A Hybrid Gene Selection Technique Using Improved Mutual Information and Fisher Score for Cancer Classification Using Microarrays

M. Anidha, K. Premalatha

Abstract**—**Feature Selection is significant in order to perform constructive classification in the area of cancer diagnosis. However, a large number of features compared to the number of samples makes the task of classification computationally very hard and prone to errors in microarray gene expression datasets. In this paper, we present an innovative method for selecting highly informative gene subsets of gene expression data that effectively classifies the cancer data into tumorous and non-tumorous. The hybrid gene selection technique comprises of combined Mutual Information and Fisher score to select informative genes. The gene selection is validated by classification using Support Vector Machine (SVM) which is a supervised learning algorithm capable of solving complex classification problems. The results obtained from improved Mutual Information and F-Score with SVM as a classifier has produced efficient results.

Keywords**—**Gene selection, mutual information, Fisher score, classification, SVM.

I. Introduction

DNA microarrays are important technology for studying gene expressions. With a single hybridization, the level of thousands of genes, or even entire genome, can be estimated from a sample of cells [1]. Microarray gene expression datasets help to study the expression levels of thousands of genes simultaneously. High-dimensional data in the input space is usually not good for classification due to the curse of dimensionality [2]. Performing feature selection helps to reduce the dimension of microarray gene expression data and thus improving the computational efficiency. Specifically, feature selection removes a huge number of irrelevant genes which improves the classification accuracy**.**

Feature selection methods can be classified into three categories depending on how they combine the feature selection search with the construction of the classification model [3]. In the filter method, feature selection, and classifier design are separated. In that, a subset of features is initially selected and then selected features are fed into classifiers that are to be trained. That is, the criteria for the selection are independent of the classifier. In the wrapper approach, on the other hand, the classification method is predetermined, and the selected features are bounded to the type of classifier adopted.

The wrapper approach uses classification accuracies to rank the discriminative power of all possible feature subsets so that the selected subset is likely to produce the best performance [4]. Embedded technique searches for an optimal subset of feature which is built into the classifier construction and can be seen as a search in the combined space of feature subsets and hypotheses [1].

Due to high computational efficiency, filter methods are very popular to high dimensional data and seem to be an appropriate method in informative genes selection from high dimensional input space and low sample gene expression profile. So far, lots of filter based gene selection methods have been implemented to identify informative genes from gene expression data sets. Statistical approaches like Relief-F Correlation and Chi-square depend on actual values of microarray gene expression data and are very sensitive to noise or an outlier of the dataset. On the other hand, information on theoretic approaches like Entropy, Information Gain, and Mutual Information (MI) are effective in gene selection as they depend on the probability distribution of gene expression values rather than its actual values [1]. So, that is easy to predict the relevance between features. Among the information theoretic measures, MI is widely used because of its non-linearity, robustness, and scalability. Owing to the empirical success of MI, many promising gene selection algorithms based on MI with different parameters have been developed [5]. Fisher score is one of the most widely used supervised feature selection methods. However, it selects each feature independently according to their scores under the Fisher criterion, which leads to a sub-optimal subset of features [11].

SVMs have a lot of features that make them efficient for microarray gene expression analysis, including their flexibility in choosing a similar function, the sparseness of solution when dealing with large data sets, the ability to handle large feature spaces, and the ability to identify outliers [12]. SVM is used in this work for microarray gene expression data classification because of the above features**.**

The remainder of this paper is organized as follows; Section II reviews briefly on some of the recent work published in the area of classification of cancer using microarray gene expression values. Section III introduces and describes the general scheme of the proposed combined data mining technique. Results of the proposed technique are presented in Section IV. Section V concludes the paper.

M. Anidha is Research Scholar with the Anna University, Chennai, India (e-mail: ani_ani_dha@yahoo.com).

K. Premalatha, Dr., is with the Bannari Amman Institute of Technology, Sathyamangalam, Erode, India (e-mail: kpl_barath@yahoo.co.in).

II. Related Work

A lot of research has addressed the topic of the classification of the microarray data by using different gene selection methods with different classifiers. Golub et al., introduced a generic approach to classifying two types of acute leukemia. In that, they used SVM as a classifier, one with Locality Preserving Projection technique (LPP) and the other with F-Score ranking feature selection technique[10]. Laiwan Chan proposed a technique of gene selection based on information theoretic combining with sequential forward floating search[9]. Xiaosheng Wang et al., revealed that feature selection method based on the α-dependent degree of the attribute in rough sets were superior to the canonical-dependent degree of attribute-based method in robustness and applicability [7]. Xiaosheng and Osamu Gotoh denoted that α-dependent degree in rough sets were used for informative gene selection and decision rule based rough sets fed into the classifiers for better performance[8]. P. Ganesh Kumar et al., proposed a novel idea of computing MI in multi-stages and classified with Artificial Neural Network (ANN) [1]. Dina.A.Salem et al., stated F-Score and entropy based hybrid gene selection technique gives better performance[13]. Ghaffai et al., proposed a new technique for selecting informative genes based on computing thresholds and discriminating capabilities of genes [14]. Hala Alshamlan et al., proposed an algorithm that comprises ABC and mRMR. The new approach is based on an SVM algorithm to measure the classification accuracy for selected genes [15]. Li-Yeh Chuang proposed a hybrid method of binary particle swarm optimization (BPSO) and a combat genetic algorithm (CGA) is to perform the microarray data selection. The K-nearest neighbor (K-NN) method with leave-one-out cross-validation (LOOCV) served as a classifier which selects effective genes for better classification performance with low error rate [16]. Fei Han et al., revealed a method of clustering all the genes by k-means method and informative gene filtering is done by values of GCS (gene-to-class-sensitivity) information. Binary Particle Swarm Optimization (BPSO) method is used further to select highly sensitive genes, and the selected genes are classified according to Extreme Learning Machine [17]. Table I shows the comparison of various feature selection techniques.

