A NOVEL MICROARRAY GENE RANKING AND CLASSIFICATION USING EXTREME LEARNING MACHINE ALGORITHM

T.REVATHI[1], DR.P.SUMATHI[2]

[1]Associate Professor & Research Scholar, Manonamniam Sundaranar University, Tirunelveli.
[2]Assistant Professor, PG & Research Department of Computer Science, Govt. Arts College, Coimbatore.
E-mail: trevathi_psg@yahoo.co.in

ABSTRACT

The research studies the suitability of Extreme Learning Machines (ELM) for resolving bioinformatics and biomedical classification problems. It is used for some direct multicategory classification to solve those problems. The cancer diagnosis has three benchmark microarray datasets for evaluated the multi-category classification performance of ELM. In order to test their overall performance, an experimental study is presented based on three gene microarray datasets found in bioinformatics and biomedical domains. This research work use this ELM for quick performance and it is better accuracies. The result can be produces with minimum time compared to artificial neural networks methods. This method shows promising classification accuracy for all the test data sets. It also shows the relevance of the selected genes in terms of their biological functions.

Keywords: *Cancer classification, DNA microarray gene expression data, Extreme learning machine.*

1 INTRODUCTION

The ratio of expression levels of a particular gene under two various experimental conditions are represented by a DNA microarray hybridization. Cancer is a diseases caused by out-of-control cell growth. The damaged cells divide uncontrollably to form lumps or masses of tissue in body called tumors. The primary tumors caused only 10% of cancer deaths. Cancer can classify 5 groups they are carcinomas, sarcomas, lymphomas, leukemia's and adenomas. More dangerous, or malignant, tumors form when two things occur: i) invasion ii) angiogenesis. The single-hidden layer feedforward networks are generalized by ELM. It may have some hidden layers but it not tuned it is also called feature mapping. The SLFNs is not used for single-hidden-layer and multi-hidden-layer. The tenet is a neural network that identify the hidden nodes when SLFNs to be tuned.

Direct gene transfer for the treatment of human diseases requires a vector. It is very efficient and safe [1].A gene microarray data analysis have coupled two-way clustering approaches. One of these is used to cluster the other, stable and significant partitions emerge is the main idea to identify subsets of the genes and samples. Based on iterative clustering this algorithm is presented and it performs such a search. This is suitable for gene microarray data. The two gene microarray data sets, on colon cancer and leukemia may use this method. It is able to discover partitions and correlations by identifying relevant subsets of the data [2]. The DNA testing may have the development of small, fast and easy-to-use devices. The preparation, operation and applications of biosensors and gene chips are discussed here. DNA biosensors and gene chips may also have some new strategies about recent trends. The process of 'Lab-on-a-Chip' may also cover in this integration of hybridization detection schemes. They conclude that DNA biosensors and gene chips are at an early stage are expected to have an enormous changes and effect for future [3]. The Clustering analysis has become a valuable and useful tool for analysis of microarray or gene expression data. In proposed a number of clustering methods are containing difficulties in meeting the requirements of automation, high quality, and high efficiency at the time simultaneously. A novel represents a efficient clustering algorithm, namely correlation search technique (CST), which fits for analysis of gene expression data. CST is used to out performing the other clustering methods include the terms of clustering quality, efficiency, and automation [4].Selection bias in gene extraction on the basis of microarray gene-expression data [5].The DNA microarrays process in cancer and other disease states are used for numerous studies. It is also called genes expressed across the gamut of human tissues. The studies of this global gene-expression pattern are performed by linking variation in the expression of specific genes, which provide clues to the potential roles of the genes and to the molecular organization of diverse cells [6]. A vector clustering was developed for a topological and dynamical characterization of the cluster structures. Each cluster are decomposed its constituent basin level cells and also enlarged clustered domain is extended naturally to serves as a basis for inductive clustering. A weighted graph preserving are simplify to construct the cluster and also to develop a robust and inductive clustering algorithm [7]. Biclustering approach is most important in microarray data analysis. Across subsets of sample, one can identify sets of genes sharing compatible expression patterns using this algorithm. The patterns are providing clues for main biological processes in various physiological states. The methods express the uses of singular value decomposition (SVD) as its framework to identify the problems of bicluster from gene expression matrix which transform into two global cluster problems. After the function of biclustering algorithm, blocks of up-regulated or down-regulated in gene expression matrix so which the genes are co-regulated and also functionally related [8]. Clustering is used to select the set of informative and relevant genes. The genetic algorithms are proposed their encoding technique to performing the set of tasks of gene selection and fuzzy clustering simultaneously. These techniques are used to minimizing the number of selected genes and the number of cluster perform automatically. The comparison of both artificial data sets and several other related feature selection and also clustering approaches using the proposed technique [9].The problem of microarray data processing is a selection of significant gene through expression patterns. This article proposed that the spectral biclustering technique for selecting relevant gene. The algorithm which are proposed to make accurate predictions while selecting a smaller subset gene. The two microarray cancer data sets, i.e., the lymphoma and the liver cancer data sets are demonstrated for unsupervised gene selection method [10].Protein sequence classification using extreme learning machine [11]. The analysis of transcriptome or gene expression data may have important step that is a Feature selection. This feature may reduce the curse of the dimensionality problem and improves the interpretability of the problem. These method may have arbitrarily or heuristic fashion, it use number of genes. The optimal number of genes to be selected in the theoretical way. They conclude that this newly developed of proposed strategy has been applied on a number of gene expression datasets and expected results have been obtained [12]. To predict protein subcellular localization they developed many computational methods. Some of the methods in that are limited to the prediction of single-location proteins. Multi-location proteins may or may not consider. Then it may have special biological functions because of that proteins with multiple locations are particularly interesting. It is essential to both basic research and drug discovery [13]. The statistical analysis may use gene subset for prediction such as survival and functional analysis for understanding biological characteristics.Agene expression datause this null space based feature selection method forclassification. By applying the information of null space of scatter matrices they discard the redundant genes. They use some method theoretically and demonstrate. It is very effectiveness on several DNA gene expression datasets. The method is easy to implement and computationally efficient [14].

