DEVELOPMENT OF AN EFFICIENT DATA MINING CLASSIFIER WITH MICROARRAY DATA SET FOR GENE SELECTION AND CLASSIFICATION

[1]A. SUMATHI, [2]S. SANTHOSHKUMAR AND [3]Dr. N. K. SAKTHIVEL

[1]School of Computing, SASTRA University, Tamil Nadu, INDIA
[2]Department of Computer Science, Government College Kumbakonam(A), Tamil Nadu, INDIA
[3]Professor, School of Computing, SASTRA University, Tamil Nadu, INDIA

Email : [1]sumathi_vijayalakshmi@yahoo.co.in, [2]santhoshsundar@yahoo.com, [3]sakthi@cse.sastra.edu

ABSTRACT

Microarray sample classification has been studied extensively using classification techniques in machine learning and pattern recognition. In a microarray chip, the number of genes available is far greater than that of samples, which is a serious problem and the gene expression reduction, is important one. Prior to sample classification, it is important to perform gene selection and more interpretable genes to be identified as biomarkers, so that a more efficient, accurate, and reliable performance in classification can be achieved. For this purpose, a hybrid scheme was proposed and this scheme is called as Single Filter – Single Wrapper Classifier (SFSW). In this technique, the Filter approach is used to select the data sets from the large Microarray and the Wrapper approach is used to classify the gene expressions from the selected data sets. From the available statistical report, it is revealed that the Filter method has a fast dimensionality reduction step for selecting a small set of genes at the cost of accuracy and the Wrapper method used for improving classification accuracy on this small set of genes at the cost of computational delay. However this approach also has some problems such as different filtered subset leads to complex evaluation. Hence an Efficient Hybrid Classifier called "Multiple-Filter-Multiple-Wrapper Technique" has been proposed with improved performance of SFSW. The use of Multiple Filters with different filter metrics ensures that useful biomarkers are unlikely to be screened out in the initial filter stage. The use of Multiple Wrappers is intended to optimize the reliability of the classification by establishing consensus among several classifiers. However, in this research work, we have identified that the Classification Accuracy of MFMW is poor due to the consideration of Indecisive Prediction Status. Hence, this Research Work is introduced an efficient classifier called *ICS4-MFMW*, which is focusing both the dimensionality reduction and Indecisive Status. This work is demonstrated its efficiency in terms of classification accuracy with Sensitivity and Specificity. From the result, it is revealed that our proposed work perform better as compared with the existing MFMW Classifier.

Key Words: *Microarray, Pattern Recognition, Filter, Wrapper, Classifier, Sensitivity and Specificity*

1. INTRODUCTION

A DNA microarray is a multiplex technology [1, 2, 6, 7, 8] used in molecular biology. It consists of arrayed series of thousands of microscopic spots. DNA micro array allows us to analyze thousands of genes in one experiment. The DNA Micro array technology allows investigators to observe simultaneously the expression behaviour of all the genes within an entire genome and can provide information on gene functions. There are many approaches available to select a gene from a genome. The Gene Selection methodology [1,3,15] based on Microarray is shown in the Figure. 1.

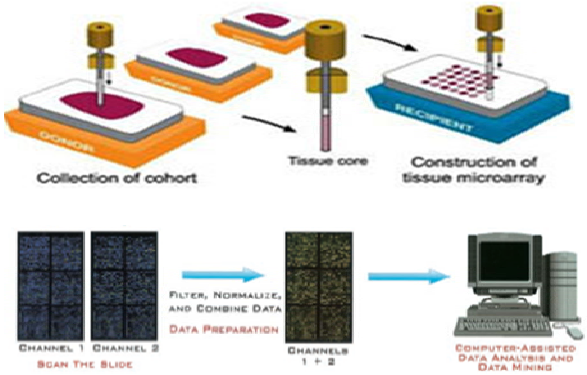

DONOR
Collection of cohort
Tissue core
RECIPIENT
Construction of tissue microarray
CHANNEL 1
CHANNEL 2
SCAN THE SLIDE
FILTER, NORMALIZE, AND COMBINE DATA
DATA PREPARATION
CHANNELS 1 + 2
COMPUTER-ASSISTED DATA ANALYSIS AND DATA MINING

