Review on Classification of Genes and Biomarker Identification

Seema S and Hamida Honnalli
Department of Computer science
MSRIT,Bangalore.

ABSTRACT

Recent advances in the DNA microarray technology have provided the ability to examine and measure the expression levels of thousands of genes simultaneously in an organism. In this technology each gene is recorded under different conditions or each gene is evaluated under a single environment but in different types of tissues. In the first case it is used in identification of functionally related genes where asthe second type of technology is helpful in classification of different types of tissues and identification of those genes whose expression levels are good diagnostic indicators. Different approaches have been applied to classify different datasets. However, the main challenges in this task is the availability of a smaller number of samples compared to huge number of genes and the noisy nature of biological data. This paper review on different techniques used to classify the genes and improved efficiency of biomarker identification due to these classifications.

Keywords: Data mining, DNA microarray, Support vector machine (SVM), Decision tree, Neural network, Biomarkers.

1. INTRODUCTION

In DNA microarray data, each data point produced by a DNA micro-array hybridization experiment represents the ratio of expression levels of a particular gene under two different experimental situations. First the experiment starts with micro-array construction, where the several thousand of DNA samples are fixed to a glass slide, each at particular position in the array.*m*-RNA samples are then collected from a population of cells subjected to various experiments. These samples are converted to cDNA via reverse transcription and are labeled with one of two different fluorescent dyes in the process.Each one of this experiment consists of hybridizing the microarray with two differently labeled cDNA samples collected at different times. One of the samples is from the reference or background state of the cell, while the other sample represents a special condition.One of the samples is from the reference or background state of the cell,where as the other sample represents a special condition set up by the experimenter. The intensity of expression of each individual gene is roughly proportional to the amount of cDNA that hybridizes with the DNA affixed to the slide. By calculating the ratio of each of the two dyes present at the position of each DNA sequence on the slide using the technology called laser scanning technology, the levels of gene expression for any pair of conditions can be calculated or measured. An experiment with *n* DNA samples on a single chip, results in a series of *n* expression-level ratios. The denominator is the expression level of the gene in the reference state of the cell, while the numerator of each ratio is the expression level of the gene in the condition of interest to the experimenter .The data from these series of *m* such experiments may be represented as a gene expression matrix, where each of the *n* rows consists of an *m*-element expression vector for a single gene. Emerging advances in micro-array "chip" technology allow the simultaneous analysis of expression patterns for thousands of gene sequences (i.e chip features) and will serve as precursors to genome-wide functional analyses. These studies help in identifying complex disease genes and bio-markers for disease diagnosis and for assessing drug efficacy and toxicity.

The development of these technologies has also provoked importance of their use in clinical trials and diagnosis. One of the best applications of gene expression analyses is bio-marker identification, thesebio-marker helps for disease risk assessment, detection, prediction response to therapy, and preventative measures. Bio-markers are expected to be more accurate, reliable, efficient for assessing disease risk and biological effect; inexpensive and simple to perform.Micro-arrays provide rapid, efficient, and systematic approaches to searching bio-markers with more accuracy for disease diagnosis and prognosis, and understanding the basic biology of a disorder. Although micro-arrays can generate a large amount of informative data, to discover a reliable and efficientbio-markers a computational and statistical methods are required.

2. BACKGROUND

With the introduction of DNA microarray and gene expression the next part of this paper reviews on various classification of genes by support vector machine (SVM), decision tree and neural network.

3. CLASSIFICATION BASED ON SVM

In paper [1] author uses gene classification usingSVM-REF, in this paper the efficiency of RFE (Recursive Feature Elimination) for SVMis compared against the "naïve" ranking on a subset of genes. They found that SVM-RFE is better than SVM without RFE and also to other multivariate linear discriminate methods, such as Linear Discriminated Analysis (LDA) and Mean Squared Error (MSE) with recursive feature elimination.

In this study, the author addresses the problem of selection of a small subset of genes from broad patterns of gene expression data, recorded on DNA micro-arrays. Author also addresses classification problems. In the problem, the features are gene expression coefficients and patterns belong to patients. Here the input is a vector called a "pattern" of *n* components called "features". *F* isthe *n*-dimensional feature space.

Guyon et.al, addresses the problems of dimensionality reduction and feature selection.To overcome the above problems Guyonet.al, used SVM for classification of gene. Although SVMs handle non-linear decision boundaries of arbitrary complexity, he limited the studyto linear SVMs because of the nature of the data sets under investigation. Guyon et.al.,used one of the variants of the soft-margin algorithm. Training consists in executing the following quadratic program:

Algorithm SVM-train:

Inputs: Training examples $\{x_1, x_2, \ldots \ldots x_l\}$and class labels$\{y_1, y_2, \ldots\ldots y_k \ldots\ldots\ldots y_l\}$

Minimize overα_k:

$$J = \left(\frac{1}{2}\right)\sum_{hk} y_h y_k \alpha_h \alpha_k (x_h \cdot x_h + \lambda\delta_{hk}) - \sum_k \alpha_k$$

Subject to:

$$0 \leq \alpha_k \leq c \; and \; \sum_k \alpha_k y_k = 0$$

Outputs: Parametersα_k.