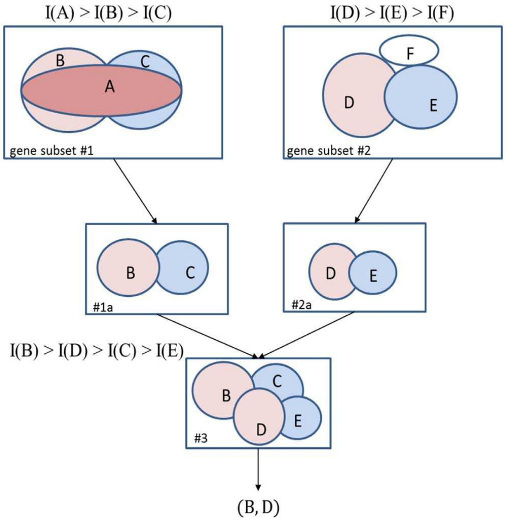

Fig. 1.An illustration of the process of gene selection using the extreme learning machine algorithm.

Two gene subsets are illustrated, with each containing three genes (h = 3). The genes in subset

#1 and subset #2 are A, B, C andD, E, F, respectively. The elliptical area in the figurecorresponds to the information I(G), pertaining to a geneor gene subset G. It can then be seen from the figure thatI(A) > I(B)> I(C) and I(D) >I(E) > I(F). Then, the problemis to select two genes (r=2) from the given genesubsets. First, the search algorithm finds two genes eachfrom the subsets that have the highest information incombination. It can be seen from the figure that though geneA (in gene subset #1) has the highest information, itsignificantly overlaps with its neighboring genes. BecauseI(B U C) > I(A U B) and I(B U C) > I(A U C) removing gene A from the subset sacrifices the least amount of information.

This yields new gene subset #1a. Most of the forwardselection and individual ranking schemes would select geneA, as it pertains to the highest information. However, theproposed scheme discards gene A because it is a redundantgene in combination with other genes. In gene subset #2, gene F does not overlap with any other genes. However,because I(D U) > I(D U F) and I(D U E) > I(E U F), genewill be discarded, as it contains minimal information. In thiscase, they selected the gene B and D is best.

