Indian Journal of Science and Technology, Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/107971, December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Feature Selection for Microarray Data using WGCNA Based Fuzzy Forest in Map Reduce Paradigm

L. Jayasurya* and S. Krishna Anand

School of Computing, SASTRA University, Thanjavur - 613401, Tamil Nadu, India;
jayasurya.usha@yahoo.com, skanand@cse.sastra.edu

Abstract

Objectives: The feature selection is one category of principally used information analysis algorithms on microarray information or any related to high dimensional information. The goal of the feature selection algorithms is to separate out a little set of informative options that best explains experimental variations. This work really investigates the feature selection drawback for microarray information with tiny samples and variant correlation. Most existing algorithms sometimes need expensive machine effort, particularly beneath thousands of cistron (Gene) conditions. Usually citron (Gene) selection methodology searches for associate best or close to best set of genes with relevance a given analysis. **Methods:** The main objective of this project is to effectively choose the foremost informative options from microarray data, whereas creating the machine expenses reasonable. This can be achieved by proposing Fuzzy Forest using Weighted Gene Correlation Network Analysis (WGCNA) that makes use of interaction between the features (or Genes). Necessary representative features (or Genes) selection measure designated from every enriched feature partition to make the reduced gene area. **Findings:** Finally, by shaping a correct regression context, the planned methodology are often simply implement to utilize the MapReduce paradigm, that considerably reduces machine load and additionally resulting in lower prediction error rates (OOB) with variable importance compared to different existing approaches. Thus, it is necessary to know the performance of random forest with microarray data and its potential use for gene selection. **Applications:** Some the major applications of WGCNA based Fuzzy Forest are genomic data analysis (including microarray data), neuroscience, bioinformatics, DNA methylation data analysis (16s rRNA gene sequencing), cancer, yeast genetics analysis, analysis of brain imaging data (functional MRI data analysis).

Keywords: Feature Selection, Fuzzy Forest, MapReduce, Microarray Data, Module Eigen Gene, Random Forest, Rhadoop, WGCNA

1. Introduction

One of the most important current challenges to several classical classification algorithms is handling giant with large high dimensional data. Examples such as embody text data, microarray data and digital images/pictures which regularly have thousands of features and many hundreds of thousands or lots of objects. Microarray data based mostly gene expression identification has been emerging as an associate economical method for diagnosing cancer, prognosis and treatment functions[1]. Now-a-days, discriminant evaluation of microarray based data is largely used for identification of disease. In[2] provided some microarray based data defined by an out-sized variety of genes aspects, a normal discriminant evaluation build a classifier supported on provided data totally differentiate between varied illness sorts. In general, a gene selection process to pick foremost demonstrative genes (features) from the total gene types is typically utilized. Generally, a number of reasons play a role in activity gene selection/extraction. The value of clinical treatments may be decreased with gene (feature) selection owing to its lower cost to centering solely on the aspect of a couple of genes for identification rather than the total gene set. Secondly, a large number of genes within the entire gene set are unit redundant. Though the training mistakes of classi-

*Author for correspondence

fier on provided data can reduces lots and lots of genes are enclosed, the theorization error once categorizing new data eventually can increase. An old gene (feature) selection activity will take away the redundant genes, reduce storage demand and machine complexness of subsequent discriminant evaluation and probably scales back the theorization mistake. Besides, gene (feature) selection will give a lot of consolidated gene pairs that may facilitate perceive the functions of explicit genes and set up the diagnosing method. The most important demerits that persist in microarray based data reasoning is that the curse of spatial property difficulty hides the helpful information of data and results in computational imbalance[3]. A decent count of feature (gene) extraction methodology might have been projected by different researchers and scholars previously. This sort of methods concentrates on procedure of compression, wherever the initial dataset gets revised. It is troublesome to process a task employing a typical system having some standard processing quality. Numerous machine learning techniques are projected within the space of bioinformatics (typically it is microarray data) by completely contrary in[4–8]. However these take up plenty of time to processing and explore massive datasets. To table this, the plan of distributed computing has been followed, wherever the data can be distributed on numerous nodes in a various cluster and processed using varied parallel processing environment[9]. The map reduce programming framework and its enforcement on Hadoop paradigm encompasses a considerable base for evaluating massive datasets, especially for large dimensional genomic based data like microarray data, in a distributed manner. Apache Hadoop[10] is free open source software and an efficient method of accumulating and process large data in a distributed mode on massive clusters of trade goods hardware. It state a master/slave design for both divided up storage and distributed evaluation, thus, carried out 2 jobs namely large data backup and quicker processing. This work, proposing a novel way for selecting top ranked features by using the new algorithm based on random forests called Fuzzy Forest, which is mainly designed to reduce the bias focused in random forests selection due to the presence of correlated features. This actually uses the recursive feature elimination, to select the important features from separate modules of correlated features, whenever the correlation within each module of features is high and the correlation between each module is less. Here, each module can be split and constructed with fuzzy measures using the Weighted Gene Correlation Network (WGCNA), which is one of data mining techniques used to perform correlation network synthesis for bulk high dimensional data. The feature selection analysis by using random forest can be parallelized using the Map reduce framework with pseudo distribution environment. It becomes possible to achieve better prediction accuracy for huge highly correlated microarray dataset using variable importance measures (MSE) and computational time for analysis is reduced by using parallel environment.

