A ROBUST MODEL FOR GENE ANLYSIS AND CLASSIFICATION

Fatemeh Aminzadeh, Bita Shadgar[1], Alireza Osareh

Department of Computer Engineering, Shahid Chamran University, Ahvaz, Iran
{f.aminzadeh, bita.shadgar, a.osareh}@scu.ac.ir

ABSTRACT

The development of microarray gene technology has provided a large volume of data to many fields. Microarray data analysis and classification has demonstrated an effective methodology for the effective diagnosis of diseases and cancers. Although much research has been performed on applying machine learning techniques for microarray data classification during the past years, it has been shown that conventional machine learning techniques have intrinsic drawbacks in achieving accurate and robust classifications. So it is more desirable to make a decision by combining the results of various expert classifiers rather than by depending on the result of only one classifier. We address the microarray dataset based cancer classification using a newly proposed ensemble classifier generation technique referred to as RotBoost, which is constructed by combining Rotation Forest and AdaBoost. The experiments conducted with 8 microarray datasets, among which a classification tree is adopted as the base learning algorithm, demonstrate that RotBoost can generate ensemble classifiers with significantly lower prediction error than either Rotation Forest or AdaBoost more often than the reverse.

KEYWORDS

Microarray Analysis, Ensembles, Cancer Classification, Rotation Forest

1. INTRODUCTION

Microarray technology has provided the ability to measure the expression levels of thousands of genes simultaneously in a single experiment. This scheme provides an effective experimental protocol for gaining insight into the cellular mechanism and the nature of complex biological process. Microarray data analysis has been developing at fast speed in recent years and has become a popular and standard way in most current genomics research works [1]

Each spot on a microarray chip contains the clone of a gene from a tissue sample. Some mRNA samples are labelled with two different kinds of dyes, for example, Cy5 (red) and Cy3 (blue). After each mRNA interacts with the genes, i.e., hybridization, the color of each spot on the chip will change. Then, the resulted image reflects the characteristics of the tissue at the molecular level [2-3].

However, the amount of data in each microarray is too overwhelming for manual analysis, since a single sample often contains measurements for around 10,000 genes. Due to this excessive amount of information, efficiently produced results require automatic computer controlled analysis of data. Many computational tools have been applied to mine through this huge amount of gene expression data to discover biologically meaningful knowledge.

One of the main important computational groups for analysis of microarray data is machine learning-based approaches. Machine learning techniques have been successfully applied to cancer classification problem using gene microarray data [4].

[1] Corresponding author

DOI : 10.5121/ijma.2011.3102

As more and more gene microarray datasets become publicly available, developing technologies for analysis of such data becomes an essential task [5, 6]. Having said that, so far various machine learning and pattern recognition methods are increasingly utilized, for instance, discriminant analysis [7], neural networks [8], and support vector machines [9–11]. A considerable amount of researches involving microarray data analysis are focused on cancer classification, aiming at classifying test cancer samples into known classes with the help of a training set containing samples whose classes are known [12]. To tackle this issue, several methods based on gene expression data have been suggested. Some of these are applicable only to binary classification, such as the weighted voting scheme of Golub et al. [13], whereas others can handle multiple classification problems. These approaches range from traditional methods, such as Fisher's linear and quadratic discriminant analysis, to more modern machine learning techniques, such as classification trees or aggregation of classifiers by bagging or boosting (for a review see [14]) [12]. There are also approaches which are able to identify test samples that do not belong to any of the known classes by imposing thresholds on the prediction strength [13, 15].

Despite these progresses, gene microarray cancer classification has remained a great challenge to computer scientists. The main challenges lie in the nature of microarray data, which is mostly high-dimensional and noisy. Natural biological instabilities are very likely to import measurement variations and bring implications to microarray analysis [4]. This makes learning from microarray data a difficult task especially under the effect of curse of dimensionality. Indeed, gene expression data often contains many irrelevant and redundant features, which in turn can affect the efficiency of most machine learning techniques.

