International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 7 (2015) pp. 16865-16872

http://www.ripublication.com

Enhanced Support Vector Machine (SVM) Algorithm For Classify Microarray Data

Dr M. Hemalatha and S.Kavitha

Associate Professor, Department of Computer Science,
Karpagam University, Coimbatore.
Research Scholar, Department of Computer Science,
Karpagam University, Coimbatore.

ABSTRACT

Recent technological developments in molecular biology have made it possible to measure the expression level of thousands of genes simultaneously. Advances in knowledge discovery have greatly improved the study of microarray data analysis and its roles in the disease identification and prediction. Usage of artificial neural networks, a part of machine learning techniques of data mining, is being extensively analyzed to extract interesting and relevant information from huge sized microarray data. In this paper, an enhanced filter-wrapper based gene selection algorithm is combined with an optimized Support Vector Machine (SVM) is proposed to classify microarray data. The working of SVM is optimized through the use of wavelet neural networks. Experimental results showed that all the proposed gene selection algorithm and the optimization operation included to enhance SVM has improved the classification performance and produces maximum efficiency with respect to classification accuracy.

INTRODUCTION

Molecular diagnostics provide insights into disease mechanisms by using new gene-based biomarkers. Existing disease diagnostic systems rely on indirect indicators that permit only general classifications of known diseases and do not take into account the alterations in individual gene expression. Analysis of microarray data allows the study of gene expression to provide way for simultaneous analysis of thousands of genes in an efficient manner (Mazandu *et al.,* 2011). The result of such analysis offers numerous opportunities to obtain insight on the state of activity of diseased cells and patient samples.

Microarray data is an arrangement of points in rows and columns. It is an extension of the concept and constitutes a very small arrangement of many points in

rows and columns. It is a collection of microscopic DNA spots attached to a solid surface such as a glass, plastic or silicon chip. It is also known as DNA Chips or Gene Chips (Alshamlan *et al.*, 2013). Microarray data analysis provides invaluable information on disease pathology, progression, resistance to treatment, and response to cellular microenvironments and helps to improve early diagnosis and innovative therapeutic approaches for cancer. As microarray technology is capable of producing massive amounts of genetic data, the need for new or enhanced techniques that can mine and discover biologically meaningful knowledge in large data sets is the current need of molecular biologists. Many researchers are using knowledge discovery techniques (data mining) for this purpose. Knowledge discovery or data mining is an interdisciplinary field that involves statistics, computer science and database management for efficient extraction of common patterns hidden behind large microarray data (Tarca *et al.*, 2013).

A typical microarray data analysis consists of three steps, namely,
(i) Identification of genetic defect(s)
(ii) Diagnostics steps to identify the type of disease
(iii) Usage of preventive medicine

Out of the above three steps, the most challenging step is the identification of genetic defects, because of the huge amount of data to be analyzed. For this purpose, several solutions have been proposed, which can be grouped into algorithms as listed below.
(i) Feature or variable selection algorithms (Identification of "marker" genes that characterize the different tumor classes)
(ii) Knowledge discovery using unsupervised (clustering) algorithms (Identification of new/unknown tumor classes using gene expression profiles)
(iii) Knowledge discovery using supervised (classification) algorithms (Classification of sample into known classes)

Of the above, this paper focuses on the third category of knowledge discovery, namely, classification. Classification has been used by several researchers to improve the performance of the classifier (De Paz *et al.*, 2011; Chen, 2012). Inspite of its wide usage, there still exists several issues which make this area an active research field.

A classification model for microarray data analysis consists of three major steps, namely, preprocessing, gene selection and identification/prediction of genetic defect. In this paper, methods that enhance the operation of the second and third step are proposed. For this purpose, a hybrid filter-wrapper method for gene selection is combined with an optimized Support Vector Machine (SVM) classification model is proposed to improve the microarray data analysis. The rest of the paper is organized as follow: Section 2 presents the proposed methodology, the results of the various experiments conducted to analyze the performance of the proposed algorithm is presented in Section 4. The work is concluded with future research directions in Section 5.

PROPOSED METHODOLOGY

Most of the classification models proposed for analysis of microarray data, try to enhance the process of classification to maximize accuracy on test data and perform feature selection and enhancement of classification as two separate steps. In this paper, an alternative method is analyzed, where the feature selection step is used to improve the performance of the classifier. Specifically, the accuracy of the classifier is improved through the identification of relevant features obtained through enhanced gene selection techniques. Thus, this paper proposes an integrated approach that combines gene selection with classification.