Figure 1 Microarray Gene Selection

However, most genes in a microarray give little benefits to the sample classification problem. Therefore, prior to sample classification, it is important to perform gene selection whereby more interpretable genes are identified as biomarkers [11,13] so that a more efficient, accurate, and reliable performance in classification can be expected. There are, in general, two approaches to gene selection, namely Filters and Wrappers. Although gene selection using Filters are simple and fast, the method has several shortcomings, For this purpose, a hybrid scheme was proposed and this scheme is called as Single Filter – Single Wrapper Classifier (SFSW)[1,14]. In this technique, the Filter approach is used to select the data sets from the large Microarray and the Wrapper approach is used to classify the gene expressions from the selected data sets. From the available statistical report, it is revealed that the Filter method has a fast dimensionality reduction step for selecting a small set of genes at the cost of accuracy and the Wrapper method used for improving classification accuracy on this small set of genes at the cost of computational delay. The SFSW approach however has its own difficulties:

- Different Filters yield different filtered subsets that may leave out some relevant biomarkers which consequently do not have a chance to be considered in the wrapper evaluation.
- Different wrappers will select different genes from the filtered set despite achieving the same training accuracy.
- Some SFSW models are better than the others in terms of attaining the required training accuracy.

Hence an Efficient Hybrid Classifier called “Multiple-Filter-Multiple-Wrapper Technique” has been proposed which improved the performance of SFSW.

2. RELATED WORKS

In this section, we will describe some filter metrics and classifiers. Also, we will illustrate some widely used ensemble method and the use in incremental induction task.

2.1 Introduction To MFMW

MFMW [1,2,5,8] hybrid model is a generalization of the SFSW model, which is shown in the Figure.2. This is using multiple filters to select genes and then combining them to provide a merged filtered subset of genes. The use of multiple filters with different filter metrics ensures that useful biomarkers are unlikely to be screened out in the initial filter stage.

The use of multiple wrappers is intended to enhance the reliability of the classification by establishing consensus among several classifiers. As a result, there is some kind of consensus among the different classifiers in the wrapper step as to which genes should be selected. Hence, the final genes selected can be considered to be more robust with a mixture of characteristics that fit several wrappers, and are therefore better qualified as biomarkers.

Furthermore, since the MFMW model already incorporates the characteristics of multiple filters and wrappers, it is no longer necessary to try different filter-wrapper combinations in order to search for a suitable combination that yields the highest classification accuracy.

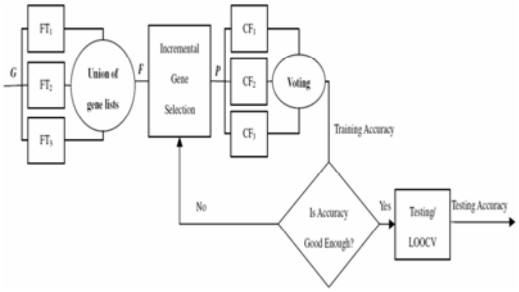


Figure 2 MFMW for gene selection

This MFMW model provides predictive accuracies that are either comparable or better than the best existing results obtained using all available SFSW methods.