There is therefore a great requirement to build up robust methods that are able to overcome the limitation of the small number of microarray input instances and reduce the influence of uncertainties so as to produce reliable classification (cancerous/non-cancerous) results. In most cases, one single classification model may not lead to high classification accuracy. Instead, multiple classifier systems (ensemble learning methods) have proved to be an effective way to increase prediction accuracy and the robustness of a learning system [16].

Although the application of multiple classifier systems (MCS) to microarray dataset classification is still a new field, recently some different MCSs have been proposed to deal with the gene microarray data classification problem. For example, Dettling et al. [17] used a revised boosting algorithm for tumor classification, Ramon et al. [18] applied Random Forest to tackle both gene selection and classification problems simultaneously, and Peng [19] designed a SVM ensemble system for microarray dataset prediction.

These techniques generally work by means of firstly generating an ensemble of base classifiers by applying a given base learning algorithm to different alternative training sets, and then the outputs from each ensemble member are combined in a suitable way to create the prediction of the ensemble classifier. The combination is often performed by voting for the most popular class. Examples of these techniques include Bagging, AdaBoost, Rotation Forest and Random Forest [20].

AdaBoost technique creates a mixture of classifiers by applying a given base learning algorithm to successive derived training sets that are formed by either resampling from the original training set or reweighting the original training set according to a set of weights maintained over the training set [20]. Initially, the weights assigned to each training instance are set to be equal and in subsequent iterations, these weights are adjusted so that the weight of the instances misclassified by the previously trained classifiers is increased whereas that of the correctly classified ones is decreased. Thus, AdaBoost attempts to produce new classifiers that are able to better predict the ''hard" instances for the previous ensemble members.

The main idea of Rotation Forest is to provide diversity and accuracy within an ensemble classifier. One possible way to promote diversity can be achieved by a principal component analysis (PCA) based feature extraction for each base classifier. Indeed, the accuracy is sought by keeping all principal components and also using the whole data set to train each base classifier. In view of the fact that both AdaBoost and Rotation Forest are successful ensemble classifier generation techniques and they apply a given base learning algorithm to the permutated training sets to construct their ensemble members with the only difference lying in the ways to perturb the original training set, it is plausible that a combination of the two may achieve even lower prediction error than either of them.

In [21] an ensemble-based technique called RotBoost which is constructed by integrating the ideas of Rotation Forest and AdaBoost is proposed. According to this study, RotBoost was found to perform much better than Bagging and MultiBoost on the utilized benchmark UCI datasets. Here, we inspired from RotBoost technique and apply it for the first time on 8 publically available gene microarray benchmark data sets. Indeed, we present a comparative study of RotBoost results with several ensemble and single classifier systems including AdaBoost, Rotation Forest, Bagging and single tree. Experimental results revealed that the RotBoost ensemble method (with several basis classifiers) perform best among the considered classification procedures and thus produces the highest recognition rate on the benchmark datasets.

The rest of this paper is organized as follows. Section 2 our proposed algorithm for an efficient classification of gene microarray data. Section 3 presents experimental results against 8 publically available benchmark gene microarray datasets. Finally, Section 4 concludes this study.

2. MATERIAL AND METHOD

This paper proposes an approach for the construction of accurate and diverse ensemble members by means of learning from the best sub-sets of initial microarray genes. The method proposed in this paper is comprised of three main stages, i.e. feature selection based on fast correlation based filter, ensemble classifier generation method using a combination of Rotation Forest and AdaBoost algorithms and evaluating the generalisation ability of various ensemble/non-ensemble classifier systems. The details of these stages are discussed in the following sections.

2.1. Datasets

In this work, we utilized 8 publicly available benchmark datasets [22]. A brief overview of these datasets is summarized in Table 1. Data pre-processing is an important step for handling gene expression data. This includes two steps: filling missing values and normalization. For both training and test dataset, missing values are filled using the average value of that gene. Normalization is then carried out so that every observed gene expression has mean equal to 0 and variance equal to 1. In summary, the 8 datasets had between 2–5 distinct diagnostic categories, 60–253 instances (samples) and 2000–24481 genes after the data preparatory steps outlined above.