Here author applied this method on two data sets both of which consist of a matrix of gene expression vectors obtained from DNAmicro-arrays for a number of patients. The first data set was taken from cancer patients with two different types of leukemia. The second data set was taken from cancerous or normal colon tissues. The problem in the first case is to distinguish between two variants of leukemia (ALL and AML).Guyon et.al., compared his work with baseline method, the proposed method eliminates gene redundancy automatically and yields better and more compact gene subsets than baseline method. In patients with leukemia, the proposed method discovered 2 genes that yield zero leave one-out error, while baseline method needs 64 genes to get the best result that is one leave-one-out error. This method is 98% accurate using only 4 genes in case of colon cancer database, while the baseline method is only 86% accurate.

The paper [25] researched on gene classification using SVM. As members of a larger class of algorithms called kernel methods. This method can be non-linearly mapped to a higher-order feature space by replacing the dot product operation in the input space with a kernel function *K*, support vector machines (SVM) is a relatively new learning algorithm. This algorithm is introduced to solve two-class pattern recognition problems using the principle called Structural Risk Minimization. Suppose given a training set in a vector space, this method finds the best decision hyper plane that separates a set of positive examples from a set of negative examples with maximum margin. The quality of a decision hyper plane is determined by the distance (referred as margin) between two hyper planes that are parallel to the decision hyper plane and touch the closest data points of each class. The decision hyper plane with the maximum margin is the best decision hyper plan. With this definition of the hyper plane, SVM is able to generalize to unseen instances quite effectively. In recently applications of SVMs to DNA microarray data analysis, kernel functions are often selected to be linear, polynomial, or Gaussian form.

In [22] proposes a new feature selection method that uses a backward elimination procedure similar to that implemented in support vector machine recursive feature elimination (SVM-RFE). At each step, the proposed approach computes the feature ranking score from a statistical analysis of weight vectors of multiple linear SVMs trained on subsamples of the original training data.CV [3] is basically a method for estimating predictive generalization error based on re-sampling. The resulting estimate of generalization error is often used for model selection by choosing the model that has the smallest estimated generalization error. Before computing the ranking score for each feature, it is important to normalize the weight vectors,

$$W_j = \frac{W_j}{\|w_j\|}$$

Authors refer to this new feature selection method as MSVM-RFE, where MSVM stands for multiple SVMs. Duanet.al.,evaluated SVM-RFE and MSVM-RFE on four gene expression datasets: Breast Cancer (Breast), Colon Tumor (Colon), ALL-AML Leukemia (Leukemia), and Lung Cancer (Lung). Information about the datasetsis as used in [20].

1. Breast dataset: With 24481 genes,number of training samples being 78,and 19 test samples.
2. Colon dataset: With 2000 genes,number of training samples being 42,and with 20 test samples.
3. Leukemia dataset: With 7129 genes,number of training samples 38,and with 34 test samples.
4. Lungs dataset: With the12533 genes, number of training samples 32, with 149 test samples.

The classification performance is improved significantly with gene selection either by SVM-RFE or MSVM-RFE on all the datasets, even though SVMs are capable of handling a large number of input variables. And the best feature subsets selected by MSVM-RFE give better classification accuracy than the best feature subsets selected by SVM-RFE.

In paper [5] authors discovered new technology called Recursive Network Elimination (RNE) with SVM. Here author demonstrate an algorithm which integrates network information with recursive feature elimination based on SVM. First, filter one thousand genes selected by t-test from training set so that only genes that map to a gene network database remain. Then to the remaining genes the Gene Expression Network Analysis Tool (GXNA) is applied to form *n* clusters of genes that are highly connected in the network. Using these clusters Linear SVM is used to classify the samples and a weight is assigned to each cluster based on its significance to the classification. The clusters with less information are removed while retaining the remainder for the next classification step. This process is repeated until an optimal classification result is attained.

Authors used three dataset for experiment,

(1) CTCL (I) includes 18 patients and 12 controls [19]

(2) CTCL (II) consistsof 58 patients and 24 controls.

(3) Lymphocytedata is from the GXNA study [17].

Samples consist of 26 healthy and 30 melanoma patients. Authorshave used three algorithms (SVM-RFE, SVM-RCE,SVM-RNE on these dataset.

This experimental result gives 100% accuracy in case of CTCL(1) dataset with 4 gene identified as biomarkers. In case of dataset CTCL(II) it gives 91% of accuracy with 5 genes identified as biomarkers.