Gene Selection

The gene selection step is used to identify the main disease genes from thousands of other regular genes in the microarray data. This step is also used to remove redundant and irrelevant data. Gene selection from microarray data is performed using filter based algorithms (Meyer *et al.*, 2008) or wrapper based algorithms (Samb *et al.*, 2012).The main disadvantage of filter approaches is the fact that they ignore the effect of the selected feature subset on the performance of the classifier to be used afterwards. Filters based method select genes simply based on the statistical scores of genes and do not take into account the gene interactions when selecting genes. On the other hand, gene sets selected by wrappers are usually with best discriminative potential for a fixed classifier selected. Though wrappers take into account the interactions between gene subset, finding such an optimum gene set requires high computational cost. To solve these issues, this paper proposes a hybrid gene selection method that combines the advantages of filter and wrapper to obtain an optimized set of genes that can improve the process of classification.

The hybrid model is built using F-Score method (filter based approach) and information gain method (wrapper based approach) (Quinlan, 1979). These algorithms were selected because of its proven success in feature selection domain. F-score is a filter model that uses the discriminative ability of each feature to identify relevant genes and features with higher F-score have better separation ability in classification problems. This feature makes it very suited for the SVM, as SVM also tries to find an optimal hyper-plane to separate two classes. However, the F-score method can examine only the discriminative ability of each individual feature and cannot identify the discriminative ability of multiple features. Hence, features with low scores will be disregarded, even if they are complementary to the top features and might be very useful. Therefore, the hybrid model also uses the information gain (IG) algorithm. IG selects features with the amount of information each feature can provide. Thus, the hybrid gene selection algorithm uses the high discriminative feature of F-score method and information gain obtained by each feature using IG method.

Usage of F-Score and IG methods, results with two feature subsets. These features are considered as the most class-related features from all features. Using all these features together as the final feature set will result in two drawbacks. (i) Increased the time complexity of the classifier and (ii) Reduced classification accuracy. Both the problems can be solved by an effective combining procedure,

which should aim at reducing redundant data. In the proposed approach, the two feature sets are combined using intersection (AND) operator. This step removes the most redundant and irrelevant features and retains only useful features.

Enhanced SVM

Support Vector Machines (SVMs) are state-of-the-art models which have been used to solve many problems. Their success is due to following three factors.

- They are maximum-margin classifiers.
- The dual form of the SVM optimization problem is quadratic and its computational complexity depends on the number of data, not on their dimensionality.
- Since the optimization of their dual form only needs the inner products of data points and not the data points themselves, kernels can easily be plugged into SVMs.

When using SVMs, three choices must be made, namely, the kernel type, the kernel parameters and the regularization parameter. These choices are critical for the quality of the results and are usually done using cross-validation. However, this process can be computationally intensive since many models have to be built. To solve this issue, this study uses ELM algorithm, the result of which is then used as kernel to SVM. The proposed hybrid model achieves two important gains, namely, fast to train (obtained through the use of ELM) and maximum-margin classifiers (obtained through the use of SVM). Steps involved in the proposed ELM-SVM are presented in Figure 1.

Integrated Approach

The integrated approach for classifying microarray data is presented in Figure 2.