Table 1. Description of 8 gene microarray datasets.

Dataset	# Total Genes (*T*)	# Instances (*n*)	# Classes (*C*)
Colon Tumor	2000	62	2
Central Nervous System (CNS)	7129	60	2
Leukaemia	6817	72	2
Breast Cancer	24481	97	2
Lung Cancer	12533	181	5
Ovarian Cancer	15154	253	2
MLL	12582	72	3
SRBCT	2308	83	4

2.2. RotBoost Ensemble Technique

As it was stated before, Rotboost is constructed by integrating the ideas of Rotation Forest and AdaBoost ensemble classifier generation techniques with the aim of achieving even lower prediction error than either of these individual techniques. Consider a training set *L* which is defined as follows:

$$L = \{(x_i, y_i)\}_{i=1}^{N} \quad (1)$$

Assume that the above training set consisting of *N* independent instances, in which each sample (x_i, y_i) is described by an input attribute vector x_i as follows:

$$x_i = (x_{1i}, x_{2i}, \ldots, x_{id}) \in R^d \quad (2)$$

and a class label y_i which takes value from the label space φ = {1, 2, …, *k*}. Now, in a typical classification problem, the goal is to use the information only from *L* to construct classifiers that have good generalization ability, namely, perform well on the previously unseen test data which are not used for learning the classifiers.

For simplicity of the notations, let *X* be an *N* x *d* matrix composed of the values of *d* input attributes for each training instance and *Y* be an *N*-dimensional column vector containing the outputs of each training instance in *L*. Alternatively, *L* can be expressed as concatenating *X* and *Y* horizontally, that is, *L* = [*X Y*]. Now, we can show the base classifiers which are included into an ensemble classifier, say, *C** by C_1, C_2, . . . , C_T [20]. Indeed, let $E = (X_1, X_2, \ldots, X_d)^T$ be the attribute set composed of *d* input attributes. Before starting on proposing the RotBoost algorithm, we briefly review the ensemble methods AdaBoost and Rotation Forest as follows.

AdaBoost [23] is a sequential algorithm in which each new classifier is built by taking into account the performance of the previously generated classifiers.
In this ensemble method, a set of weights $w_t(i)$ (i = 1, 2, ...*N*) are maintained over the original training set *L*. The weights initially set to be equal (namely, all training instances have the same importance). In subsequent iterations, these weights are adjusted so that the weight of the instances misclassified by the previously trained classifiers is increased whereas that of the correctly classified ones is decreased. In this way, the difficult input samples can be better predicted by the next trained classifiers [20].

In AdaBoost, the training set L_t utilized for learning each base classifier C_t is acquired by either resampling from the original training set *L* or reweighting the original training set *L* according to the updated probability distribution w_t maintained over *L*. Here, the resampling scheme is applied as it has less complexity for implementation. Indeed, each base classifier C_t is assigned to a weight in the training phase and the final decision of the ensemble classifier is obtained by weighted voting of the outputs from each ensemble member.

Rotation Forest is another proposed ensemble classifier generation method [24] in which the training set for each base classifier is constructed by incorporating PCA to rotate the original feature axes. On the other hand, in order to create the training data for a base classifier, the feature set E is randomly split into K subspaces and then PCA is applied to each of these subspaces. To retain the variability information in the data all principal components are preserved. Thus, K axis rotations take place to form the new attributes for a base classifier.

The main idea of Rotation Forest is to simultaneously preserve individual accuracy and diversity within the ensemble individual base classifiers. To be more specific, diversity is promoted through doing feature extraction for each base classifier and accuracy is obtained by keeping all principal components and also using the whole data set to train each base classifier.