The remaining section of this paper is sorted out as follows. Related work is discussed in Section 2. The proposed model is formulated in Section 3. Experimental results are examined in Section 4. Conclusions are provided in Section 5.

Random forest is the best approach which makes the classification decisions by selecting results of each and every decision tree. Generally, good generalization accuracy with an ensemble learner have two major properties, they are: 1. High accuracy of individual component learner and 2. Component learners have high diversity. This is mainly producing classifiers from the random samples of training data, unlike the other methods like bagging and boosting creating the standard classifiers from randomly selected subspaces of each data. In[11] proposed an improved random forest for feature weighting techniques, which uses a novel feature weighting method for the selection of subspace that improves the performance of classification on high dimensional data. The information gain ratio or x^2-test is used for the calculation of weights of feature. In[12] discussed a hierarchical sampling strategy by using these weights to choose the feature subspaces for random forest in classification complications. In[13] proposed a new clustering method which mainly used stratified idea. However, in suggested the random forest implementation which is mainly based on the binary classification setting and it uses splitting criteria for the linear discriminant analysis. Despite the favorable aspects, this method is not found to be efficient on high dimensional datasets with several classes. In order to get over this drawback, in[14] proposed the feature weighting method for sampling subspace to deal with data like microarray. Besides, variance analysis has been mainly utilized to calculate weights for features. This method might be compared with the method of Adaboost which was proposed by[15] in which the training samples to

be selected reported to example weights that were figured out from the output of the past classification. Later to this, in[16] proposed a method by using PCI and decision forest methodology, selecting the features for sentiment analysis, which is validated using twitter dataset to increase prediction of classifiers. In[17] proposed a new procedure which slightly differs from previous methods for partition a subset of features. Here, all uninformative features (noise) will be removed from the existing system and the better feature set, that is largely related to response feature found using method like statistical. In the later part of the year[18,19], proposed a novel approach for evaluating correlation and to optimize storage and calculation of correlation an effective algorithm based on MapReduce. This algorithm is generally used the standard method for optimizing correlation measure for large throughput microarray dataset. The existence of a massive number of inconsiderable and non related features (genes) decreases the quality of computation of illness such as 'cancer'. For this reason, in[20] proposed that the proper tests namely Friedman test, Kruskal–Wallis test, ANOVA test based on map reduce was declared to choosing related variables from microarray dataset. To assort this microarray dataset map reduce based proximal Support Vector Machine classifier has been declared. The four slave (data) nodes in Hadoop clusters and a formal system were used to test the performance of the algorithms. The best performing machine learning algorithms such as random forests classifiers, that are amongst the ensembles of decision trees and have been well employed for genetic diverse prioritization in case-control studies. In order to generate rankings of genes in associated studies along with multi variable quantitative traits RFs is applied and to validate genetic law of similarity measures which are mainly used as prognostic of the trait. For this, proposed Parallel Random Forest Regression (PARFR) which takes the merits of MapReduce programming model and applied to the study of genome wide association on Alzheimer's disease. The proposed method by them applied to the quantitative data that really comprise of high dimensional neuro imaging phenotype mainly describes modifications in human brain structure longitudinally and also gives SNPs ranking, which is associated to this data. In[21,22] discussed that by using two methods statistical and informative based methods for selection of genes where the score of gene relevance for each gene is evaluated without considering correlation of genes. Our proposed WGCNA based Fuzzy Forest algorithm in MapReduce environment falls into second category. Some of irrelevant features accommodate with relevant features (genes), which severely impact the accuracy of machine learning. Hence, feature subset selection algorithm should have high capability to determine and remove necessary amount of the irrelevant and redundant data/information as much as possible. Moreover, in general "new feature subsets includes highly correlated features with (predictive of) the class, thus far uncorrelated features with (not predictive of) one another". With this view in perspective, here the proposed work is to develop a novel algorithm that could effectively and efficiently handle with features of both impertinent and redundant features and gain a well feature subset by using the weighted network analysis with fuzzy measures to detect the modules from a correlated network. Based on these selection features, random forest classification/ Regression ensemble learning method is used to obtain the predictions for new data and besides, it will be helpful for additional experimental variations.