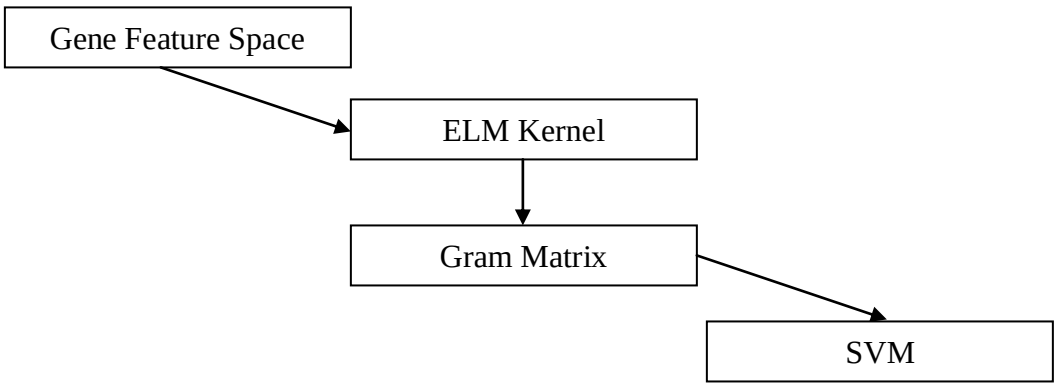

Figure 1 : ELM-SVM Classification Model

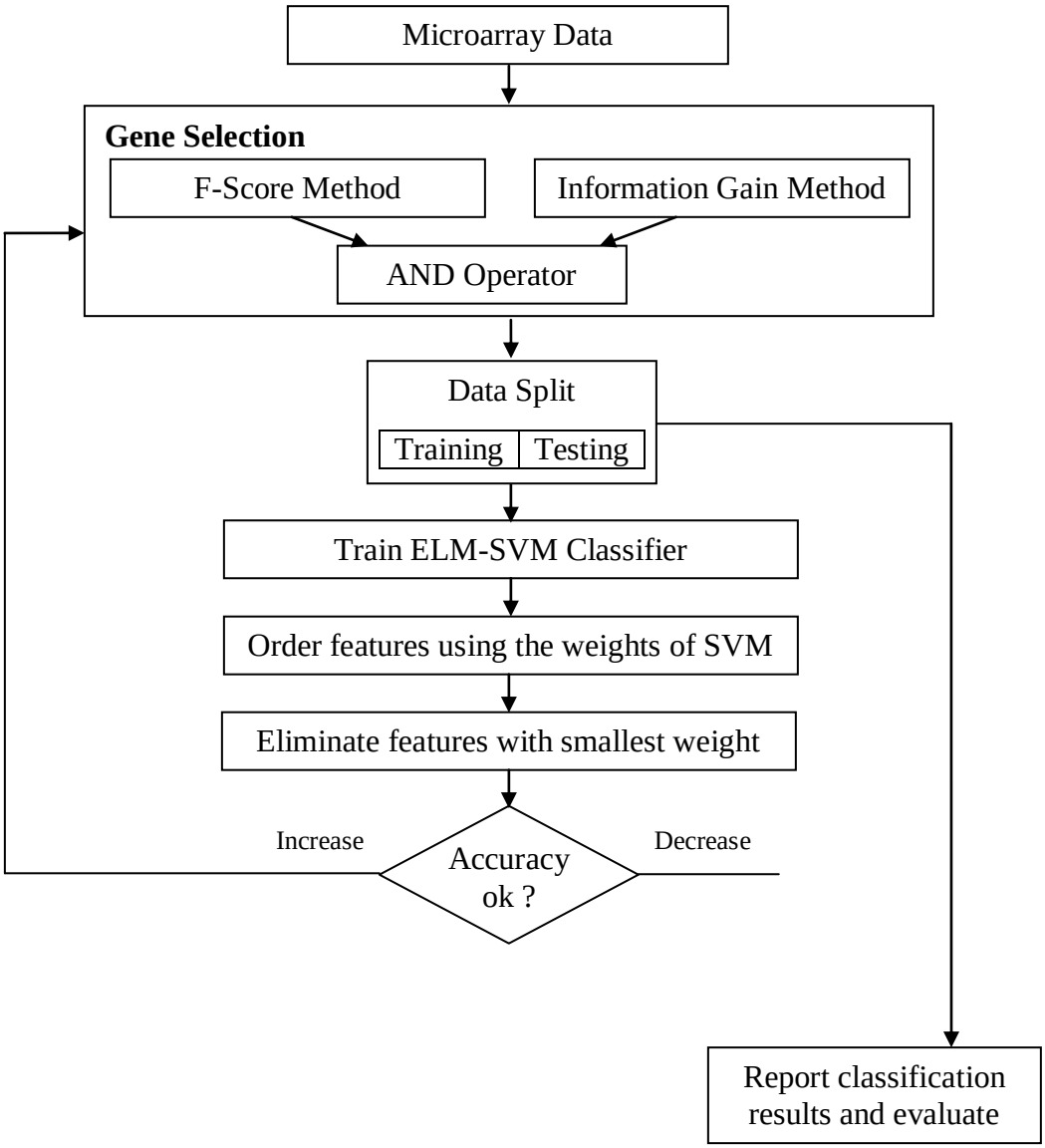

Figure 2 : Proposed Integrated Classification Approach

This approach begins by selecting optimal genes using the hybrid gene selection algorithm (Section 2.1). The resultant feature set is split into training (70%) and testing (30%) set using hold-out method. The holdout method randomly partitions the dataset into two independent sets, training and testing. The training set is then used to train the ELM-SVM classifier. A feature removal algorithm is performed based on the SVM weights. The features are first ordered using SVM weights. Features with smaller weights are removed to form an optimized feature set. The ELM-SVM classifier is then evaluated. If the accuracy increases, the new feature set is again used as input to gene selection step and the whole process is performed again. If it decreases, the removed feature set is inserted into the feature set and the process

stops. Then the testing dataset is used to evaluate the proposed ELM-SVM model.