2.2 Filter Metric

Gene expression data have thousands of features, it is nearly impossible to represent and understand their complex relationships directly. One way to do so is by filtering, whereby the “goodness” of a gene is evaluated by measuring the relationship between gene expression and the class label using statistical techniques. In this work, the authors Yukyee Leung and et.al. adopt the SNR (Signal-to-Noise Ratio) feature selection to reduce the dimensionality of the data. Signal-to-noise ratio is a measure of how the defected signal compared with the other background noise. In bio informatics signal refers to useful information conveyed by a gene, and noise to anything else on the gene. Three of the most commonly used metrics are

- *SNR*
- *TS and*
- *Pearson Correlation Coefficient (PC)*

$$SNR = 10\log_{10}\left(\frac{P(signal)}{P(noise)}\right) \qquad \text{.........(1)}$$

2.3 Classifier

The aim of supervised classification is to develop a decision rule to discriminate between samples of different classes based on the gene expression profile. Discovery of significant classification rules to accomplish the classification task is suitable for bio-medical research. Two widely used classifiers are considered, namely

- *k-NN and*
- *SVM*

k-NN classifies samples based on closest2 training examples in the gene space. All training samples are mapped into a multidimensional gene space and it is partitioned into regions by class labels of the training samples. The purpose of this algorithm is used to classify a new object based on training samples.

Support Vector Machine[1,4,10,12,15] (SVMs) is a set of related supervised learning methods that analyze data and recognize patterns, used for classification. Support Vector Machine (**SVM**) is a non-linear **classifier** method which is often reported as producing better classification results compared to other methods. The main idea of this method is to non-linearly map the input sample data to some high dimensional space, where the data can be linearly separated, thus providing higher classification (or regression) accuracy. SVM then automatically discovers the optimal separating hyper plane. SVMs are rather interesting in that they enjoy both a sound theoretical basis as well as state-of-the-art success in real-world applications.

2.4 Decision Tree Induction

Decision trees are usually used for gaining information for decision making. For inductive learning, decision tree is attractive for the following reasons:

- Decision tree is good for unobserved instance, only if the instances are described in terms of features that are correlated with the target concept.
- The decision tree methods are efficient in computation that is proportional to the number of observed training instances.
- The result of decision tree provides a representation of the concept that understandable to human. The two types of algorithms used in decision trees are incremental and non – incremental decision trees.

Non Incremental decision algorithm is a type of decision tree algorithms which derives its classes from a fixed set of training instances. The various non–incremental Classifiers are ID3, C 4.5 and C5.0. These classifiers can't be used to improve the accuracy. Thus Incremental Decision Trees are built, which are giving higher accuracy. A few Incremental Decision Trees are ITI and ICS4.

3. IDENTIFIED PROBLEMS

From the Previous Section, it is observed that various Classification Mechanisms such as kNN, SVM, ITI, MFMW and ICS4 have been proposed to improve the performance of classification accuracy. The Multiple Filter Multiple Wrapper (MFMW) approach is intended to optimize the classification accuracy and from our experimental results, it is observed that this MFMW achieves higher accuracy. This MFMW however considers a few indecisive status. ie this MFMW couldn't classify either as True Negative or True positive for few samples, instead it is classifying those samples as indecisive. This is the major drawback and issue of this system. Thus this research work introduced an efficient classifier called ICS4-MFMW, which is improving the classification accuracy of MFMW.

4. PROPOSED TECHNIQUE

As stated in the previous section, to overcome the identified problem of MFMW, this work is introduced an efficient hybrid new classification method called ICS4-MFMW which improved the classification accuracy in terms of Specificity and Sensitivity.

4.1 Use of Cascading-and-Sharing method in Incremental Tree Induction with MTMW

Among many classification approaches, using classification rules derived from the decision tree induction may helpful to perform this work and previous studies show that it is powerful. We define a rule as a set of conjunctive conditions with a predictive term. The general form of rules is presented as:

IF condition$_1$ & condition$_2$ & ... & condition$_m$, THEN a predictive term

The predictive term in a rule refers to a single class label. For useful clinical diagnosis purpose, using those rules we can address issues in understanding the mechanism of a disease and improve the discriminating power of the rules. A significant rule is one with a largest coverage which the coverage satisfies a given threshold. For example, the given threshold is 60%, if one rule's coverage is larger than 60% then it is called significant rule. However, using traditional ensemble method to build and refine the tree committee and derive significant classification rules is still impossible. So, this work is taken the concept of ICS4, which is a new incremental decision learning algorithm which uses the skeleton of ITI and accepts the cascading and sharing ensemble method of CS4 to break the constraint of singleton classification rules by producing many significant rules from the committees of decision trees and combine those rules discriminating power to accomplish the prediction process.