2. Proposed Models

The proposed methodology is WGCNA based Fuzzy Forest in MapReduce framework to handle the unbalanced data and to select the top features from each gene module. Thereby, it is possible to eliminate the unimportant features from each module and select top features from them. The random forests are fit for selected features, which can be used to obtain predictions for new data and it will be helpful for additional experimental variations. By this work, it is possible to achieve better prediction accuracy for huge highly correlated microarray dataset using variable importance measures (MSE) and computational time for analysis is reduced by using parallel environment. Initially, Data analysis is used to collect usable information from a data source such as websites, database etc. in data analysis the collected information should be associated to specific domain. The dataset for the reasoning, which perform as necessary input to models is acquired from UCLA human genetics dataset repository and National Center of Biotechnology Information (NCBI, GEO: http://www.ncbi.nlm.nih.gov/geo/). The microarray dataset used here includes F2 mouse intercross gene expression details with clinical traits and gene annotations, having high correlation between each gene modules and also to illustrate the idea of regression analysis.

2.1 WGCNA based Analysis

Coefficient of Correlation networks are progressively getting utilized in biological information applications. For instance, weighted gene correlation network analysis may be a systems biological methodology for depicting the correlation structure among features across microarray gene samples. Weighted Gene Correlation Network Analysis (WGCNA) is used locating clusters of extremely correlative features, for iterating clusters using the module Eigengene methodology, for connecting modules to at least each other and to external gene sample traits (by use of method module Eigengene) and for scheming module membership measures. WGCNA is broadly used data mining methodology especially in the area of biological networks study based on the pairwise correlation between each features. Generally, it is used for high dimensional dataset, in particular genomic applications. The WGCNA includes the various function used for finding the correlation between the gene modules, which in turn used to eliminate the irrelevant features and taking care for relevant features in each modules. These functions of WGCNA are illustrated in the form of flowchart using Figure 1. Initially the gene correlation network has been constructed, for network nodes corresponds to genes(features) and relation capabilities which are depicted by pairwise correlations between expression data file. In distinction to unweighted networks, the weighted networks utilize soft thresholding of Pearson correlation matrix for depicting the affiliation capabilities between two genes (features). To build weighted correlation network, gene co-expression similarity measure is utilized to connect every gene-gene relationship. The adjacency matrix is then built by utilizing soft power adjacency function as:

$$c_{ij} = Power\left(a_{ij}, \beta\right) \equiv \left(a_{ij}\right.^{\beta} \quad (1)$$

Where, a_{ij} explains the correlation similarity and Cij stands for the adjacency output which calculates the connection capabilities. The power β is selected by using the scale free topology criterion. Table 1 describes that the first column listing power beta, second column provides the scale free topology index results, next column gives the slope of line, fourth column gives about the truncated exponential scale free model, the left over columns gives the mean and median connectivity of network. At once the network has been created; the logical step is often module identification. Modules can be described as clusters of closely interconnected genes. Module finding is supported on the topological overlap matrix (TOM),