In [6] authors have introduced a novel method for cancer classification using expressions of few genes. This method used three datasets such asLymphoma, Liver and Leukemia datasets from micro array gene expression data. For both selection and classification this method uses the same classifier .The classifiers as Support vector machines-one against all (SVM-OAA), *k*-nearest neighbor (KNN) and Linear Discriminant analysis (LDA) were compared with one another.

Gene ranking can be performed by the use of Analysis of Variance (ANOVA). The classifier is validated using 5 fold cross validation (CV) technique. It includes the process such as pre-processing the gene expression data, top ranked gene selection, gene subset ranking, gene combination, gene selection using SVM and classification using SVM, KNN, LDA and finally testing data can be predicted. The classifiers SVM-OAA performed well on the lymphoma data.Same accuracy is achieved on liver and leukemia by KNN and SVM-OAA classifiers.

The experiment is repeated for top 10,20,30,50 and 100 genes and they found that the accuracy level for the lymphoma data varied from 80.65% to 100%, for Leukemia data accuracy is 89.54% to 100%, and for the Liver data ranging from 95.67% to 100%.

In[23] new approach on gene classification has been discovered. This approach proposes a Genetic Algorithm (GA) with Support Vector Machines (SVM) for the classification of high dimensional Microarray data. This approach involvespre-filtering technique which is based on fuzzy logic. The propose method involve 3-stages,

Stage 1:Pre-processing by fuzzy logic.This stage is to reduce the dimension of the initial problem by eliminating gene redundancy.

Stage 2:Gene subset selection by GA/SVM.Genes obtained from stage 1 is subjected with wrapper approach, which combines GA and SVM, and performs gene subset selection. This SVM evaluation is done on set of training data.

Stage 3: Classification. Reduced set of relevant gene obtained from stage 2 is used in final step for gene selection and classification. This newselection and classification is now done on set oftest data**.**Wrapper GA/SVM algorithm, uses a SVM classifier to evaluate the quality of a gene subset. Suppose a chromosome *x*that represents a gene subset, apply a (LOOCV) Leave-One-OutCross-Validation method to calculate the average accuracy (rate of correct classification) of a SVM trained with this gene subset. For each chromosome *x*,

$Fitness(x) = accuracySVM(x)$.

Leukemia dataset and the Colon cancer datasetare used for classification. The Leukemia dataset consists of 72 samples with 7129 gene expression levels. Colon cancer datasetcontainsthe expression of 6000 genes with 62 cell samples.Experimental results that for the Leukemia dataset, obtain a classification rate of 100% using 25 gens. Classification rate of 99.41% (with 10 genes) is reached for the Colon dataset.

In [7] authors researched on gene classification. The objective is to propose efficient cancer classification techniques which provide reliable and significant classification accuracy. To achieve this optimal classification the first research goal is to find the smallest set of genes that can ensure high accuracy in classification using supervised machine learning algorithms. The proposed method involves two steps. In the first step, with the help of Analysis of Variance (ANOVA) ranking scheme some important genes will be chosen. The second step, the classification capability is tested for all simple combinations of those important genes using a better classifier. The proposed method initially uses Support Vector Machine classifier. Then Modified Extreme Learning Machine classifier is used for increasing the classification accuracy over SVM.

The proposed method is applied on lymphoma data set and liver cancer dataset. The expression data of 4026 genes are included in the entire data set. Next, 20 genes with highest ANOVA is picked. The propose shows 100% accuracy with MELM method where as 98.7% accuracy with SVM.

In [8] the authors researched on effective classification of genes by Blending of LPP and SVM.In the proposed work, firstly, the LPP is used to reduce high dimensionality of the microarray gene data because LPP preserves the locality of neighborhood relationship. Secondly, the SVM is applied on the dimensionality reduced gene data for classification. In the process of information retrieval in DNA microarray technology, gene classification is tough task, if the data is highly dimensional and small in size.This technique is consisting of two steps, dimensionality reduction and SVM-based classification. In the dimensionality reduction, the high dimensional gene data is converted to low dimensional data. The resulting low dimensional data is then classified using a SVM. The SVM is trained by the gene data of different classes. Once the SVM is trained by the low dimensional gene data of various classes, it will be used to classify any of the similar gene expression data. So, before classification the SVM is trained with the aid of the gene data of different classes.

For the experiment purpose sample of human acute leukemia is used. The high dimensional gene expression data has been subjected to LPP-based dimensionality reduction and so a dimensionality reduced gene data with dimensions, 38×38 (i.e. ns =38) has been obtained. The accuracy obtained by this method is 97.29%.

