ISSN XXXX XXXX © 2016 IJESC

Research Article Volume 6 Issue No. 12

Microarray Gene Expression Data Classification with Random Forest

Süleyman Aytekin[1], B. Sıddık Yarman[2], İnci Zaim Gökbay[3]
Department of Biomedical Engineering, Namık Kemal University, Tekirdağ, Turkey[1]
Department of Electrical-Electronics Engineering, Istanbul University, Istanbul, Turkey[2]
Informatics Department, Istanbul University, Istanbul, Turkey[3]

Abstract:
Cancer diagnosis is a major clinical topics of gene expression microarray technology. We are researching to improve system for cancer diagnostic model based on microarray data. We studied an area of classification algorithms, gene selection method, and cross-validation using 9 different Tumor datasets. Multi-Category Support Vector Machine was found to be the better method than other learning algorithms such as K-Nearest Neighbors and Neural Networks. Gene selection techniques are shown that classification performance significantly increases. So many researchers analiyze how to select a small number of informative genes from thousand of genes. In this paper, we observed that Random Forest algorithm with gene selection algorithm which is called CfsSubSetEval performs successfully than other algorithms which are used numerously in the literature.

Keywords: Microarray gene expression, random forest, gene selection

I. INTRODUCTION

Microarray gene expression are becoming so important for clinical decision support especially prediction of clinical results of cancer and other diseases. Microarray technology has the potential to provide accurate and objective cancer diagnosis due to its high capability of measuring expression levels of tens of thousands genes simultaneously. Researchers are seeking to improve and apply the best classification algorithms to maximize benefits of this technology. They also have tried to analiyze thousands of genes simultaneously to obtain significant information about specific cellular functions of gene(s) which can be used in cancer diagnosis [1]. A necessary prerequisite for the creation of clinically successful microarray-based diagnostic models is a solid understanding of the relative strengths and weaknesses of avaliable classification and methods.Prior research suggest that support vector machines (SVMs), k-nearest neighbors, back propagation neural network, probabilistic neural networks, weighted voting methods and decision trees are significantly outperforming. The comprehensive study in [2] shows that SVM can outperform K-nearest neighbors and neural network in gene expression cancer diagnosis. The gene selection from microarray gene expression data is very hard due to high dimension. Choose the significant subset of genes with hig classification accuracy is needed. Such methods are not enough for doctors to identify a small subset of biologically related genes for cancers[3]. In the application of microarray data, how to select small number of informative genes from thousands of genes that may ensure to the occurence of cancers is a significant issue. To achieve powerful gene selection from thousands of genes that can ensure in identifying cancers. In feature selection method, the best related genes are choosen in the space of all feautre subset and other genes are ignored [4]. Statistical methods employs to collect the distinctive characteristic of genes in discriminating the tarhgeted class [5-7]. Genetic algorithms (GAs) [8] are usually used as the search engine for feature subset selection and such estimation of distrubition algorithm with support vector machine (SVM) [9-16], K nearest neighbors/genetic algorithms (KNN/GA) [17,18], genetic algorithms-support vector machine (GA-SVM) [19], neural networks [20,21], nearest shrunken centroids [22], logistic regression [23] are used. In the recent years, random forest algorithm [24] is more popular within the bioinformatics community for classification of microarray gene expression data [25-27]. The random forest algorithm has a number of attractive properties making it well fro classification of microarray expression data. It is appropriate when there are so many predictors than observations. It is based on ensemble learning that allows the algorithm to learn accurately both simple and complex classificaiton functions and it is appropriate for both binary and multicategory classification tasks. In this paper, we analyzed microarray gene selection with and classificaiton with random forest algorithm and compare random forest algorithm success with SVM, K-NN and navie bayes methods which are often used in the numerous research. Following second part express methodology of our study and third part result of experience and the last one is conclusion.

II. METHODS

A. *Microarray datasets*

Among 9 datasets used in [28], we choose dificult one which make SVMs generate “bad” classification performance.

- 9 Tumors [29]: the dataset comes from a study of 9 human tumor types: NSCLC, colon, breast ovary, leukemia, renal, Melanoma, prostate, and CNS. There are 60 samples, each of which contains 5726 genes.