2 PROPOSED FEATURE SELECTION ALGORITHMS

In this section, they describe the proposed feature selection of algorithm. The main aspects of the algorithm namely, Extreme Learning Machine are discussed in Sections 2.1.

2.1 Extreme Learning Machine

ELM was originally proposed for standard single hidden layer feedforward neural networks.

1. A unified learning platform with widespread type of feature mappings is provided by ELM. It is applied in regression and multi-class classification applications directly;
2. ELM has milder optimization constraints compared to SVM, LS-SVM and PSVM;

It obtain a best solution and complexity when compared to ELM, SVM, etc., then

ELM has been successfully used in the following applications: Biometrics, Bioinformatics, Image processing, Signal processing etc., ELM is a simple tuning-free three-step Algorithm. The learning speed of ELM is extremely fast. The traditional classic gradient-based learning algorithms which only work for differentiable activation functions but it also facing several issues like local minima, overfitting, etc.,

Algorithm ELM: Given a training set ℵ = {(xi,ti)|xi ∈R^n,ti∈R^m,I = 1,…,N}, activation function g(x), and hidden node number N,

Step 1: Randomly assign input weight wi and bias bi, I = 1,…N.

Step 2: Calculate the hidden layer output matrix H.

Step 3: Calculate the output weight β

(β = H+T)

Where

T =[t1,…,tN]^T.

The ELM learning algorithm looks much simpler and it gives accurate result when compare to other algorithms. Machine Learning is about building programs with tunable parameters. Extreme Learning Machines (ELM) provides efficient solutions. ELM possesses unique features to deal with regression and (multi-class) classification tasks. Consequently, ELM offers advantages such as fast learning speed, ease of implementation.

3 EXPERIMENTAL RESULTS

In this section, they described about the experimental setup of data sets.

3.1 Data Sets Used In Experimental Setup

There are three DNA microarray gene expression data sets areutilized in this work to show the effectiveness of theproposed method. The descriptions of the data sets are givenas follows:
SRBCT data set: the small round blue-cell tumor dataset consists of 83 samples, each containing 2,308 genes. This is a 4-class classification problem. The tumors are Burkitt 's lymphoma (BL), the Ewing family of tumors (EWS), Neuro Blastoma (NB), and rhabdomyosarcoma (RMS). Thereare 63 samples for training and 20 samples for testing. Thetraining set consists of 8, 23, 12, and 20 samples of BL, EWS, NB, and RMS, respectively. The test set consists of 3, 6, 6, and5 samples of BL, EWS, NB, and RMS, respectively.
MLL data set [2]: this data set contains three classes ofleukemia, namely acute lymphoblastic leukemia (ALL), myeloid/lymphoid leukemia (MLL), and acute myeloid leukemia (AML). The training set contains 57 leukemiasamples (20 ALL, 17 MLL, and 20 AML), whereas the testset

contains 15 samples (4 ALL, 3 MLL, and 8 AML). Thedimension of the MLL data set is 12,582.

We select the best genes for each of the data sets. These genes are selected using the training samples only. They use ELM algorithm to select the best features from the training samples.

Table 1Summary of the Data Sets Used in the Experimentation

Data Set	Class	Dimension (Number of genes)	Training Sample	Test Sample
SRBCT	4	2308	63	20
MLL	3	12582	57	15

In this above table for the SRBCT data set, the classification accuracy ranged between 85 and 100 percent; for the MLL data set, the classification accuracy ranged between 80 and 100 percent; and for the Prostate Tumor data set, the classification accuracy ranged between 76.47 and 100 percent.

It may have number of algorithms they selected the top-4 genes**.**

Table 2Comparison of the Methods on the SRBCT Data Set

Methods (Feature Selection+Classification)	**No.of Selected Genes**	**Classification**
Information Gain+SVM	**350**	**50%**
Proposed Feature Selection+SVM	**200**	**63%**
Information Gain+ELM	**150**	**70%**
Proposed Feature Selection+ELM	**4**	**93%**

A. Testing Accuracy

Table 3Comparison of Testing Accuracy

NO OF GENES COMBINATIONS	**ACCURACY**	
	SVM	**ELM**
100, 2	87.12	89.54
100, 3	86.12	89.14

This table 3 shows the accuracy for SVM and ELM. It is clear from the above table the proposed approach ELM gives improvedaccuracy than SVM.