The detailed steps of Rotation Forest are described in [20]. It has been already pointed out by many researchers that [24], for an ensemble classifier to achieve much better generalization capability than its component members, it is essential that the ensemble classifier consists of highly accurate base members which at the same time disagree as much as possible. It has also been noted by [25] that the prediction accuracy of an ensemble classifier can be further improved on condition that the diversity of the ensemble members is increased whereas their individual errors are not affected.

When employing the above proposed RotBoost algorithm to solve a classification task, some parameters required to be defined beforehand. As with the most ensemble methods, the values of the parameters S and T that, respectively, specify the numbers of iterations done for Rotation Forest and AdaBoost should be fine tuned by the user and the value of K (or M which represents the number attributes in each subspace) can be selected to be a moderate value according to the size of the feature set E. Since the good performance of an ensemble method largely depends on the instability of the used base learning algorithm [26], the base classifier can be therefore generally chosen to be either a decision tree or an artificial neural network [27] which is instable in the sense that small variations in its training data can lead to large changes in the constructed decision boundary. Here, we utilized decision trees as the individual base classifiers of the final constructed RotBoost ensemble predictor.

2.3. Gene Selection

As it was already pointed out, generalisation ability of the RotBoost ensemble model can be highly affected by the presence of thousands of genes many of which are unnecessary from the classification point of view. Thus, if RotBoost applied to classify a typical microarray dataset, a rotation matrix with thousands of dimensions is required for each tree, which this in turn requires a very high computational complexity. As only a small subset of genes are of interest in practice, therefore, a key issue of microarray data classification based on RotBoost ensemble model is to accomplish an efficient dimension reduction process to identify the smallest possible set of genes that can achieve good predictive accuracy.

Two broad categories of optimal feature subset selection have been proposed: filter and wrapper. In filter approaches, features are scored and ranked based on certain statistical criteria and the features with highest ranking values are selected. Frequently used filter methods include t-test, chi-square test, mutual information, Pearson correlation coefficients and PCA [28].

In contrast, in wrapper approaches, feature selection is “wrapped” in a learning algorithm. The learning algorithm is applied to subsets of features and tested on a hold-out set, and prediction accuracy is used to determine the feature set quality. Since exhaustive search is not computationally feasible, wrapper methods must employ a search algorithm to search for an optimal subset of features.

In this work, we consider fast correlation-based filter (FCBF) method which has been successfully used for gene selection and demonstrated to attain promising performance [29]. FCBF, is a fast correlation-based filter method which begins by selecting a subset of relevant features whose *C*-correlation are larger than a given threshold γ, and then sorts the relevant features in descending order in terms of *C*-correlation. Using the sorted feature list, redundant features are eliminated one-by-one in a descending order. A feature is redundant only if it has an approximate Markov blanket [4]. The remaining feature subset thus contains the predominant features with zero redundant features in terms of *C*-correlation**.**

3. EXPERIMENTAL RESULTS

We apply the RotBoost method to eight well-known cancer datasets described in section 2.1. The corresponding classification task is to classify the normal and tumor samples. The datasets are first pre-processed and then to reduce the computational complexity and select the most informative genes, FCBF is applied separately to these datasets. Table 2 summarized the identification numbers (IDs) of those genes selected by FCBF method.

Table 3 shows the number of genes which are selected by FCBF for each individual microarray gene datasets. As it can be seen, the number of selected genes is different and depends on the processed gene dataset.

Table 2. The IDs of the genes selected by FCBF method.