The dataset are avaliable on the website of GEMS in the format of either GEMS or MATLAB mat file.The gene expression data are normalized by being rescaled to between 0 and 1. It is also for the aim of speeding up the training of classification algorithm.

B. *Attribute evaluater*

We used CfsSubSetEval which is default Attribute Evaluator in Weka. It evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subset of features that are highly correlated with the class while having low intercorrelation are preferred [30]. This method evaluates the worth of a subset of descriptors by considering the individual predictive ability of each one along with the degree of redundancy between the descriptors. Subsets of descriptors that are highly correlated with the property/activity values and having low intercorrelation are preferred. Fort this problem, BestFirst (search method) and CfsSubSetEval (attribute evaluater) combination is as best variable selection techniques-genetic algorithm or simulated annealing – but it is much quicker. This is why these default settings were selected for application.

- An attribute subset is good if the attributes it contains are
  - Highly correlated with the class attribute
  - Not strongly correlated with one another

- Goodness of an attribute subset (1);

$$\frac{\sum_{all\ attributes\ x} C(x, class)}{\sqrt{\sum_{all\ attributes\ x}\sum_{all\ attributes\ y} C(x, y)}} \quad (1)$$

- C measures the correlation between two attributes

- An entropy-based metric called the "symmetric uncertainty" is used

C. *Cross-validation design*

We used 9-fold cross-validation to estimate the performance of the classification algorithms. For four classification methods, we obtained cross-validation performance. We built a classification model with the best

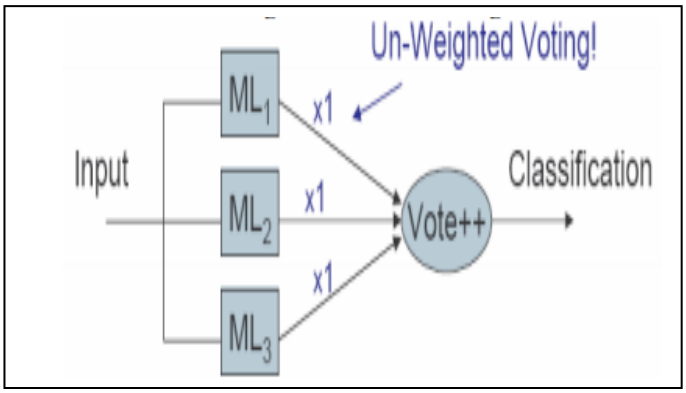

Figure.1. Meta learners [32]

parameters on the training set and applied this model to the testing set.

D. *Random forest classifier*

Random forests (RF) is a classification algorithm that uses an ensemble of unpruned decision trees, each of which is built on a bootstrap sample of the training data using randomly selected subset of variables [24]. It is a class for constructing a forest of random trees. This algorithm has a number of properties making it an appealing technique for classification of microarray gene expression data. The random forest machine learner, is a meta learner; meaning consisting of many individual learners (trees). The random forest uses multiple random trees classifications to votes on an overall classification for the given set of inputs. In general in each individual machine learner vote is given equal weight. In Breiman's later work, this algorithm was modified to perform both un-weighted and weighted voting. The forest chooses the individual classification that contains the most votes. Fig. 1 below is a visual reoresentation of the un-weighted random forest algorithm [31]. Individual random tree machine learners are grown in the following manner [31]:

1. A dataset is formed by sampling with replacement members from training set.
2. A random number of attributes are chosen for each tree. These attributes form the nodes and leafs using standard tree building algorithms
3. Each tree is grown to the fullest extent possible without pruning.This process is repeated to develop multiple individual random trees learners. After the development of the tree, the out-of-bag examples are used to test the individual's trees as well as the entire forest.