4. CLASSIFICATION USING DECISION TREES

Decision trees are one of the most powerful and popular approaches in knowledge discovery and data mining, the technology of exploring large and complex bodies of data in order to discover useful patterns. It enables modeling and knowledge extraction from the abundance of data available. Decision trees, originally implemented in decision theory and statistics, which are highly effective tools in areas such as data mining,, machine learning, and pattern recognition, information extraction. Benefits of decision trees in data mining that:

- Decision trees are Self-explanatory and easy to follow when compacted
- It handles a variety of input data
- It processes the datasets that may have errors or missing values
- High prognostic performance for a relatively small computational effort
- Available in many data mining packages over a variety of platforms Useful for various tasks, for example classification, clustering and feature selection

In [17] research on gene classification is shown using decision tree technique. It was found that present decision tree methods perform poorly for classifying gene expression data. To address this problem, authors introduced a new technique for building decision trees that is better suited to this scenario. Method is based on consideration of the area under the Receiver OperatingCharacteristics (ROC) curve [18], to help determine decision tree characteristics, such as stopping criteria node and selection.

The main contributions in there paper is as follows:

- Development of a decision tree induction technique,ROC-tree, which in a novel way uses the area under the ROC curve (AUC) to select nodes of the tree. The aim of this development is to address the problem of classification performance of standard decision tree classifiers used for gene expression datasets.
- Purpose of an AUC-based criterion to stop growing the tree.
- An experimental investigation which demonstrates that *ROC-tree* outperforms well known techniques in terms of accuracy as well as overall AUC value.

The steps involve in algorithm for building the decision tree using the ROC measure.

- Selecting Nodes of the Tree
- Splitting Threshold
- Stopping Criterion
- Labeling the Leaf Nodes.

Authors used 12 datasets for analysis, of which 6 are GE datasets and 6 are non-GE datasets with having rather different characteristics. In the experimental analysis 10 techniques are used on datasets. The performance of the *ROC-tree* classifier, a 10-fold cross validation (CV) scheme is used 10 times for all 12 datasets.

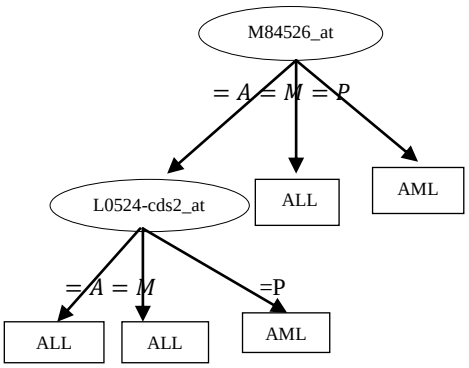
In [24] the authors have researched on gene classification. The main aim of this research is to construct classifiers that can be human readable as well as robust in performance in microarray data using decision trees. In this research they have used a real world leukemia microarray experiment performed in [19]. Leukemia cancer is a cancer of bone marrow or blood cells, i.e. a generalized neoplasic proliferation or a hematopoietic cells accumulation with or without peripheral blood involvement. There are four main types of leukemia:

- Chronic Myeloid Leukemia (CML),
- Chronic Lymphoblastic Leukemia (CLL)
- Acute Myeloid Leukemia (AML),
- Acute Lymphoblastic Leukemia (ALL).

The original dataset provided by [19] contains nominal (discretized) and continuous values of gene expression data. Therefore the author has performed experiments using both sorts of values. They defined three different datasets: two of them contain nominal (*S*) and continuous (*S*″) values both without any further data preprocessing; the third one (*S*′) contains nominal values preprocessed as follows. There are three nominal values: '*A*': gene is absent or not expressed; '*P*': gene is expressed or present and 'M': the level of the expression is marginal among '*A*' and '*P*'. Let S_1 denote the original training set with 38 examples and s1 denotes the original test set with 34 examples, both of them containing *'M', 'A'* and *'P'* values. Dataset S'_1 is equivalent to S_1 but *'M'* values were replaced by '?' and analogously for s'_1 from s_1.

Datasets S''_1 and s''_1 correspond to training and test sets respectively, both containingcontinuous values. From previous it is found that a feature selection method increases the number of rules in classifier, for this reason they adopted the different strategy. This approach have generated 30 decision trees $T_1,T_2,\ldots..,T_{10}$, and T'_1,T'_2 $\ldots\ldots T'_{10},T''_1,T''_2\ldots\ldots\ldots T''_{10}$. Considering individual trees, the best performance was achieved by T''_3 with 5.88% error rate and AUC 0.94, followed by both T_1 and T'_2 with 8.82% error rate and AUC 0.90.

The experimental results show three decision trees with three most significant genes identified by authors. One of the decision tree identified as significant gene by author is shown in figure 1.