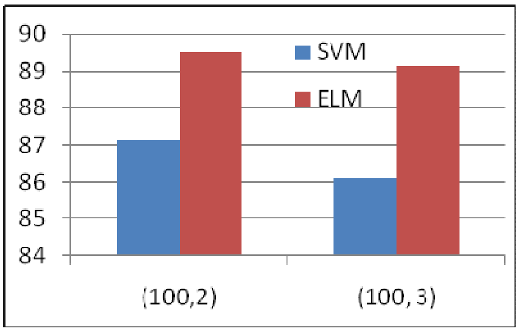

Fig2. Chart for Comparison of Testing Accuracy

This chart shows that the accuracy of ELM is greater than SVM using testing datasets.

4 CONCLUSION

The thousands of gene in microarray data that can directly contribute to determined the class membership of each pattern. The results confirmed that the ELM classifier is a promising candidate for improving Accuracy and Minimum Sensitivity. This research focuses on the approach for developing a cancer classification system by proposed ELM techniques and its training is lesser in time. Extreme Learning Machine is a significantly fast rather than using a gradient descent method and also back propagation and SVM technique. The Algorithm of proposed ELM for feature selection looks simpler than most training methods by using the SRBCT and MLL datasets. Finally, ELM technique has better generalization performance than SVM and BP.

REFERENCES

[1] Gao, X., and L. Huang. "Cationic liposome-mediated gene transfer." *Gene therapy* 2.10 (1995): 710-722.

[2] Getz, Gad, Erel Levine, and EytanDomany. "Coupled two-way clustering analysis of gene microarray data." *Proceedings of the National Academy of Sciences* 97.22 (2000): 12079-12084.

[3] Wang, Joseph. "Survey and summary from DNA biosensors to gene chips." *Nucleic Acids Research* 28.16 (2000): 3011-3016.

[4] Ambroise, Christophe, and Geoffrey J. McLachlan. "Selection bias in gene extraction on the basis of microarray gene-expression data." *Proceedings of the National Academy of Sciences* 99.10 (2002): 6562-6566.

[5] Liu, Bing, Chunru Wan, and Lipo Wang. "Unsupervised gene selection via spectral biclustering." *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*. Vol. 3. IEEE, 2004.

[6] Shyamsundar, Radha, *et al*. "A DNA microarray survey of gene expression in normal human tissues." *Genome biology* 6.3 (2005): R22.

[7] Tseng, Vincent S., and Ching-Pin Kao. "Efficiently mining gene expression data via a novel parameterless clustering method." *Computational Biology and Bioinformatics, IEEE/ACM Transactions on* 2.4 (2005): 355-365.

[8] Wang, Dianhui, and Guang-Bin Huang. "Protein sequence classification using extreme learning machine." *Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on*. Vol. 3. IEEE, 2005.

[9] Lee, Jaewook, and Daewon Lee. "Dynamic characterization of cluster structures for robust and inductive support vector clustering." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 28.11 (2006): 1869-1874.

[10] Yang, Wen-Hui, Dao-Qing Dai, and Hong Yan. "Biclustering of microarray data based on singular value decomposition." *Emerging Technologies in Knowledge Discovery and Data Mining*. Springer Berlin Heidelberg, 2007. 194-205.

[11] Mukhopadhyay, Anirban, UjjwalMaulik, and Sanghamitra Bandyopadhyay. "Simultaneous informative gene selection and clustering through multi objective optimization." *Evolutionary Computation (CEC), 2010 IEEE Congress on*. IEEE, 2010.

[12] Sharma, Alok, et al. "Strategy of finding optimal number of features on gene expression data." *Electronics letters* 47.8 (2011): 480-482.

[13] Wan, Shibiao, Man-WaiMak, and Sun-Yuan Kung. "mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines." *BMC bioinformatics* 13.1 (2012): 290.

[14] Sharma, Alok, et al. "Null space based feature selection method for gene expression data." *International Journal of Machine Learning and Cybernetics* 3.4 (2012): 269-276.