$\varphi = \left[\rho_{ij}\right]$, where ρij is symmetric and lies between 0 and 1. To utilize this in hierarchical clustering, it will be changed to dissimilarity standard by deducting this from 1 like,

$$d_{ij}^{\rho} = 1 - \rho_{ij} \quad (2)$$

Based on this dissimilarity measure, Figure 2 explains that the hierarchical clustering can be utilized to differentiate one module from other one. To find biological based or clinical based important modules and genes is a main task for lot of correlation network analysis. In organic improvement analysis, a gene importance measure can able to point out path (way) membership. The module significance measurement can be defined as a mean gene (feature) significance across the module genes. In that module, genes with large interconnectivity be given having massive independent co-efficient of correlation with body weight of mice (GSweigh). For a provided physiological trait data, the one can able to describe an evaluation of gene importance by shaping implicit measure of coefficient of correlation of both traits and gene expression values. The bodyweight of mice will be useful to describe a gene significance of with gene expression as:

$$GSweigh(i) = \left(cor\left(y(i), weigh\right)\right(\quad (3)$$

Where, y (i) is gene expression data of gene i. This is mainly used to evaluate the coefficient of correlation between external samples and the mice expression data and it can be identified by using the colors for each module in the network. Then, if each module is detected with high correlated traits, then one can easily select a representative trait by picking the gene (trait), which is nearest to each module called Eigengene. The Eigengene for a given module will be described as first principal component of weighted standard expression profile. It will condense each module into one domain (or profile). Find the modules that are related to clinical trait of interest by using Eigengene. The adjacency between the Eigengene and the sample gene data is referred as Eigengene significance. The weighted Eigengene correlation network can be described as:

$$c_{SR} = \left(cor\left(EG^{(S)}, EG^{(R)}\right)\right)^{\beta} \quad (4)$$

Where, $EG^{(S)}$ and $EG^{(R)}$ denotes the two distinct modules Eigengenes. Here the module Eigengenes calculate 18 module from the given set and the highly correlated modules with body weight are represented as brown (331 genes), blue (483 genes) and red (223 genes). In few applications it should be profitable to outline endless fuzzy evaluation of module membership for each and every node. Such calculation is significantly helpful to establish nodes that lie close to the limit of a module, or nodes which are in between two or a lot of modules. The module membership of node i in a module S can be described as:

$$K_{cor,i}^{(s)} = cor\left(y(i), EG^{(S)}\right) \quad (5)$$

Where, y (i) represents salience of node i, $EG^{(S)}$ denotes module Eigengene of module S. The module membership calculation $K_{cor,i}^{(s)}$ which lies in between [-1, 1] and determines how much the node i is near to module s, s=1,….,S. $K_{cor,i}^{(s)}$ is mentioned as module Eigengene related connection measurement KME.

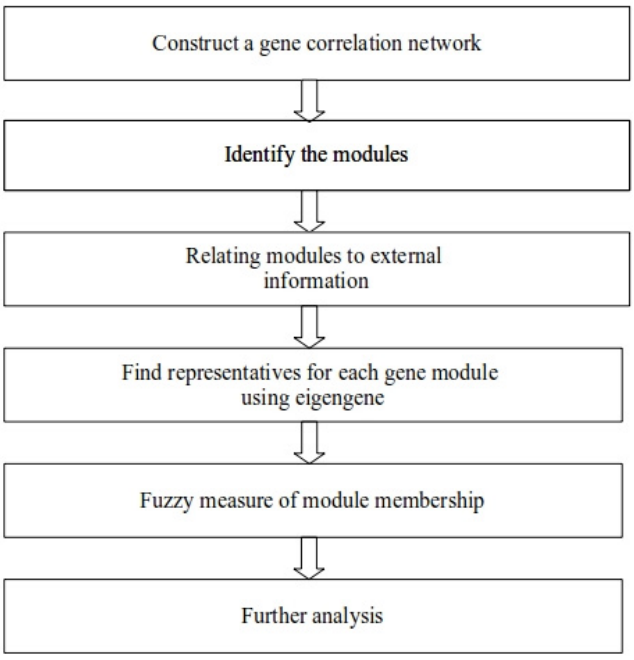

Figure 1. Overview of WGCNA.