T_1

Figure 1:Decision Tree (Decision tree identified byOscar Picchi et. al.,in [24])

In [26] authors presented a classification method based on decision rule using single gene. They used three dataset for experiment. They used machine learning method for feature selection[4] and classification. In comparisons with other this method is simple, effective and robust. Authors estimated the classification accuracy rate by testing on independent samples, which is more unbiased than the cross validation. Three datasets used here are leukemia, lung cancer and prostate cancer.

1. Leukemia dataset: With the 7129 genes, class ALL/AML, number of training samples 38,with 34 test samples.
2. Lung cancer dataset: With the 12533 genes,class MPM ADCA, number of training samples 32, with 149 test samples.
3. Prostate cancer dataset: With the 12600 genes, class Tumor/Normal, number of training samples 102, with 34 test samples.

Authors used decision table to represent every cancer classification related microarray. In the Decision table, *m* represents samples and *n represents* genes. Each sample is assigned to one class label. Each class is a decision attribute and each gene is a condition attribute. *g(x, y)* signifies the expression level of gene *y* in sample *x*.
Three maintask done here is data processing, gene selection, gene classification In gene selection informative gene is selected by α depended degree [21].This begin with α=1, then gradually decrease α value. In the worst case, stop attempts at the point of α=0.7, which is the lower bound. The genes with $\gamma_P(D, \alpha)$ =1 are picked out. Next, perform the classification based on the decision rules induced by the selected genes, and apply the classifiers for independent test sets to validate the classification performance.
In leukemia data samples nearly 8 genes are identified with higher accuracy, in which two genes have accuracy of 94-100% and other two genes have accuracy between 91-100%.

In lung cancer dataset nearly 25 genes with higher accuracy is selected based on α value. In which 1 gene with accuracy98-100% is identified.In the prostate cancer dataset, only when α=0.8, 11 genes are marked, with accuracy 91%.

In [16] authors applied the Data Mining Techniques for Cancer Classification using Gene Expression Data. He used t- Statistics (t-GA) based genetic algorithm for Feature selection from microarray dataset. The decision-based classifier is used which is applied on the top data sets. Colon, Leukemia, Lymphoma, Lung and Central nervous system (CNS) is selected as datasets. The performance of t-GA is compared with the previously used gene selection methods such as GA, t-Statistic, Info Gain and GS. The experimental result shows when applying the decision tress based classifier in all of these data sets with the scoring scheme t-GA provides highest accuracy than that of GA,Statistic, Info Gain and GS.

5. CLASSIFICATION BASED ON NEURAL NETWORK

The term neural network was traditionally used to refer to a network or circuit of biological neurons. The usage of the term often refers to artificial neural networks, which are consisting of artificial neurons or nodes. Artificial neural networks are composed of interconnecting artificial neurons such as programming constructs that mimic the properties of biological neurons. ANN may either be used to gain an understanding of biological neural networks, as well as for solving artificial intelligence problems without necessarily creating a model of a real biological system. Real biological nervous system is highly complex: artificial neural network algorithms attempt to abstract this complexity and focus on what may hypothetically matter most from an information processing point of view. Good performance such as measured by good predictive ability and low generalization error, or human error patterns, or performance mimicking animal, can then be used as one source of evidence towards supporting the hypothesis that the abstraction really captured something important from the point of view of information processing in the brain. One more incentive for these abstractions is to reduce the amount of computation required to simulate ANNs, so as to allow one to experiment with larger networks and train them on larger data sets.

In [11] a neural network is used for cancer type classification. The method consists of three major steps:First step is principle component analysis, 2nd step is relevant gene selection and last step is artificial neural network prediction. Dimensionality reduction is major problem involve in gene classifications. Principle component analysis [12] is to overcome this problem, which helps to avoid "over fitting" error in the supervised regression model. They observed that insertion of class labels into the reduction process does not provide optimal performance but introduces bias in the data. Thus, class labels are excluded from the dimensions that undergo reduction.

A model dependent analysis method is used for checking the relevancy of each gene, which is defined through sensitivity function. For a data set of N samples and K classes denoted as $c_1, c_2, \ldots, c_k$, the sensitivity of a gene $g_i$with respect to the class labels is defined as,

$$S_i = \frac{1}{N}\frac{1}{K}\sum_{j=1}^{N}\sum_{m=1}^{K} |\partial\, c_m / \partial g_i|$$

This formula gives the importance of a gene with respect to the total classification. In addition, they also specified sensitivity of each gene g_i with respect to each class, $c_j$defined as,

$$S_{ij} = \frac{1}{N}\frac{1}{K}\sum_{m=1}^{N} |\partial c_j / \partial g_i|$$

Where$c_j$is the j^{th} class label and $g_i$is the i^{th} gene. For each S_{ij}, they also defined a sign that tells if the largest contribution to the sensitivity is due to positive or negative terms. The $S_i$and S_{ij}values of genes are calculated, and genes are ranked both according to their importance with respect to the total classification and to their importance with respect to each individual cancer class. The class prediction was done using an Artificial Neural Network (ANN) classifier [13]. It was observed that selection of 96 genes gives the best performance for the data set they used (88 samples of 6567 genes in which 63 are used in the training process and 25 used in the test).

