both OOB error rate and the prediction error rate are used to determine the best subset of genes. GSRF has some limitations for real data because it needs two datasets for implementation. Data with unbalanced samples of SNP is not appropriate to deal with classification error in VarSelRF from genome-wide association studies.

Calle et al. [13] suggested an alternative importance measure of predictive accuracy by replacing misclassified error of VerSelRF with AUC. ***Genuer et al. [14]*** has developed a new heuristic method to calculate the variable selection in RF. The basic workflow of VerSelRF was followed in this method. All the features are ranked by VIMP. Instead of removing 20% of the features at each iteration it removes all the unimportant variables in single instance by applying a threshold for minimum prediction value from CART fitting, ‘m’ important variables are kept in the beginning of the procedure. The iterative RF has now implemented it starting from the most important variable and the iteration continues increasing till all the variables are selected till ‘m’ in a uniform fashion. Based on the OOB error the final model is selected. All the above mentioned methods for variable selection are empirically performing well, but the major concern is

that all are adapting the same ranking approach and also ranking itself is a major issue than variable selection.

2.3 RF prediction: The primary goal of genomic data analysis is prediction of genes. Prediction of disease status like tumor with genomic markers. Random forest plays an important role in predicting high throughput genomic platforms and acts as important predicting tools for large datasets**. *Wu et al. [15]*** used Random forest algorithm to separate early stage ovarian cancer samples from normal tissue samples based on mass spectrometry data and further compared with other classification algorithms like Support Vector Machine (SVM), bagging and boosting classification trees, k-nearest neighbor (KNN) classifier, quadratic discriminate analysis (QDA), linear discriminate analysis (LDA), RF outperformed the other methods in terms of prediction error rate. ***Lee et al. [16]***used seven microarray gene expression datasets for classification with RF and then compared the results with the following techniques LDA (Linear Discriminant Analysis), QDA (quadratic discriminant analysis), logistic regression, (PLS) partial least square, KNN, neural network, SVM, among all these tree based techniques RF showed the best performance with five micro array gene expression datasets for survival outcomes, RFS displayed favorable results compared with supervised principal component analysis , nearest shrunken cancroids and boosting. RF and RFS are capable of accurate prediction when compared to the state of the art methods as discussed above. However, the results are encouraging but the next stage of comparative analysis for RF is theoretical nature focusing on rate of convergence. Such comparison should be done both with traditional large samples $n \to \infty$ and in setting where the features space is allowed to increase $p \to \infty$. The later study is important as the high dimensional scenario of high throughput genomic data. RF is now well known for its performance with large datasets but also if the theoretical properties are studied then it will have a deeper understanding of RF and also it would guide ways to improve it in genomic applications. Different modified versions of RF are noted which are proposed to improve the prediction performance especially for larger datasets. ***Chen et al. [17]*** proposed a new method which resulted in good prediction accuracy and interpretation, the method is called path-way based predictor instead of individual gene for cancer survival prediction using RSF. The results are based on empirical process. The ways to improve the performance of RF depends on deeper understanding of theoretical properties such as rate convergence. Biological questions are broadly answered by RF with respect to prediction, Pathway signaling and cell functions play a significant role in Protein-Protein interactions, structural biology and bioinformatics are greatly influenced with the field of PPI interaction. In the recent study it is learned that RF plays an important role in predicting PPI when compared to other methods **[18]**. Binding sites prediction from sequence annotation is another important area for structural bioinformatics. RF has been successfully applied to predict protein–DNA binding sites **[19]**, protein–RNA binding sites **[20]**, protein–protein interaction sites **[21]**, and protein–ligand binding affinity **[22]**. Based on sequence information, RF was shown as a promising tool for predicting protein functions **[23]**. MicroRNAs (miRNAs) are post-transcriptional regulators that target miRNAs for translational repression or target degradation. RF was implemented to classify real or pseudo miRNA precursors using premiRNAs like hairpins, and it achieved high specificity and sensitivity **[24]**. Glycosylation is one of the post-translational modifications (PTMs) for protein folding, transport, and function. ***Hamby and Hirst [25]*** utilized RF to predict glycosylation sites based on pair wise sequence patterns and observed improved accuracy.