III. RESULT

We used random forest algorithm for classification and compare it with support vector machine (SVM), k-nearest neighbor (K-NN) and Navie Bayes classification algorithm. We also used CfsSubSetEval algorithm for gene selection and 58 genes in 5726 are selected. There are 60 samples for 9 different tumor types. Table-I shows the classification results of the full dataset training. All algorithms are successful with %100 and K-NN is the fastest algorithm. Table-II shows the

TABLE.1.FULL DATASET TRAINING

Method	***Accuracy (%)***	***Time (s)***
Random Forest	100.0	1.93
SVM	100.0	0.55
K-NN	100.0	**0.52**
Navie Bayes	100.0	3.81

TABLE.II. 9-FOLD CROSS-VALIDATION WITH GENE SELECTION

Method	***Accuracy (%)***	***Time (s)***	***Selected gene***
Random Forest	**79.31**	0.48	58
SVM	74.14	< 0.72	
K-NN	69.0	< 0	
Navie Bayes	74.14	< 0	

TABLE.III. 9-FOLD CROSS-VALIDATION WITHOUT GENE SELECTION

Method	***Accuracy (%)***	***Time (s)***
Random Forest	46.55	2.15
SVM	**62.07**	0.51
K-NN	41.37	< 0
Navie Bayes	43.1	0.34

Classification results with 9-fold cross-validation under gene selection. Random forest algorithm has the most accurate result with 79.31 than other classification methods. Table – III shows the classification results applying 9-fold cross-validation without gene selection. SVM is the best with %62.07 for microarray gene expression data classification.

IV. CONCLUSION

Based on statistical analysis, random forest (RF) algorithm with gene selection outperforms other popular classifier. It was successfully applied to cancer gene expression data.

V. REFERENCES

[1] Alba E, et al: Gene selection in cancer classification using PSO/SVM and GA/SVM hybrid algorithms. IEEE C Evol Computat. 2007, 9: 284-290.

[2] Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S: A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 2005, 21: 631–643.

[3] Li S, Wu X, Tan M: Gene selection using hybrid particle swarm optimization and genetic algorithm. Soft Comput. 2008, 12: 1039-1048. 10.1007/s00500-007-0272-x.

[4] Ahmad A, Dey L: A feature selection technique for classificatory analysis. Pattern Recogn Lett. 2005, 26: 43-56. 10.1016/j.patrec.2004.08.015.

[5] Su Y, Murali TM, et al: RankGene: identification of diagnostic genes based on expression data. Bioinformatics. 2003, 19: 1578-1579. 10.1093/bioinformatics/btg179.

[6] Kahavi R, John GH: Wrapper for feature subset selection. Artif Intell. 1997, 97: 273-324. 10.1016/S0004-3702(97)00043-X.

[7] Li X, Rao S, Wang Y, Gong B: Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. Nucleic Acids Res. 2004, 32: 2685-2694. 10.1093/nar/gkh563.

[8] Zhao XM, Cheung YM, Huang DS: A novel approach to extracting features from motif content and protein composition for protein sequence classification. Neural Netw. 2005, 18: 1019-1028. 10.1016/j.neunet.2005.07.002.

[9] Brown MP, et al: Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A. 2000, 97: 262-267. 10.1073/pnas.97.1.262.

[10] Evers L, Messow CM: Sparse kernel methods for high-dimensional survival data. Bioinformatics. 2008, 24: 1632-1638. 10.1093/bioinformatics/btn253.

[11] Hua S, Sun Z: A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol. 2001, 308: 397-407. 10.1006/jmbi.2001.4580.

[12] Oh JH, Gao J: A kernel-based approach for detecting outliers of high-dimensional biological data. BMC Bioinforma. 2009, 10: S7-

[13] Saeys Y, et al: Feature selection for splice site prediction: a new method using EDA-based feature ranking. BMC Bioinforma. 2004, 5: 64-10.1186/1471-2105-5-64.

[14] Zhu Y, Shen X, Pan W: Network-based support vector machine for classification of microarray samples. BMC Bioinforma. 2009, 10: S21-

[15] A. I. Su, et al, "Molecular classification of human carcinomas by use of gene expression signatures", Cancer Res., Vol. 61, pp. 7388-7393, 2001.

[16] S. Ramaswamy, et al, "Multiclass cancer diagnosis using tumor gene expression signatures", Proc. Natl Acad. Sci. USA, Vol. 98, pp. 15149-15154, 2001.

[17] Li L, Darden TA, et al: Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. Comb Chem High T Scr. 2001, 4: 727-739.

[18] S. L. Pomeroy, et al, "Prediction of central nervous system embryonal tumour outcome based on gene expression", Nature, Vol. 415, No. 6870, pp. 436-442, 2002.