Data Set	IDs of selected genes with FCBF Method
Colon Tumor	1671,765,625,1423,1772,1042,1153,1635,1900,279,576,682,1328,1560
CNS	2474,7016,5507,2996,5528,612,2032,400,1971,2735,1320,6810,2089,2404,11,2142,3113,4509,18,844,360,3420,6485,4484,2695,3185,2426,2202
Leukaemia	1834,4847,1882,3252,2288,6855,1685,6376,2354,4373,4366,758,1829,2128,2020,1779,1926,1674,2111,538,2497,5501,1630,7119,4951,2441,1239,1904,4438,1087,683,4190,4664,6277,3172,3482,1120,4232,2517,6169,5376,2733,4898,5984,4342,4593,620,6184,2626,412,1924
Breast Cancer	3463,377,8782,1889,8910,7448,10889,17595,15102,15906,2663,19856,16616,10643,275,18109,9445,12553,12429,11536,5861,1505,21304,21545,20866,7814,13800,2882,12520,20341,18820,6757,20317,1126,571,7081,7509,14532,3524,20342,1229,23161,1355,4248,644,2713,14374,15635,3697,15387,1007,5393,23207,10876,462,5280,4583,24107,21818,14991,719,18767,6592,15813,11853,18539,2583,12259,11182,7295,4351,216,5052,10997,56,14447,22612,5984,7790,20891,3190,8074,7655,17787,4618,16894,590,12572,24298,407
Lung Cancer	3191,9038,10188,10891,10175,5533,7568,8890,12052,1682,4983,9770,10138,9093,11300,3120,5292,2870,3875,11942,8294,4282,8457,9609,2383,9470,9311,8745,7361,7298,9170,10426,6422,9134,4115,11468,9937,4525,8683,12021,6949,4733,6174,12375,8199,8786,6897,9897,6513,8429,6796,10787,8762,3692,5108,10128,6620,9989,6185,7087,4321,6814,9910,5407,10862,7124,192,11646,6944,4473,3776,4397,1776,4772,4943,6319,5561,4778,2919,7905,7328,7721,6784,4879,6304,7162,1591,3104,10847,8533,10376,5600,9357,11841,12283,3321,4061,5619,4693,8157
Ovarian Cancer	1679,2237,1684,1736,1677,6782,545,2528,182,1733,1823,2666,5534,1702,187,2306,13170,8840,7508,13978,1494,6499,7781,13261,360,14,63,2199,2233,10408
MLL	2592,11297,5370,7666,6067,8428,9741,10457,7136,8165,11366,8423,3882,4602,5460,5801,10797,7232,7716,9668,6718,3399,9845,6416,5083,9478,9929,4745,10454,7961,4804,8370,3675,6413,6337,6294,1176,4660,7930,10274,8455,11282,7155,7070,7946,10530,5772,8050,8281,12174,6615,3804,1259,11044,6830,9085,4347,7131,317,9716,8769,4431,3361,4050,8711,445,5591,4892,12355,7007,11884,8725,7489,3869,4276,9153,5961,338,1587,3712,6809,10496,3806,4346,9870,9952,11635,10363,1876,5549,8919,3306,12031,11851,659,3330,3793
SRBCT	1601,742,1003,1389,509,1708,2050,1645,2162,1613,153,1194,1980,417,1884,256,1434,380,1662,2198,1700,251,2303,1536,1795,1207,867,1655,1158,1159,1673,2168,368,667,365,2199,1112,326,2230,1489,2159,1105,819,558,1888,799,1208,607,1768,188,2186,373,2301,1479,774,454,156,733,2235,2049,1760,1210,1942,1634,67,672,490,979,823,1924,3,1120,437,2000,117,1775,314,1829,1962,159,1464,746

Table 3. Number of genes selected for each dataset by FCBF gene selection method.

Datasets	Initial Gene Numbers	FCBF Selected Numbers
Colon	2000	14
CNS	7129	28
Leukaemia	7129	51
Breast	24481	90
Lung	12553	100
Ovarian	15154	30
MLL	12582	97
SRBCT	2308	82

The experimental settings are as follows. In all the ensemble methods, a classification tree [30] is always adopted as the base learning algorithm because it is sensitive to the changes in its training data and can still be very accurate. The parameters included in classification tree algorithm, such as the number of training instances that impure nodes to be split must have, are all set to be the default values of the Weka Toolbox. In order to provide a fair comparison, for all other utilized techniques such as Rotation Forest, AdaBoost and Bagging, 100 trees are trained to constitute the corresponding ensemble classifiers. With respect to RotBoost, the number of iterations done for Rotation Forest and AdaBoost are both chosen to be $S = T = 10$ (to balance the trade-off between these two algorithms) so that an ensemble classifier created by it also consists of 100 trees. As for the parameter M (namely, the number of attributes contained in each attribute subset) included in RotForest and RotBoost, the optimum value is experimentally found to be 3.