$$Mean = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) \quad \text{………. (2)}$$

$$Radius = \frac{\sqrt{\sum_{i}^{n} (x_i - x_o)^2}}{n} \quad \text{…………(3)}$$

Based on this idea, this work is implemented the ICS4 with MFMW and MFMW separately. This software framework of this work has been developed with Java Programming Language and BioWEKA and it has the following features and modules.

- Gene/Cancer classification, which is used to find Cancer Pattern.
- Cancer Intensity Prediction, which is used to measure the Accuracy and Error Rate of this proposed and existing system.

For the Gene Classification[9], this work constructed Confusion Matrix for both the MFMW-SVM and MFMW-ICS4. From the confusion matrix, the Specificity, Sensitivity, Accuracy Rate and Error rate have been calculated. For measuring accuracy rate and Error Rate, the following mathematical model is used.

$$\text{Accuracy Rate} = \frac{T_P + T_N}{T_P + T_N + F_P F_N} \quad \text{………. (4)}$$

$$\text{Error Rate} = \frac{F_P + F_N}{T_P + T_N + F_P F_N} \quad \text{……….. (5)}$$

4.2 Performance Analysis

This work studied for both the existing and proposed technique thoroughly. For study, our work used DNA microarray data sets that have been used in the diagnosis of COL62 cancer.

For this experiment, this work has taken 18 normal data sets and 18 tumor data sets and each containing 7457 data.

The predicted results of MFMW-SVM classifier and MFMW-ICS4 classifiers are shown in the Figure 3 and Figure 4. The MFMW-SVM is compared the MFMW-ICS4 for the Cancer Pattern COL62 with the corresponding Normal Pattern and the classification accuracy of these classifiers have been shown in the Figure 5.

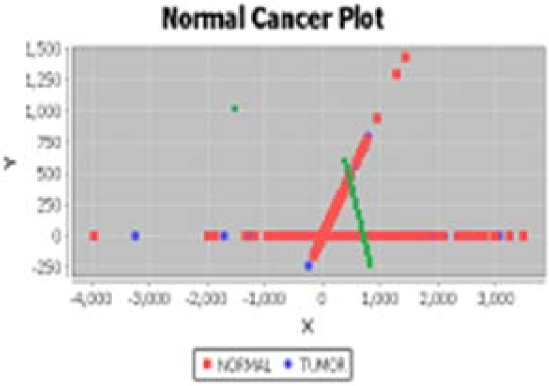


Figure 3 Gene Classification using MFMW-SVM

Dataset	Classifier	Accuracy Rate	Error Rate
COL 62	MFMW-SVM	95.224265	.04775733
	MFMW-ICS4	96.794664	.03205355

Table 1 : Classification Accuracy and Error Rate

The confusion matrix of MFMW-SVM and MFMW-ICS4 Classifiers are shown in the Figure 6 and Figure 7. The Accuracy Rate and Error Rate of the proposed Classifier are compared with existing classifier MFMW-SVM which is shown in the

Table 1.