In [9] authors proposed a novel radial basis functions (RBF) neural network for cancer classification using expression of very few genes. This technique was applied to the three data sets the lymphoma, the small round blue cell tumors (SRBCT), and the ovarian cancer.T-test scoring method is use for gene ranking to measure the discriminative ability of genes. RBF neural networked used only 9 genes for the lymphoma data, this approach also obtained 100% accuracy in the SRBCT data set with 8 and the ovarian data with 4 genes. RBF method includes two steps;first select some genes with the greatest discriminative ability in the training data. In the second step, use the selected genes to train RBF neural network and subsequently use the trained network to classify the testing data.Therefore, the RBF neural network consumes fewer genes as well as it also reduces the gene redundancy for cancer classification using micro array data compared to the previous nearest shrunken centroids.

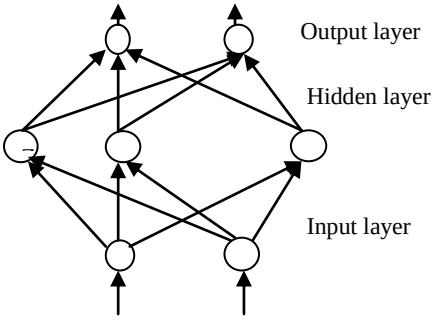


Figure 2:RBF neural network

An RBF neural network has three layers as shown in Figure 2.

- Input layer
- Hidden layer that includes some radial basis functions, also known as hidden kernels,
- The output layer.

An RBF neural network can be considered as a mapping of input domain X onto the output domain Y. The most commonly used kernel function for RBF neural networks is Gaussian kernelfunction,

$$G(\|\vec{x} - \vec{t_i}\|) = \exp\left(-\frac{\|\vec{x} - \vec{t_i}\|^2}{2\sigma_i^2}\right)$$

Where$\sigma_i$is the radius of the kernel *i*. The steps involve in constructing an RBF neural network includes:

(a) Determining the positions of all the kernels t_i,

(b) Determining the radius of each kernel

(c) Calculating the weights between the kernels and the output nodes.

The lymphoma data set entire data set includes the expression data of 4026 genes. The SRBCT data contains the expression data of 2308 genes. The ovarian data contains 125 samples; the entire data set includes the expression data of 3363 genes.

In paper [10] the authors proposed there work on gene classification.Given an enhanced wavelet neural network for early analysis of cancer patients using clustering algorithms. A variety of clustering algorithms, such as K-means (KM), symmetry-based K-means (SBKM),Fuzzy C-means (FCM), symmetry-based Fuzzy C-means (SBFCM) and modified point symmetry-based K-means (MPKM) is used by author in his work. Based on these algorithms he caused the translation parameter. The data sets such as LEU, SRBCT, GLO and CNS collected for the development of cancer classification in the use of micro array gene expression data. Feature selection from micro array gene data set is performed by using T-Test. The highest classification can be achieved with the use of MPKM algorithms in all the three data sets. The experimental results showed that the proposed classifiers achieved a superior accuracy, which ranges from 86% to 100%. Performance comparisons are also done with some other classifiers, which show this proposed approach outperforms most of them.

In [14] authors established a classifier called Semi supervised Ellipsoid ARTMAP (SsEAM) for multi class cancer. Informative gene selection has been completed by Particle swarm optimization. The classifier such as Semi supervised Ellipsoid ARTMAP is a neural network architecture that is embedded in Adaptive Resonance Theory classification tasks have been performed by clustering data that are attributed with the same class label. An evolutionary algorithm-based technique called PSO for global optimization used to point out whether the genes are designated or not. The data set used in his work is NC169 data from the national cancer institute [15], Acute Leukemia Data and acute lymphoblastic leukemia (ALL) data set. Classification accuracy for three of these data sets has been computed using EAM, SsEAM, PNN, ANN, LVQ1 and KNN. PSO and Fisher Criterion based on the classifier have made a Gene selection. Compared with other machine learning technique SsEAM with PSO performed well on all of these three data sets as well as classification accuracy also is different and significant.

NCI60 Data

The data set includes 1,416 gene expression profiles for 60 cell lines in a drug discovery screen by the National Cancer Institute the best result author obtain with ssEAM/PSO is 87.9 percent (79 genes are selected by PSO).