EXPERIMENTAL RESULTS

Several experiments were conducted using two datasets, 3-class leukemia dataset (Golub *et al.*, 1999) and Breast Cancer Dataset (van't Veer *et al.*, 2002). The leukemia dataset is available at http://www.genome.wi.mit.edu. The second dataset was downloaded from http://www.rii.com/publications/2002/vantveer.htm. The experiments were designed to study the impact and importance of gene selection on classification. This was carried out by comparison of classification performance while using all genes (without gene selection) and while using the proposed gene selection algorithms integrated with the classification task. Two parameters, namely, accuracy and error rate, were used during performance evaluation. The results pertaining to accuracy and error rate are presented in Figures 3 and 4 respectively.

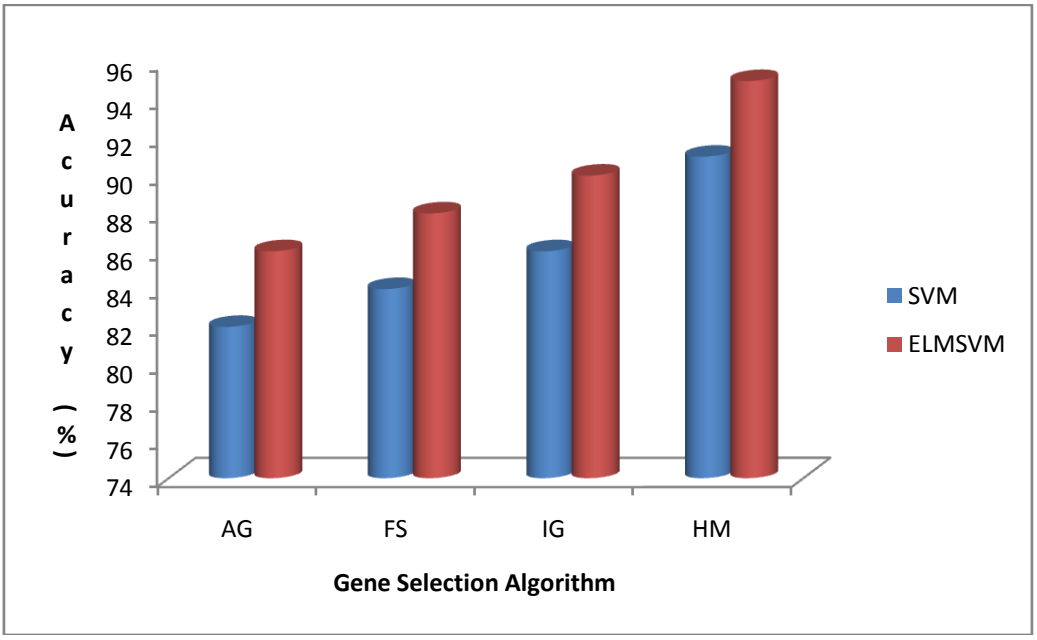

Figure 3: Classification accuarcy

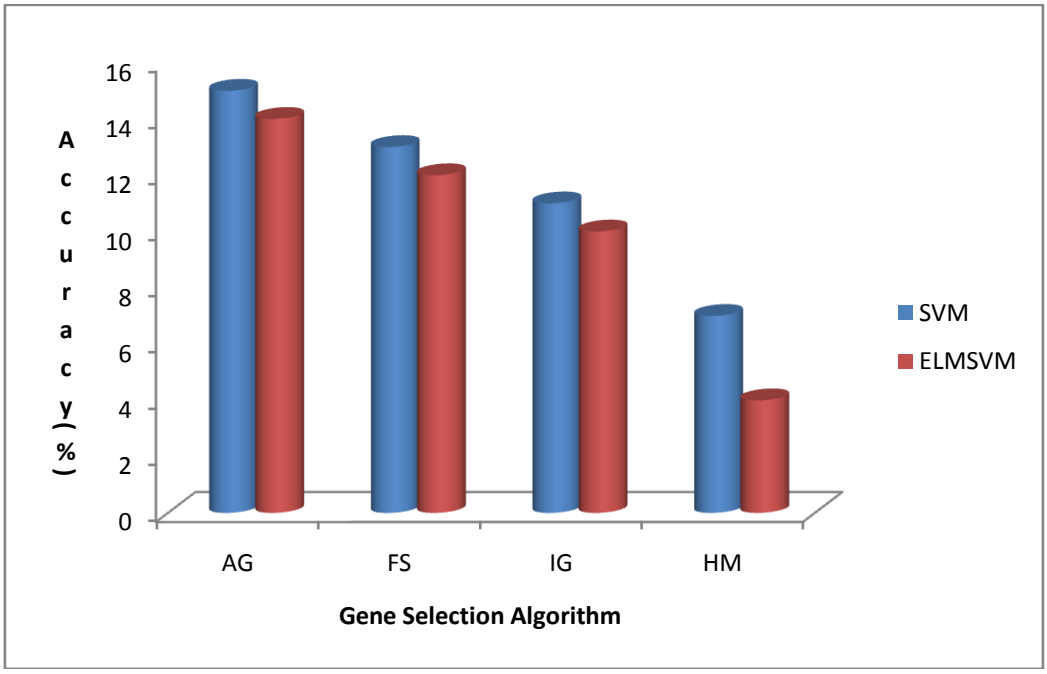

Figure 4: Error Rate of Classification

The results clearly show that the proposed integrated approach combining gene selection algorithm with ELM-SVM classifier, has improved the microarray data classification performance. The result show that both accuracy and error rate of microarray classification has improved by the usage of the integrated approach when compared with the existing approaches.

CONCLUSION

Microarray datasets tend to be small in sample size due to the cost associated with the assays. There are many more gene expression measurements (e.g. 54,000 transcripts) available than samples. This huge number makes gene selection a crucial step while designing and building a classifier. In this research work, the gene selection method is integrated with knowledge discovery algorithm (classification) is proposed and analyzed. The method propose an enhanced gene selection method that combines the advantages of filter and wrapper based method, which are then combined with ELM-SVM classifier. The process of SVM is enhanced through the usage of ELM kernels. Experimental results proved that all the proposed method are efficient and enhance the process of microarray data classification and can safely be used by molecular biologists during meaning knowledge discovery and analysis.