2.2 Fuzzy Forest based Error Rate Prediction

Fuzzy Forest is an expansion of random forests which is planned to return low biased variable importance ranking, whenever there is high correlation among the features (variables). Fuzzy Forest also permits for ease integration along with the WGCNA using the function "wff". Whenever the function wff is called, WGCNA with some parameters used to divide the variables into different modules, thereby the clusters are roughly uncorrelated with each other. Based on previous discussion, the Fuzzy Forest is the extension of random forest and in this point of view random forest constructs an ensemble method of regression tree and everyone will be severally learned on boot-strapped method. The random forest projects that forest error rate depend on two factors:

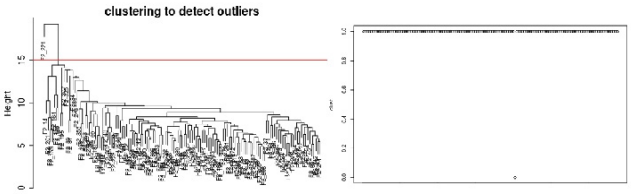

Figure 2. Clustering output to detect outliers and sample clusters plot.

- The relativity between any of two trees in forest and by maximizing the coefficient of correlation will increment of error rate of forest.
- The power of every separate tree in forest. If the tree is having less error rate, then it is a strong classifier. Improving the power of trees reduces error rate of forest. To get an unbiased calculation of classification/regression error as trees and which are connect to the forest, by using OOB (Out-Of-Bag) error prediction method. This is utilized to get an estimation of error prediction for its tree and also overall estimation. It is also used to obtain the variable importance estimation. The Out-Of-Bag error rate is essential to choose an adequate huge amount of trees to undertake best performance and constant ranking. Here the OOB error rate predicted by initializing n tree = 5000 and the number of iterations is 5, for the Substance BxH of F2 intercross expression profile with clinical trait body weight to find the correlated genes from each module. The correlation coefficient between the two error based on Substance BxH is 0.9518 describes using the Table 2. By this, some of the genes having the high variable importance with the OOB (MSE) for error prediction.

Table 1. Results of soft threshold for connectivity of network

x	Power	soft.R.sq	Slope	truncated.R.sq	mean.k.	median.k.	max.k.
1	1	0.0278	0.345	0.458	748.01	763.0000	1211.0
2	2	0.1260	-0.557	0.844	255.00	251.1000	573.0
3	3	0.3400	-1.031	0.973	111.00	103.4300	323.0
4	4	0.5066	-1.426	0.973	56.54	48.2400	212.0
5	5	0.6810	-1.722	0.941	32.22	26.1100	135.0
6	6	0.9020	-1.501	0.962	19.91	13.5000	94.8
7	7	0.9210	-1.677	0.918	13.24	8.6800	84.1
8	8	0.9041	-1.724	0.878	10.25	5.3900	76.35
9	9	0.8590	-1.724	0.836	6.88	3.5600	70.51
10	10	0.8330	-1.668	0.831	5.19	2.3800	65.8

2.3 Hadoop Implementation for Feature Selection

Fuzzy Forest implementation is like the normal random forest implementation and builds trees in sequential manner. But, the sequential approach is greatly in efficient, in particular whenever each and every tree consists an enormous count of variables and lot of trees are necessary in order to get effective measures of variable importance and based on the error rate prediction (MSE). Here the parallel portion of Fuzzy Forest enforced using MapReduce model for deploying huge Hadoop clusters. In the proposed methodology the Fuzzy Forest first screens out unimportant from each of modules via feature elimination method. For this, it has some of tuning parameters such as mtry factor, n tree factor, drop fraction, keep fraction, node size and finally for the selection of features using number selected. Based on this the important top ranked features can be selected with variable importance and module membership. The MapReduce model in the Hadoop environment consists of 3 forms, the map form, shuffle form and reduce form. Every map and reduce phases have <key, value> pairs as an input and output. The shuffle part recombining the result of map part to input for reduce phase equally by using rmr2 library. The input file is generally stored in the HDFS (Hadoop Distributed File System) and then that file can be used for further process. The map phase here having every input key fit to specific tree ID and their value is NULL, till loading of entire data to build trees. The output of map phase includes; 1. The sample identifier (key) and important feature list based on OOB factor with module membership values, 2. Variable identifier (key) and important variables with mean-square-error (value), which can be used to get the top ranked features with variable importance gain at reduce part, 3. Example pair identifier and different features (values), that is utilized to match with the feature list to produce important features. All kind of these output from each mappers will be shuffled sorted and send to reducer by Hadoop. Hadoop task first distributes the large dataset to each map task using the distributed cache mechanism. In this implementation work, every map job loads the boot-strapped version of dataset for further analysis. The reducer job is to compute the top ranked features for mice expression profile with clinical trait as body weight, which all are having high correlation between them based on the variable importance score.