In many earlier works, researchers typically split the original dataset into two parts i.e. a training set and a test set in a random fashion. Gene selection is then performed on the training set and the goodness of selected genes is assessed from the unseen test set [4]. However, due to the small number of instances in gene microarray datasets, such an approach is now recognized by the community as unreliable. Instead, Ambroise and McLachlan [31] suggested splitting the data using 10-fold cross validation or 0.632+bootstrap. Indeed, a comparative study of several different error estimation techniques on microarray classification [32] also suggests that 0.632+ bootstrap may be more appropriate than other estimators including re-substitution estimator, cross-validation, and leave-one-out estimation.

Thus, in this work we employed a balanced 0.632+bootstrap technique to evaluate the performance of the gene selection algorithm considered in this study. The .632+bootstrap requires sampling a training set with a replacement manner from the original dataset. The test set is then made by those samples excluded from the training dataset. Finally, the 0.632+bootstrap is repeated n times and the final bootstrap error is estimated as follows:

$$E = {}^{1}\!/_{n} \sum_{i=1}^{n} \left(0.368\alpha_i + 0.632\beta_i\right) \tag{3}$$

where α_i and β_i are the training error and test error on the ith resampling stage. Following the work in [19], Here, the bootstrap samples are experimentally formed with $n = 15$ replicates. Therefore, each sample in the original dataset is made to appear exactly 15 times in the balanced bootstrap training samples. It is worth to note that, the feature selection is then performed using only the training samples. Finally, the test error (classification accuracy) is estimated on the unseen test samples using Equation (3).

Table 3 shows the means of classification accuracy for each classification method on the considered datasets, where the values following ‘‘±” are their respective standard deviations. In order to see whether RotBoost is significantly better or worse than other ensemble/non-ensemble methods including single tree, Rotation Forest, AdaBoost and Bagging from the statistical viewpoint, a one-tailed paired t-test was performed with significance level $\alpha = 0.05$ and the results for which a significant difference with RotBoost was found are marked with a bullet or an open circle next to them. A bullet next to a result indicates that RotBoost is significantly better than the corresponding method. An open circle next to a result denotes that RotBoost performs significantly worse than the corresponding method. In the triplet labeled ‘‘Win–Tie–Loss” in the last row of Table 3, the first value is the number of data sets on which RotBoost performs significantly better than the corresponding algorithm; the second one is the number of data sets on which the difference between the performance of RotBoost and that of the corresponding algorithm is not significant; and the third one denotes the number of data sets on which RotBoost behaves significantly worse than the compared algorithm.

Table 4. Means of classification accuracy for each classification method against 8 different gene microarray datasets. ‘‘•”specifys that RotBoost is significantly better and ‘‘^{o}” points out that RotBoost is notably worse at the significance level $\alpha = 0.05$.

Dataset	**RotBoost**	**Single Tree**	**Rotation Forest**	**AdaBoost**	**Bagging**
Colon	95.48±0.61	93.80±0.82•	95.21±0.43	94.97±0.63•	94.92±0.50•
CNS	94.80±0.59	89.92±0.61•	92.37±0.83•	95.09±0.64	93.50±0.79•
Leukemia	98.75±0.31	96.60±00.46•	97.97±0.38•	98.22±0.55•	97.47±0.51•
Breast	94.39±0.49	88.50±0.72•	92.60±0.63•	94.89±0.47^{o}	92.74±0.45•
Lung	98.11±0.17	94.36±0.42•	97.56±0.23•	98.08±0.39	97.08±0.37•
Ovarian	99.82±0.08	99.37±0.12•	99.77±0.07•	99.57±0.11•	99.36±0.08•
MLL	98.86±0.23	96.03±0.59•	97.61±0.31•	97.63±0.45•	97.08±0.55•
SRBCT	99.50±0.31	93.96±0.59•	97.44±0.41•	98.16±0.39•	96.46±0.58•
Win tie loss		8/0/0	7/1/0	5/2/1	8/0/0

