Microarray Gene Expression Data Classification with Random Forest
Süleyman Aytekin¹, B. Siddık Yarman², İnci Zaim Gökbay³
Department of Biomedical Engineering, Namık Kemal University, Tekirdağ, Turkey¹
Department of Electrical-Electronics Engineering, Istanbul University, Istanbul, Turkey²
Informatics Department, Istanbul University, Istanbul, Turkey³

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 analyze 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 numerous 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 analyze 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 available 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 high 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 occurrence 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 chosen in the space of all feature subset and other genes are ignored [4]. Statistical methods employs to collect the distinctive characteristic of genes in discriminating the targetted class [5-7]. Genetic algorithms (GAs) [8] are usually used as the search engine for feature subset selection and such estimation of distribution 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 for 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 classification functions and it is appropriate for both binary and multiclassification tasks. In this paper, we analyzed microarray gene selection with and classification 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 difficult 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 available 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 evaluator

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 evaluator) 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);

\[
\sum_{all\; attributes} C(x, class) / \sqrt{\sum_{all\; attributes} \sum_{all\; attributes} C(x, y)}
\]

(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 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 representation 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 leaves 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.I. 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   |     |
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


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


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

İnci Zaim Gökbay Assistant Professor, Informatics Department, Istanbul University