3. Experimental Results

For intent of estimating the performance and effectiveness of discussed WGCNA based Fuzzy Forest algorithm, validating whether or not the strategy is possibly useful in preparation, letting some other researchers to confirm the results, the F2 mouse intercross gene expression data set is used. This dataset consists of Female mice liver expression profile has 3600 genes (features) with 143 variables, which describes the physiological and metabolic traits of F2 mouse intercross gene expression details also it has clinical trait data as gene trait data, which has 361 observations with 27 variables contains related information

of mice expression profile and the gene annotation data, which contains related gene information of selected probe and has 23388 genes with 34 variables. The dataset used here is having high correlation between each gene modules and also to illustrate the idea of regression analysis.

In order to do the best usage of data and get steady results, initially module Eigengene method is defined to find the representatives of gene expression profile for each module. Following this, by correlate the gene based expression salience with module Eigengene of a provided module; it is possible to express a "fuzzy" determination of membership for each gene based on module. Using these methods based on the WGCNA allows to partition the covariates into distinct modules, such that the clusters are approximately uncorrelated with one another. Figure 3 and Figure 4 shows the results of fuzzy membership values for each gene (feature) based modules with some color traits. In this module Eigengenes calculate 18 module from the given set and the highly correlated modules with body weight are represented as brown (331 genes), green (483 genes) and red (221 genes). The module membership of green color trait with gene significance for body weight (GS_Weigh) of mice has high correlation of (cor = 0.21) and p = 5.2e-05. Based on the fuzzy membership for each gene in a module, Fuzzy Forest is applied, which mainly planned to return low biased variable importance ranking, whenever there is high correlation among the features (variables). Fuzzy Forest also permits for ease integration along with the WGCNA using the function "wff". To get an unbiased calculation of classification/regression error as trees and which are connect to the forest, by using OOB (Out-Of-Bag) error prediction method. This was utilized to get an estimation of error prediction for its tree and also overall estimation. It is also used to obtain the variable importance estimation. Here the OOB error rate predicted by initializing n tree = 5000 and the number of iterations is 5 (Mtry), for the Substance BxH of F2 intercross expression profile with clinical trait body weight to find the correlated genes from each module. The correlation coefficient between the two error based on Substance BxH is 0.9518 describes using the Figure 5. By this, some of the genes is having the high variable importance with the OOB (MSE) for error prediction. Further the parallel portion of Fuzzy Forest enforced using MapReduce based model for deploying huge Hadoop clusters. The MapReduce model in the Hadoop environment consists of 3 phases, the map part, shuffle part and reduce part. The map part here having every input key fit to specific tree ID and their value is NULL, till loading of entire data to build trees. The output of map phase includes; 1. The sample identifier (key) and important feature list based on OOB factor with module membership values, 2. Variable identifier (key) and important variables with mean-square-error (value), which can be used to get the top ranked features with variable importance gain at reduce phase, 3. Example pair identifier (key) and different features (values) that are utilized to match with the feature list to produce important features. All kind of this output from each mapper will shuffle, sorted and send to reducer by Hadoop. The reducer job is to compute the top ranked features for mice expression profile with clinical trait as body weight, which all are having high correlation between them based on the variable importance score which describes in Figure 6. The parallel Fuzzy Forest analysis is implemented in R language with the help of R Hadoop. It could run the jobs in a standalone, pseudo distributed mode. By using parallel implementation, the accuracy with 97% for the feature with their corresponding module membership is achieved by taking less computational time.

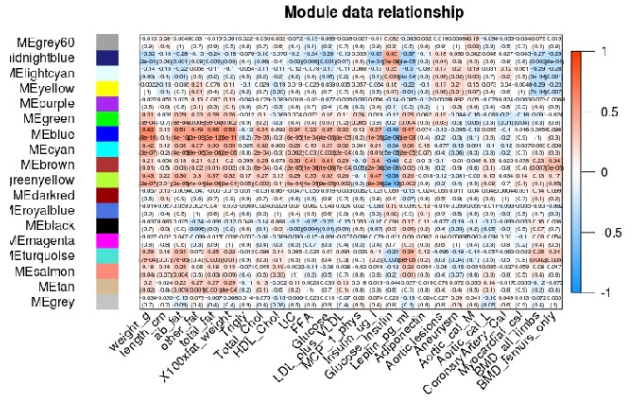

Figure 3. Module trait relationship with fuzzy module membership.