Because of the large dimensionality of inherent modeling of gene-gene interaction and searching the loci in gen-gene interaction, statisticians have to impose methodological and computational challenges. Since the genome wide scans are commonly available sophisticated and powerful methods are required to handle this

huge amount of data in a feasible gene-gene interaction. The major solution for the dimensionality of the data is to remove the data by preliminary screening and select the best candidate for further analysis. A data reduction technology based on RF to improve the power of MDR. In an era with large datasets the software should be capable to handle this large datasets. MDR has been programmed to deal with data sets of 500K SNPs for 4000 subjects, but the power of MDR in this setting is not clear. The performance of MDR in large-scale studies is evaluated by calculating the proportion of simulated data sets in which MDR proposes the underlying epitasis model as the best model. As no permutation tests are run, these percentages overestimate the power of MDR and cannot be compared with our results. Prescreening the data to narrow .RF analyses are performed using Java code based on the RFs software. Software for the combined method RF couple+MDR was implemented in C++. Simulations are run on Intel Xeon X3220 2.4 Ghz processors **[26].**

The intention to develop a new technology for cell tumor and cancer classification leads to the development of gene chip. It is the process of repeated partitioning of RF trees from micro array data entry to classify cell tumor and cancer. The procedure is to form the forest of classification trees and compare the performance with extend alternatives to improve the classification and prediction accuracy. Two published datasets are used to form the deterministic forest which resembled same as random Forest and all the forests are far better than the single tree. To compare the performance of our forest constructions with random forests, individual trees, and other commonly used methods of classification and discrimination, we use two published and frequently used data sets. The first data set is on leukemia and can be downloaded at http:www-genome.wi.mit.edu_cancer. It includes 25 mRNA samples with acute myeloid leukemia (AML), 38 samples with B cell acute lymphoblastic leukemia, and 9 samples with T cell acute lymphoblastic leukemia. Expression profiles were assessed for 7,129 genes for each sample. We analyzed the data with 3,198 genes by removing the genes with at least eight missing values among all 84 samples. This lymphoma data set is available at http:llmpp.nih.gov_lymphoma **[29]**.

The major and common task in most gene expression studies for sample classification is to identify and select the most relevant genes. Researchers and scientists strive hard to detect these relevant genes which should be smaller but also giving good prediction accuracy. Microarray data classification is done with Random Forest algorithm which is well studied because of its excellent performance even when the prediction variables are noisy and also RF works well when the study is done for no. of variables greater than the no. of sample and also with the problem with more than 2 classes are required and also because of the variable importance measure. Thus the importance of Random Forest algorithm for the study of micro array data for selection of possible use of gene selection.

A new method is described for classification of microarray data for selection of gene problem based on Random Forest algorithm, Nine microarray datasets are used to classify the gene expression and compared to other classification methods including DLDA, KNN, SVM, Random Forest outperformed all the other methods yielding small sets of genes which also preserves the prediction accuracy. Because of its performance and features, random forest and gene selection using random forest should probably become part of the "standard tool-box" of methods for class prediction and gene selection with microarray data Random forest has excellent performance in classification tasks, comparable to support vector machines. All simulations and analyses were carried out with R **[27]**, using packages Random Forest (from A. Liaw and M. Wiener) for random forest. The microarray and simulated data sets are available from the supplementary material web page **[28]**.

3. Discussion and Conclusion:

Effective statistical analysis for complex and high dimensionality genomic data requires powerful and flexible statistical learning tools. Random Forest has proved to be an effective tool for classification of such complex applications. Variable selection and accuracy detection are the two most important aspects for classifying large

datasets like genome data with feature interactions, and the correlation property of RF helps in detecting the related genes and predict the accurate gene for disease and tumor. Still rigorous theoretical work is needed in RF. Improvement in developing a forest is still underway especially with small sample size and large features space settings are not fully understood and could reveal many insights to improve the forest. Theoretical analysis will focus on asymptotic rate of convergence. Theoretical analysis would result in answering the practical questions such as determining optional tuning values for RF parameters such as mtry and node size and this would help seek improvement in developing forest with improved performance. Furthermore most of the information about the data is provided by trees and forests which aren't the case with other methods for example proximity is the unique way to quantify nearness of data points in high dimensions to get the information about the near point in the high dimensionality data could be the future study. By studying the splitting behavior of the variable the interactions between the variable could be explored. Higher order interaction between the variable could be explored by higher order sub trees such analysis could be the starting point for peering inside the black-box of RF.

4. References:

[1] L. Breiman, Random forests, **Mach. Learn**. 45 (1) (2001) 5–32.

[2] L. Breiman, **Bagging predictors, Mach. Learn**. 24 (2) (1996) 123–140.

[3] H. Ishwaran, U.B. Kogalur, E.Z. Gorodeski, A.J. Minn, M.S. Lauer, **High-dimensional variable selection for survival data**, J. Am. Stat. Assoc. 105 (489) (2010) 205–217.

[4] L. Breiman, J.H. Friedman, R. Olshen, C. Stone, **Classification and Regression Trees**, Wadsworth, Belmont, Calif., 1984

[5] G. Biau, L. Devroye, G. Lugosi, **Consistency of random forests and other averaging classifiers**, J. Mach. Learn. Res. 9 (2008) 2015–2033.