REFERENCES

[I] Mazandu, G.K., Opap, K. and Mulder, N.J. (2011) Contribution of microarray data to the advancement of knowledge on the Mycobacterium tuberculosis interactome: use of the random partial least squares approach, Infect Genet Evol., Vol. 11, No. 4, Pp. 725-33.

[II] Alshamlan, H.M., Badr, G.H. and Alohali, Y. (2013) A study of cancer microarray gene expression profile: Objectives and approaches, Proceedings of the World Congress on Engineering, Vol. II, Pp. 1-6.

[III] Tarca, A.L., Lauria, M., Unger, M., Bilal, E., Boue, S., Dey, K.K., Hoeng, J., Koeppl, H., Martin, F., Meyer, P., Nandy, P.,Norel, R., Peitsch, M., Rice, J.J., Romero, R., Stolovitzky, G., Talikka, M., Xiang, Y., Zechner, C. and Improver DSC Collaborators (2013) Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge, Oxford Journals, Life Sciences & Mathematics & Physical Sciences Bioinformatics, Vol.29, No.20, Pp.1-8.

[IV] De Paz, J.F., Bajo, J., Vera, V., Corchado, J.M. (2011) MicroCBR: A case-based reasoning architecture for the classification of microarray data, Applied Soft Computing, Vol. 11, Issue 8, Pp. 4496-4507.

[V] Chen, C.K. (2012) The classification of cancer stage microarray data, Computer Methods and Programs in Biomedicine, Vol. 108, Issue 3, Pp. 1070-1077.

[VI] Meyer, P.E., Schretter, C. and Bontempi, G. (2008) Information-theoretic feature selection in microarray data using variable complementarity, IEEE Journal of Selected Topics in Signal Processing, Vol. 2, No. 3, Pp. 261-274.

[VII] Samb, M.L., Camara, F., Slimani, Y. and Esseghir, M.A. (2012) A Novel RFE-SVM-based Feature Selection Approach for Classification, International Journal of Advanced Science and Technology Vol. 43, Pp. 27-36.

[VIII] Quinlan, J.R. (1979) Expert systems in the microelectronic age, Edinburgh University Press, Scotland: Edinburgh, Pp. 168–201.

[IX] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., 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, Vol, 286, Pp. 531-537.

[X] van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002) Gene expression profiling predicts clinical outcome of breast cancer, Nature, Vol. 415, Pp. 530–536.