Table 2. Fuzzy Forest analysis of error rate prediction (OOB)

Features	%Inc MSE
MMT00067008	0.99151687
MMT00051244	0.94596869
MMT00030931	0.90860944
MMT00192541	0.90004320
MMT00065159	0.86805021
MMT00058021	0.73982567

MMT00043149	0.71762960
MMT00061509	0.71758201

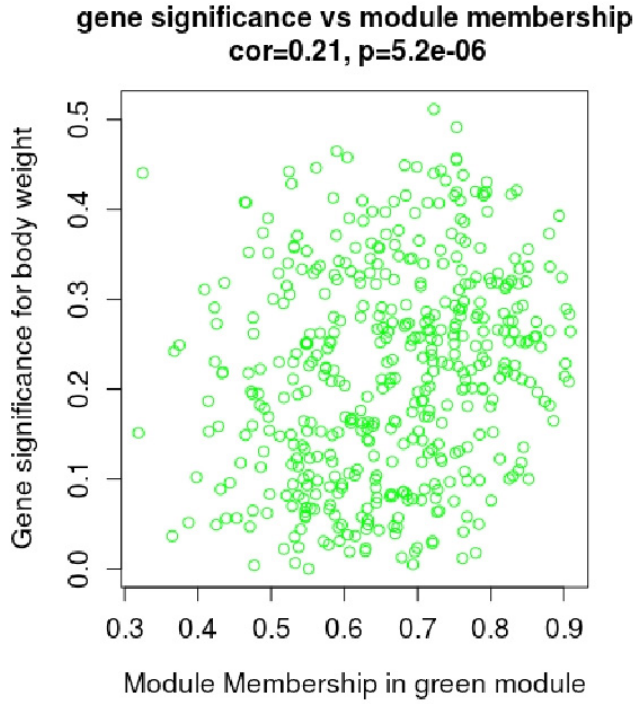

Figure 4. Module membership with gene significance.

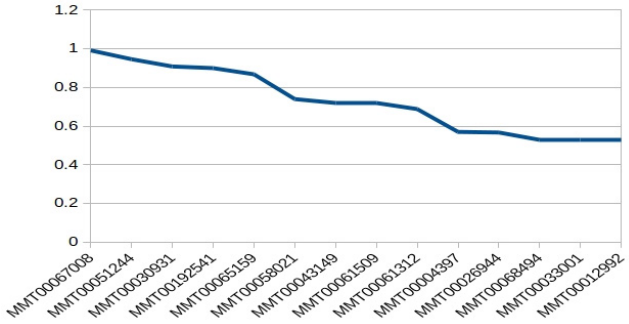

Figure 5. Fuzzy Forest analysis of error rate prediction (OOB).

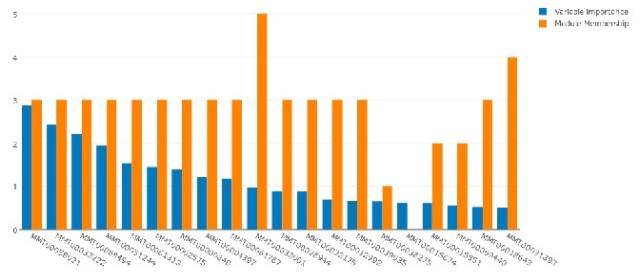

Figure 6. Top ranked selected genes with variable importance using MapReduce.

4. Conclusion

In this work, feature (gene) selection for microarray dataset using WGCNA based Fuzzy Forest in MapReduce paradigm is proposed to enhance the feature (gene) selection method by selecting important features from separate modules of correlated features, whenever the correlation within each module of features is large and the correlation between each modules is less along with variable importance and fuzzy measures for module membership. The results demonstrate the feasibility of proposed methodology for gene selection problem when using large correlated dataset. The proposed methodology for feature selection helps biologists to detect various types of diseases whenever having the large set of cell structure with their clinical traits and annotations (have high correlation with each other). By using the MapReduce paradigm the results can be obtained quickly thereby the feature selection can be efficient for further analysis. Further, this piece of work can be stretched by viewing some of the relevant machine learning approaches like Logistic regression, SVM, Naive Bayes and other similar methods using Map reduce environment in fully distributed mode with master and slave nodes.