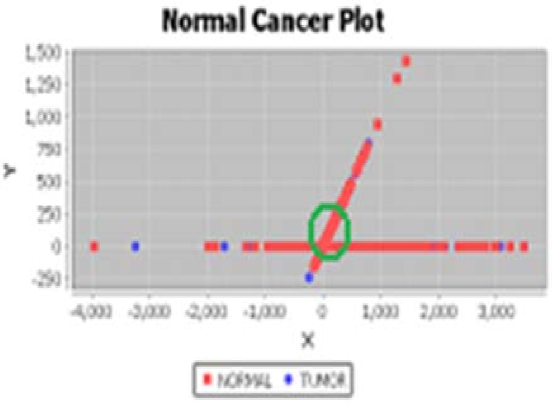


Figure 4 Gene Classification using MFMW-ICS4

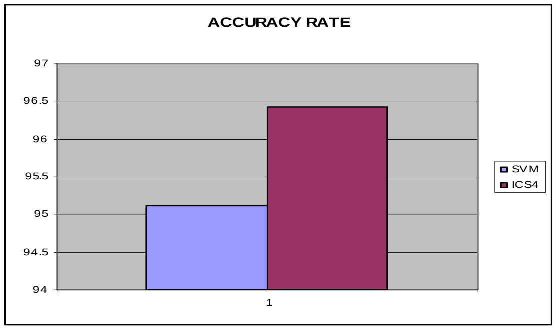
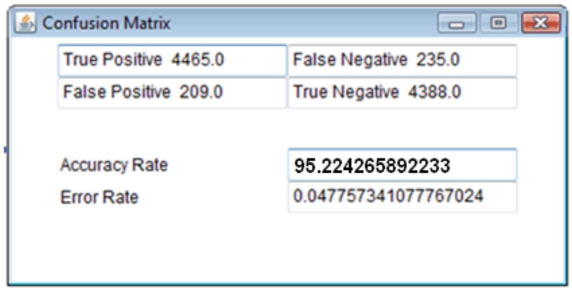
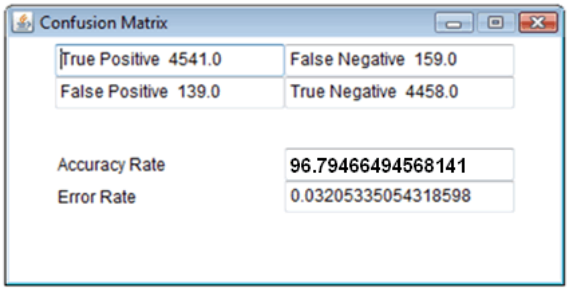


Figure 5 Comparison of Accuracy of MFMW-SVM and MFMW-ICS4


Confusion Matrix

True Positive 4465.0	False Negative 235.0
False Positive 209.0	True Negative 4388.0
Accuracy Rate	95.224265892233
Error Rate	0.047757341077767024

Figure 6 Confusion matrix of MFMW-SVM


Confusion Matrix

True Positive 4541.0	False Negative 159.0
False Positive 139.0	True Negative 4458.0
Accuracy Rate	96.79466494568141
Error Rate	0.03205335054318598

Figure 7 Confusion Matrix ICS4

The performance of the MFMW-SVM and our proposed Classifier, MFMW-ICS4 is measured interms of Classification Accuracy and Error Rate, which are shown in the Figure.5. From our experimental result, it is revealed that our proposed work giving more accuracy in terms of sensitivity, specificity, Accuracy and Error Rate as compared with existing.

5. CONCLUSION

Microarray sample classification has been studied extensively using MFMW-SVM and MFMW-ICS4. From our experimental result, it is established that our proposed work achieves high classification accuracy interms of Sensitivity, Specificity, Accuracy and Error Rate.

REFERENCES

[1] Yukyee Leung and Yeungsam Hung, "A Multiple Filter Multiple Wrapper to gene selection and microarray data classification", IEEE/ACM Transcations computational Biology and Bioinformatics, VOL. 7, NO. 1, JANUARY-MARCH 2010.

[2] Minghao Piao, Jong Bum Lee, Khalid E.K. Saeed, and Keun Ho Ryu, Discovery of significant classification rules from Incrementally inducted decision tree ensemble for diagnosis of disease". 2009.

[3] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. chummer, and D. Haussler, "Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data," Bioinformatics, vol. 16, no. 10, pp. 906-914, 2000.