Acute Leukemia Data

This datasets consist of 72 samples that contain three different leukemia types,i.e 25 acute myeloid leukemia (AML),38m B-cell acute lymphoblastic leukemia (ALL), and nine T-cell ALL These samples are divided into two groups, 38 for training and 34 for testing .The best classification accuracy is achieved by ssEAM when 63 or 97 genes are selected with PSO,and the accuracy with 97%.

All Data

This data set consists of six different acute lymphoblastic leukemia subtypes; the best classification accuracy is achieved by ssEAM when 95 genes are selected with PSO,with efficiency of 98%.

6. CONCLUSION

The above survey concludes that better neural network techniques can be incorporated with the present research work for less complexity and better learning capacity. And it also concludes thatSVM has performed better in almost all the datasets. Different classification methods have different performance on different datasets. Researchers have also used the ensemble of classifier to exploit the characteristic of each classifier and the combined effect improves the performance for any dataset in use. Ensemble of feature selection can also be used to obtain the more significant genes for better classification. Table 1 reviews the accuracies obtained for different classifier (SVM, Decision trees and neural networks) on different cancerous microarray gene expression datasets. The table 1 also lists the count of the number of marker genes identified by these algorithms.

Table 1: The table shows the accuracy and biomarkers obtained for different classification methods.

Author	Techniques	Accuracy (%)	Number of gene Identified
R. Mallika, And V. Saravanan	Novel Method-SVM OAA	Lymphoma=100% Liver data=97.44% Leukaemiadata=95.83%	10
RuiXu, Anagnostopoulos	Semisupervised Ellipsoid ARTMAP(neural network architecture)	NCI60 Data=89.7% AcuteLeukemia=97% All Data=98%	NCI60 Data=79 Acute Leukemia=63-97 All Data=95
Jinn	*T-Genetic Algorithm*-decision tress based classifier	97.70%	Unknown
Osareh, A. Shadgar, B.	SVM, K-nearest neighbors and probabilistic neural networks	98.80%	Unknown
Zainuddin. Z and Pauline	MPKM-Wavelet Neural Networks	88-100%	Unknown
Lipo Wang, Feng Chu,	Fuzzy, neural networks-RBF	Lymphoma=100% SRBCT=100% Ovarian=100%	Lymphoma=9 SRBCT=8 Ovarian=4
XIAOSHENG WANG and OSAMU GOTOH	Single gene decision rule	Leukemia=94-100% Lung cancer= 98-100% Prostate cancer =91%	Leukemia=8 Lung cancer=25 Prostate cancer=11
MarufHossainMd.Rafiul Hassan	Decision tree with ROC	89.0%	Unknown
Malik Yousef, Mohamed Ketany	SVM-RNE	CTCL(I)=100% CTCL(II)=91% Lymphocyte =80%	CTCL(I)=4 CTCL(II)=5 Lymphocyte=13
Oscar PicchiNetto, Ricardo Nozawa	Decision tree	Leukemia =95%	(AML AND ALL) DT(nominal)=2 DT(continuous)=1
A. Bharathi and A.M. Natarajan	ANOVA Ranking Scheme-SVM	97.91%	Unknown
ISABELLE GUYON	SVM	Leukemia=100% Colon=98%	Leukemia=2 colon cancer=4
Edmundo Bonilla Huerta, B´eatrice Duval, and Jin-Kao	Hybrid GA/SVM	Colon=99.41% Leukemia=100%	Leukemia =25 Colon=10
Jagath and Haiying Wang	MSVM-RFE	95%	Breast=161 Colon=3 Leukemia=37 Lung dataset=3

7. REFERENCES

[1] Guyon, I., J. Weston, S. Barnhill and V. Vapnik, 2002,"*Gene selection for cancer classification using support vector machines",*Machine Learning, Volume 46, pp 389- 422.

[2] M.P.S.Brown, W.N.Grundy, D.Lin, N.Cristianiti, C.W. Sugnet, T.S.Furey, JrM.Ares, and Haussler, "*Knowledge-based analysis of microarray gene expression data by using support vector machines*", National Academic of Science., USA, 97(1):262-267,2000.

[3] M. Stone, "*Cross-valedictory choice and assessment of statistical predictions*," *Journal of the Royal Statistical Society:Series B*, Vol. 36, no. 1, pp. 111–147, 1974.

[4] I. Guyon and A. Elisseeff, "*An introduction to variable and feature selection*," *Journal of Machine Learning and Research*, no. 3, pp. 1157–1182, Mar. 2003.

[5] Malik Yousef, Mohamed Ketany, Larry Manevitz, Louise C Showe and Michael K Showe"*Classification and biomarker identification using gene networkmodules and*

support vector machines", *BMC Bioinformatics* 2009, Volume 10,pp:337

[6] R. Mallika, And V. Saravanan, "*An SVM Based Classification Method For Cancer Data Using Minimum Microarray Gene Expressions*", World Academy Of Science, Engineering And Technology 62, pp.543-547,2010.