5. References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science. 1999 Oct; 286(5439):531–7.
2. Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd C, Pohl U, Hartmann C, McLaughlin ME, Batchelor TT, Black PM, Von Deimling A, Pomeroy SL, Golub TR, Louis DN. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Research. 2003 Apr; 63(7):1602–7.
3. Lee G, Rodriguez C, Modabhusi A. Investigating the efficacy of non-linear dimensionality reduction schemes in classifying gene and protein expression studies. IEEE/ACM Transactions on Computational Biology Bioinformatics. 2008 Jul-Sep; 5(3):368–84.
4. Lee KE, Sha N, Dougherty ER, Vannuia M, Mallick BK. Gene selection: A bayesian variable selection approach. Bioinformatics. 2003 Jan; 19(1):90–7.
5. Peng Y, Li W, Liu Y. A hybrid approach for biomarter discovery from microarray gene expression data for cancer classification. Cancer Informatics. 2007 Feb; 2:301–11.

6. Wang L, Chu F, Xie W. Accurate cancer classification using expressions of very few genes. IEEE/ACM Transactions on Computational Biology Bioinformatics. 2007 Jan-Mar; 4(1):40–53.
7. Deb K, Raji Reddy A. Reliable classification of two-class cancer data using evolutionary algorithms. Bios Systems. 2003 Nov; 72(1-2):111–29.
8. Hemandez JCH, Duval B, Hao JK. A genetic embedded approach for gene selection and classification of microarray data. Springer Berlin Heidelberg; 2007 Apr. p. 90–101.
9. Map-Reduce for machine learning on multicore. 2009. Available from: https://www.cs.duke.edu/courses/spring09/cps296.3/lectures/07-mapreduce-ml-multicore.pdf
10. Apache Hadoop. 2016. Available from: https://en.wikipedia.org/wiki/Apache_Hadoop
11. Xu B, Huang JZ, Williams G, Wang Q, Ye Y. Classifying very high-dimensional data with random forests built from small subspaces. International Journal of Data Warehousing and Mining. 2012; 8(2):1–20.
12. Ye Y. Stratified sampling for feature subspace selection in random forests for high dimensional data. Pattern Recognition. 2013 Mar; 46(3):769–87.
13. Chen X, Ye Y, Xu X, Huang JZ. A feature group weighting method for subspace clustering of high-dimensional data. Pattern Recognition. 2012 Jan; 45(1):434–46.
14. Amaratunga D, Cabrera J, Lee YS. Enriched random forests. Bioinformatics. 2008 Jul; 24(18):2010–4.
15. Qiu D, Ye Wang. Identify cross-selling opportunities via hybrid classifier. International Journal of Data Warehousing and Mining. 2008; 4(2):1–8.
16. Islam AT, Jeong BS, Bari AG, Lim CG, Jeon SH. Map reduces based parallel gene selection method. Applied Intelligence. 2015 Mar; 42(2):147–56.
17. Nguyen TT, Huang JZ. Unbiased feature selection in learning random forests for high dimensional data. Scientific World Journal. 2015; 2015:1–18.
18. Wang S. Optimising parallel r correlation matrix calculations on gene expression data using MapReduce. BMC Bioinformatics. 2014 Nov; 15(1):1–9.
19. Ludwig SA. MapReduce-based fuzzy c-means clustering algorithm: Implementation and scalability. International Journal of Machine Learning and Cybernetics. 2015 Dec; 6(6):923–34.
20. Kumar M, Kumar Rath S. Classification of microarray data using kernal fuzzy inference system. International Scholarly Research Notices. 2014 Aug; 2014:1–18.
21. Jotheeswaran J, Koteeswaran S. Feature selection using random forest method for sentiment analysis. Indian Journal of Science and Technology. 2016 Jan; 9(3):1–7.
22. Das K, Ray J, Mishra D. Gene selection using information theory and statistical approach. Indian Journal of Science and Technology. 2015 Apr; 8(8):695–701.