[6] Y. Lin, Y. Jeon, **Random forests and adaptive nearest neighbors**, J. Am. Stat. Assoc. 101 (474) (2006) 578–590.

[7] C. Strobl, A.L. Boulesteix, A. Zeileis, T. Hothorn, **Bias in random forest variable importance measures: illustrations, sources and a solution**, BMC Bioinformatics8 (2007) 25.

[8] C. Strobl, A.L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, **Conditional variable importance for random forests**, BMC Bioinformatics 9 (2008) 307.

[9] M.H. Wang, X. Chen, H.P. Zhang, **Maximal conditional chi-square importance in random forests**, Bioinformatics 26 (6) (2010) 831–837.

[10] R. Diaz-Uriarte, S. Alvarez de Andres, **Gene selection and classification of microarray data using random forest**, BMC Bioinformatics 7 (2006) 3.

[11] B. Efron, R. Tibshirani, **Improvements on cross-validation: the .632+ bootstrap Method**, J. Am. Stat. Assoc. 92 (1997) 548–560.

[12] H. Jiang, Y. Deng, H.S. Chen, L. Tao, Q. Sha, J. Chen, C.J. Tsai, S. Zhang, **Joint analysis of two microarray gene-expression data sets to select lung adenoid carcinoma marker genes**, BMC Bioinformatics 5 (2004) 81.

[13] M.L. Calle, V. Urrea, A.L. Boulesteix, N. Malats, Auc-rf: **a new strategy for genomic profiling with random forest**, Hum. Hered. 72 (2) (2011) 121–132.

[14] R. Genuer, J.M. Poggi, C. Tuleau-Malot, **Variable selection using random forests, Pattern Recognit**.Lett. 31 (14) (2010) 2225–2236.

[15] B.L.Wu, T. Abbott, D. Fishman, W. McMurray, G.Mor, K. Stone, D.Ward, K.Williams,H.Y. Zhao, **Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data**, Bioinformatics 19 (13) (2003) 1636–1643.

[16] J.W. Lee, J.B. Lee, M. Park, S.H. Song, **An extensive comparison of recent**

classification tools applied to microarray data, Comput. Stat. Data Anal. 48 (4) (2005) 869–885.

[17] X. Chen, L.Wang, H. Ishwaran, **An integrative pathway-based clinical-genomic modelfor cancer survival prediction**, Stat. Probab. Lett. 80 (17–18) (2010) 1313–1319.

[18] N. Lin, B. Wu, R. Jansen, M. Gerstein, H. Zhao, **Information assessment on predicting protein–protein interactions**, BMC Bioinformatics 5 (2004) 154.

[19] J.S. Wu, H.D. Liu, X.Y. Duan, Y. Ding, H.T. Wu, Y.F. Bai, X. Sun, **Prediction of DNA binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature**, Bioinformatics 25 (1) (2009) 30–35.

[20] Z.P. Liu, L.Y. Wu, Y. Wang, X.S. Zhang, L. Chen, **Prediction of protein–RNA binding sites by a random forest method with combined features**, Bioinformatics 26 (13)(2010) 1616–1622.

[21] M. Sikic, S. Tomic, K. Vlahovicek, **Prediction of protein–protein interaction sites insequences and 3D structures by random forests**, PLoSComput. Biol. 5 (1) (2009)
e1000278.

[22] P.J. Ballester, J.B. Mitchell, **A machine learning approach to predicting protein– ligand binding affinity with applications to molecular docking**, Bioinformatics 26
(9) (2010) 1169–1175.

[23] K.K. Kandaswamy, K.C. Chou, T.Martinetz, S. Moller, P.N. Suganthan, S. Sridharan, G.Pugalenthi, Afp-pred: **a random forest approach for predicting antifreeze proteins from sequence-derived properties**, J. Theor. Biol. 270 (1) (2011) 56–62.

[24] P. Jiang, H. Wu, W. Wang, W. Ma, X. Sun, Z. Lu,**Mipred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features**, Nucleic Acids Res. 35 (2007) W339–W344.

[25] S.E. Hamby, J.D. Hirst, **Prediction of glycosylation sites using random forests**, BMC ,Bioinformatics 9 (2008) 500.

[26] Lizzy De Lobel, Pierre Geurts, Guy Baele, Francesc Castro-Giner, Manolis Kogevinas7 and Kristel Van Steen", **A screening methodology based on Random Forests to improve the detection of gene–gene interactions**", European Journal of Human Genetics (2010) 18, 1127–1132.

[27] R Development Core Team: **R: A language and environment for statistical computing.** 2004 [http://www.R-project.org].R Foundation for Statistical Computing, Vienna, Austria [ISBN 3-900051-00- 3].

[28]http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html].