As can be seen from Table 4, RotBoost performs significantly better than both Single Tree and Bagging algorithms. When compared with Rotation Forest, the statistically significant difference is favorable in 7 datasets, and tie has been occurred against colon dataset. Indeed, RotBoost is seen to outperform AdaBoost in most cases although the advantage of RotBoost is not significant in 1 set and tie is occurred for the remaining 2 datasets.

4. CONCLUSIONS

In this paper, we applied RotBoost ensemble technique to tackle the microarray data classification problem. This ensemble classifier generation method is a combination of Rotation Forest and AdaBoost techniques which in turn preserve both desirable features of an ensemble architecture i.e. diversity and accuracy. Here, we utilized decision tree classifiers as our base learners. To cope with curse of dimensionality of gene microarray datasets, FCBF filter method is first employed to select a small subset of most informative genes. Then, the RotBoost was operated on the selected gene subsets.

To evaluate the effectiveness of RotBoost algorithm other ensemble/non-ensemble techniques including single tree, Rotation Forest, AdaBoost and Bagging were also considered and their performances compared against RotBoost. The experimental results show that Rotboost ensemble with several basis classifiers is a robust method for microarray classification, which achieved the highest accuracy for majority of the analysed benchmark datasets.

In fact, RotBoost is found to perform much better than the other examined counterparts. By the way, the improvement of generalization ability achieved by RotBoost is obtained with negligible increase in computational costs. Indeed, RotBoost provides a potential computational benefit over AdaBoost in that it can be executed in a parallel manner.

REFERENCES

[1] C. Shang, Q. Shen, (2005) “Aiding Classification of Gene Expression Data with Feature Selection”, *International Journal of Computational Intelligence Research*, Vol. 1, No. 1, pp 68-76.

[2] H. Lodish, A. Berk , S.L. Zipersky, P. Matsudaira, D. Baltimore, J.E. Darnell, (2003) *Molecular Cell Biology*, Demos Medical Publishers.

[3] A. Lehninger, M. Co, D. Nelson, (2000) *Principals of Biochemistry*, Worth Publishers.

[4] Z. Zexuan, Y. Ong, M. Dash, (2007) “Markov Blanket-Embedded Genetic Algorithm for Gene Selection”, *Pattern Recognition*, Vol. 49, No. 11, pp 3236-3248.

[5] E. Dougherty, A. Datta, (2005) “Genomic Signal Processing: Diagnosis and Therapy”, *IEEE Transactions Signal Process*, Vol. 22, pp 78-92.

[6] E. Dougherty, A. Datta, S. Chao, (2005) “Research Issues in Genomic Signal Processing”, *IEEE Transactions Signal Process*, Vol. 22, pp 46-68.

[7] S. Dudoit, J. Fridlyand, T. Speed, (2000) “Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data”, *Journal of American Statistics Association*, Vol. 97, pp 77-87.

[8] J. Khan, J. Wei, M. Ringner, L. Saal, M. Landanyi, F. Westermann, (2001) “Classification and Diagnostic Prediction of Cancers using Gene Expression Profiling and Artificial Neural Networks”, *Nature Medicine*, Vol. 7, pp 670-3.

[9] T. Furey, N. Cristianini, N. Duffy, D, Bednarski, M. Schummer, D. Haussler, (2000) “Support Vector Machine Classification and Validation of Cancer Tissue Samples using Microarray Expression Data”, *Bioinformatics,* Vol. 16, pp 906-14.

[10] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, (2002) “Gene Selection for Cancer Classification using Support Vector Machines”, *Machine Learning*, Vol. 46, pp 389-422.