[7] A. Bharathi and A.M. Natarajan "*Efficient Classification of Cancer using SupportVector Machines and Modified Extreme LearningMachine based on Analysis of Variance Features*", American Journal of Applied Sciences Volume 8, No.12, 1295-1301, 2011.

[8] J. Jacinth Salome and R. M. Suresh "*An Effective Classification Technique for Microarray Gene Expression by Blending of LPP and SVM*", European Journal of Scientific Research, Volume 64, No.1 (2011), pp. 32-41

[9] Feng Chu and Lipo Wang, "*Applying Rbf Neural Networks To Cancer Classification Based On Gene Expressions*," International Joint Conference on Neural Networks, July 16-21,2006.

[10] Zainuddin. Z and Pauline. O, "*Improved wavelet neural network for early diagnosis of cancer patients using microarray gene expression data*", International Joint Conference on Neural Networks, 2009. IJCNN 2009.

[11] J. Khan, J. Wei, M. Ringner, L. Saal, and et. al.," *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks*", Nature Medicine,7(6):673-9, 2001.

[12] Jian J. Dai, Linh Lieu, David Rocke," Dimnesion reduction for classification with gene expression microarray data", Statistical applications in Genetics and molecular biology, Volume 5, Issue 1, 2006, pp 1-19

[13] Farid E Ahmed,"*Artificial neural networks for diagnosis and survival prediction for colon cancer*",Molecular cancer 2005, Volume 4, issue 29.

[14] RuiXu, Anagnostopoulos, G.C. And Wunsch, D.C.I.I.,"*Multiclass Cancer Classification Using Semi supervised Ellipsoid ARTMAP and Particle Swarm Optimization with Gene Expression Data", IEEE/ACM Transactions On Computational Biology* And Bioinformatics, Vol.4, No.1, pp. 65-77, 2007.

[15] U. Scherf, D. Ross, M. Waltham, L. Smith, J. Lee, L. Tanabe, K. Kohn, W. Reinhold, T. Myers, D. Andrews, D. Scudiero, M. Eisen, E. Sausville, Y. Pommier, D. Botstein, P. Brown, and J. Weinstein, "*A Gene Expression Database for the Molecular Pharmacology of Cancer*," Nature Genetics, vol. 24, pp. 236-44, 2000.

[16] Jinn-Yi Yeh, Tai-Shi Wu, Min-Che Wu and Der-Ming Chang, "*Applying Data Mining Techniques for Cancer Classification from Gene Expression Data*", International Conference on Convergence Information Technology, 2007.

[17] M. MarufHossain ,Md. Rafiul Hassan andJames Bailey," *A Novel Decision Tree Induction Algorithm Based on Receiver Operating Characteristics to Classify Gene Expression Data*," The University of Melbourne, Australia.

[18] D. M. Green and J. M. Swets' "*Signal detection theory and psychophysics*" JohnWiley& Sons Inc., New York,USA, 1966, ISBN:0-471-32420-5.

[19] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M.,Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomfield,C. D., and Lander, E. S. (1999). "*Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*." Science, Volume 286, pp 531–537

[20] J. Li and H. Liu. (2002) Kent Ridge Bio-Medical Data Set Repository: http://sdmc.lit.org.sg/GEDatasets/Datasets.html

[21] Wang, X.,Gotoh O, "*Microarray-Based Cancer Prediction Using Soft Computing Approach*", *Cancer Informatics*, Volume 7, pp123–139, 2009.

[22] Kai-Bo Duan, Jagath C. Rajapakse, HaiyingWang,and Francisco Azuaje, *"Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data*", IEEE transactionsonnanobiosciences, Volume 4, No. 3, pp. 228-234,September 2005.

[23] Edmundo Bonilla Huerta, B´eatrice Duval, and Jin-Kao Hao "*A Hybrid GA/SVM Approach for Gene Selection and Classification of Microarray Data*", Evolutionary Workshops2006, Lecture Notes in Computer Science, pp. 34–44, 2006

[24] Oscar PicchiNetto, S´ergio Ricardo Nozawa Rafael Andr´es Rosales Mitrowsky, Alessandra Alaniz Macedo1, Jos´e Augusto Baranauskas, "*Applying Decision Trees to Gene Expression Data from DNA Microarrays: A Leukemia Case Study*", 20-23 July 2010, ISSN 2175-2761.

[25] Junying Zhang, Richard lee, Yue Joseph Wang, "*Support Vector Machine Classifications for Microarray Expression Data Set*," IEEE, 2003. Page 67 ISBN:0-7695-1957-1

[26] Wang X,Gotoh O." *Accurate molecular classification of cancer using simple rule",* 2009 Oct 30; Volume 2, No. 64,2009.