[4] Li, J.Y., Liu, H.A., Ng, S.-K., Wong, L.: See-Kiong Ng, Limsoon Wong: Discovery of significant rules for classifying cancer diagnosis data. Bioinformatics 19, 93–102 (2003)

[5] Bogdan Done, Purvesh Khatri, Arina Done, and Sorin Dr_aghici, “Predicting Novel Human Gene Ontology Annotations Using Semantic Analysis,” IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 7, No. 1, pp. 91-99, January-March 2010.

[6] Santanu Ghorai, Anirban Mukherjee, Sanghamitra Sengupta and Pranab K. Dutta, “Cancer Classification from Gene Expression Data by NPPC Ensemble,” IEEE Transactions On Computational Biology And Bioinformatics, pp. 1-6, 2009.

[7] Steen Knudsen, Medical Prognosis Institute, “Cancer Diagnostics with DNA Microarrays,” A John Wiley & Sons, Inc., Publication, 2006.

[8] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring,” Science, vol. 286, no. 5439, pp. 531-537, 1999.

[9] L. Li, C.R. Weinberg, T.A. Darden, and L.G. Pedersen, “Gene Selection for Sample Classification Based on Gene Expression Data: Study of Sensitivity to Choice of Parameters of the GA/ KNN Method,” Bioinformatics, vol. 17, no. 12, pp. 1131-1142, 2001.

[10] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler, “Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data,” Bioinformatics, vol. 16, no. 10, pp. 906-914, 2000.

[11] M.M. Xiong, L. Jin, W. Li, and E. Boerwinkle, “Tumor Classification Using Gene Expression Profiles,” Biotechniques, vol. 29, pp. 1264-1270, 2000.

[12] Y. Wang, I.V. Tetko, M.A. Hall, E. Frank, A. Facius, K.F.X. Mayer, and H.W. Mewes, “Gene Selection from Microarray Data for Cancer Classification—A Machine Learning Approach,” Computational Biology and Chemistry, vol. 29, no. 1, pp. 37-46, 2005.

[13] M. Xiong, X. Fang, and J. Zhao, “Biomarker Identification by Feature Wrappers,” Genome Research, vol. 11, pp. 1878-1887, 2001.

[14] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification. Wiley, 2000.

[15] http://www.cancer.gov/

AUTHOR PROFILES:

SUMATHI. A. received the M.Sc., Degree in Computer Science from Bharathidasan university, Tamil Nadu, India in 2005 and her M.phil., Degree in Computer sience from Periyar University, Tamil Nadu, India in 2007. She has joined at SASTRA University as an Assistant Professor in 2007. She has currently doing her M.Tech. degree in Computer Science & Engineering in SASTRA University. Her research interests include Data Mining and their applications and Computational Biology, Object oriented Programming and Image Enhancement.

SANTHOSH KUMAR. S received the M.Sc.(IT) Degree in Information Technology from Bharathidasan University, Tamil Nadu, India in 2002 and he has completed his M.Phil., degree in Computer Science from Periyar University. He is a Ph.D. Research Scholar at PRIST University, Thanjavur, Tamil Nadu. He is currently working as a Lecturer in Government College (Autonomous), Kumbakonam, Tamil Nadu. His research interests include Data Mining and their Applications, Computational Biology, Artificial Intelligence and Neural Networks.

SAKTHIVEL. N. K. received the Ph.D., Degree in Computer Science (Intelligent Routing Technique) from SASTRA University, Tamil Nadu, India in 2006. He is currently a Professor with School of Computing, SASTRA University, Tamil Nadu, India. His research interests include Next Generation Networks, High Performance QoS Routing, Wireless Sensor Networks, Bioinformatics and Computational Biology, Semantic Web Services, Data Mining and their Applications. He is the member of Computer Society of India(CSI), India, Advanced Computing and Communication Society, Bangalore. He has published more than 25 Technical Papers in International Journals and Conferences.