[11] K. Mao, (2004) “Feature Subset Selection for Support Vector Machines Through Discriminative Function Pruning Analysis”, *IEEE Transaction on Systems*, *Man, and Cybernetics*, Vol. 34, pp 60-67.

[12] H. Wong, H. Wang, (2008) “Constructing the Gene Regulation-Level Representation of Microarray Data for Cancer Classification”, *Journal of Biomedical Informatics*, Vol. 41, pp 95-105.

[13] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, (1999) “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring”, *Science,* Vol. 286, pp 531-537.

[14] S. Dudoit, Y. Yang, M. Callow, T. Speed, (2002) “Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments”, *Statistica Sinica*, Vol. 12, pp 111-139.

[15] Y. Lee, C. Lee, (2003) “Classification of Multiple Cancer Types by Multicategory Support Vector Machines using Gene Expression Data”, *Bioinformatics,* Vol. 19, pp. 1123-1139.

[16] R. Banfield, L. Hall, O. Bowyer, W. Kegelmeyer, (2007) “A Comparison of Decision Tree Ensemble Creation Techniques”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 29m No. 1, pp 173–180.

[17] M. Dettling, P. Buhlmann, (2003) “Boosting for Tumor Classification with Gene Expression Data”, *Bioinformatics*, Vol. 19, No. 9, pp 1061–1069

[18] D. Ramon & A. Sara, (2006) “Gene Selection and Classification of Microarray Data using Random Forest”, *Bioinformatics*, Vol. 7, No. 3, pp 120-132.

[19] Y. Peng, (2006) “A Novel Ensemble Machine Learning for Robust Microarray Data Classification”, *Computers in Biology and Medicine*, Vol. 36, pp 553-573.

[20] Y. Freund & R. Schapire, (1996) “Experiments with a New Boosting Algorithm”, *Proceedings of 13th International Conference on Machine Learning*, pp 148-156.

[21] C. Zhang & J. Zhang, (2008) “RotBoost: A Technique for Combining Rotation Forest and AdaBoost”, *Pattern Recognition Letters*, Vol. 29, pp. 1524-1536.

[22] T. Zhang & M. Ogihara, (2004) “A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification based on Gene Expression”, *Bioinformatics*, Vol. 20, pp 2429-2437.

[23] Y. Freund & R. Schapire (1997) “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”, *Journal of Computer and System Sciences*, Vol. 55, No. 1, pp 119–139.

[24] J. Rodreguez, L. Kuncheva, C.Alonso, (2006) “Rotation Forest: A New Classifier Ensemble Method”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 10, pp 1619-1630.

[25] A. Krogh, J. Vedelsby, (1995) “Neural Network Ensembles, Cross Validation, and Active Learning”, In: G. Tesauro, D. Touretzky, T. Leen, (Eds.), *Advanced Neural Information Processing Systems*, Vol. 7. MIT Press, Cambridge MA, pp 231-238.

[26] T. Dietterich, (2000) “An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting and Randomization”, *Machine Learning*, Vol. 40, No. 2, pp 139-157.

[27] L. Hansen, P. Salamon, (1990) “Neural Networks Ensembles”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 10, pp 993-1001.

[28] I. Gheyas, (2010) “Feature Subset Selection in Large Dimensionality Domain”, *Pattern Recognition*, Vol. 43, No. 1, pp 5-13.

[29] L. Yu, H. Liu, (2004) “Efficient Feature Selection via Analysis of Relevance and Redundancy”, *Journal of Machine Learning Research*, Vol. 5, pp 1205-1224.

[30] L. Breiman, (2001) “Random Forests”, *Machine Learning*, Vol. 45, No. 1, pp 5-32.

[31] C. Ambroise, G. McLachlan, (2002) “Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data”, *Proceeding of National Science*, Vol. 99, pp 6562-6566.

[32] U. Braga-Neto, E. Dougherty, (2004) “Is Cross-Validation Valid for Small-Sample Microarray Classification? ”, *Bioinformatics*, Vol. 20, No. 3, pp 374-380.