[19] Li L, Jiang W, et al: A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics. 2005, 85: 16-23. 10.1016/j.ygeno.2004.09.007.

[20] J. Khan, et al, "Classification and diagnosis prediction of cancers using gene expression profiling and artificial neural networks", Nat. Med., Vol 7, pp. 673-679,2001.

[21] D. Berrar, et al, "Multiclass cancer classification using gene expression profiling and probabilistic neural networks", In Proceedings of the Pacific Symposium on Bio computing (PSB), Lihue, Hawaii, January 3, 2003.

[22] R. Tibshirani, et al, "Diagnosis of multiple cancer types by shrunken centroids of gene expression", Proc. Natl Acad. Sci. USA, Vol. 99, pp. 6567-6572, 2002.

[23] J. Zhu and T. Hastie, "Classification of gene microarrays by penalized logistic expression", Biostatistics, Vol. 5, pp.427-443,2004.

[24] Breiman L: Random forests. *Machine Learning* 2001, 45: 5–32.

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

[26] Lee JW, Lee JB, Park M, Song SH: An extensive comparison of recent classification tools applied to microarray data.*Computational Statistics & Data Analysis* 2005, 48: 869–885.

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

[28] A. Statnikov, et al, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis", Bioinformatics, vol 5, pp. 631-643, 2005

[29] J. E. Staunton, et al, "Chemosensitivity prediction by transcriptional profiling", Proc. Natl Acad. Sci. USA, Vol. 98, pp. 10787- 10792, 2001.

[30] M. A. Hall (1998). Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand.

[31] Frederick Livingston, Machine Learning Journal Paper, 2005

[32] M. White, ECE591Q-Machine Learning-Lecture slides, 2005

Süleyman A. Aytekin B.Sc. in Electronics and Communication Engineering, Kocaeli University, Kocaeli, Turkey, 2011; M.Sc. in Electronics and Communication Engineering, Yıldız Technical University (YTU), İstanbul, Turkey, 2015; Ph.D. candidate in Biomedical Engineering, Istanbul University, Istanbul, Turkey. Research Assistant, Biomedical Engineering, Corlu Engineering Faculty, Namık Kemal University. Research in computational bioinformatics, biomedical signal and speech processing, brain computer interfaces.

İnci Zaim Gökbay Assistant Professor, Informatics Department, Istanbul Universitt

B. Sıddık Yarman B.Sc. in Electrical Engineering (EE), Istanbul Technical University (I.T.U.), Istanbul, Turkey, February 1974; M.E.E.E in Electro-Math Stevens Institute of Technology (S.I.T.) Hoboken, NJ, June 1977; Ph.D. in EE-Math Cornell University, Ithaca, NY, January 1982. Member of the Technical Staff (MTS) at Microwave Technology Centre, RCA David Sarnoff Research Center, Princeton, NJ (1982–1984). Associate Professor, Anadolu University, Eskişehir, Turkey, and Middle East Technical University, Ankara, Turkey (1985–1987). Visiting Professor and Research Fellow of Alexander Von Humboldt, Ruhr University, Bochum, Germany (1987–1994). Founding Technical Director and Vice President of STFA Defense Electronic Corp., Istanbul, Turkey (1986–1996). Full Professor, Chair of Division of Electronics, Chair of Defense Electronics, Director of Technology and Science School, Istanbul University (1990–1996). Founding President of Işık University, Istanbul, Turkey (1996–2004). Chief Advisor in Charge of Electronic and Technical Security Affairs to the Prime Ministry Office of Turkey (1996–2000). Chairman of the Science Commission in charge of the development of the Turkish Rail Road Systems of Ministry of Transportation (2004). Young Turkish Scientist Award, National Research Council of Turkey (NRCT) (1986). Technology Award of Husamettin Tugac Foundation of NRCT (1987). International Man of the Year in Science and Technology, Cambridge Biography Center of UK (1998). Member Academy of Science of New York (1994), Fellow of IEEE (2004). Four US patents (1985–1986), More than 100 technical papers, technical reports in the field of Design of Matching Networks and Microwave Amplifiers, Mathematical Models for any Systems, Speech and Biomedical Signal Processing (since 1982). Prof. Yarman has been back to Istanbul University since October 2004.