III. Proposed Methodology

A. System Description

The proposed classification system receives pre-processed high dimensionality microarray data set as an input. The first step is reducing the total number of genes in the input dataset to a smaller subset using combined MI and Fisher score ranking techniques for gene selection. Then, these significant genes shall be used by the SVM for classification. At this point one can measure and record the test classification accuracy which is equal to the number of correctly classified test samples divided by the total number of introduced test samples.

B. Concept of MI

MI of two random variables is a quantity that measures the mutual dependence of the two variables (features). It is the reduction in the uncertainty of one random variable with the knowledge of the other.

The initial uncertainty about a gene A is given by the entropy H(A).

$$H(A) = -\sum_{i=1}^{n} p(a_i)\log(\ p(a_i)) \tag{1}$$

where $p(a_i)$ are the probabilities for the different values of gene A.

TABLE I
Various Feature Selection Techniques

Year	Authors	Data Sets Used	Techniques	% Accuracy
1999	T. R. Golub et al., [10]	Acute Leukaemia	LPP and F-score with SVM	100
2005	Laiwan Chan [9]	Leukaemia, Ovarian, Lung and Lymphoma	Information Theoretic with sequential forward floating search	98.87
2009	Xiao sheng Wang and Osamu gotoh [8]	Colon, CNS, Prostate, Lung, Breast and Leukaemia Tumor datasets from Kent Ridge Bio-Medical Repository	α depended degree in rough sets	91.93,91.67,98.04,100
2010	Xiao sheng Wang and Osamu gotoh [7]	Colon, CNS, Prostate, Lung, Breast and Leukaemia Tumor datasets from Kent Ridge Bio-Medical Repository	Simple Rule based system based on rough sets	91.93,91.67,98.04,100
2011	P. Ganesh Kumar and T. Aruldoss, Albert Victorie [1]	Colon cancer, Lymphoma, Prostate cancer, Leukemia, Rheumatoid Arthritis versus Control (RAC), Rheumatoid Arthritis versus Osteoarthritis (RAOA), Ovarian cancer, Breast cancer, Pancreatic cancer, and Lung cancer	Multistage MI with ANN	98.3,95.3,98.5,98.6,97.1,96.8,99.6,96.3,96.2,98.9
2011	Dina A. Salem et al., [13]	Leukemia & Lymphoma from Broad Institute of MIT	F-score with entropy based method with SVM	97,78
2012	Li-Yeh Chuang et al., [16]	Leukemia, Breast 2 class, Breast 3 class, NCI60, Adenocarcinoma, Brain, Colon, Lymphoma, Prostate, and Srbct	ABC-mRMR with SVM	98.8,93.5,99.8,99.8,99.8,100,99.6,100,99.9
2014	Hala Alshamlan et al., [15]	Colon, Leukemia, Lung, Prostate	BPSO & CGA with KNN	100,100,100,98.29
2014	Han F et al., [17]	Leukemia, colon, SRBCT, Lung, Brain cancer, Lymphoma	GCS & BPSO with ELM	100,97.3,100,96.88,86.07,85.05

The joint entropy provides the amount of relevance between two features [6] and is given by:

$$H(A,B) = -\sum_{i=1}^{n} p(a_i,b_i)\log(p(a_i,b_i)) \tag{2}$$

The MI of two features which provides a measure of the relevance between the two genes can be calculated as:

$$MI=H(A)+H(B)-H(A, B) \quad (3)$$

In this method the level of discretization of different values of each gene is improved. Many levels of discretization are introduced to improve the efficiency of MI Algorithm so that the relevance between any two features/genes is effectively improved which is efficient for selecting highly informative genes.

C. Concept of F-Score

The concept of Fisher score is represented as the distances between features. The distances between features in different classes are as large as possible while the distances between features in the same class are as small as possible.

Let μ_{i0}, μ_{i1} and σ_{i0}, σ_{i1} are the mean and standard deviation of class 0 and class 1 of the i-th feature. Then the Fisher score of the i-th feature is computed below:

$$F(x)=\sum|\mu_0-\mu_1|/(S_0+S_1) \quad (4)$$

After computing the Fisher score of each feature, the top-m ranked features with large scores are selected. The features selected by the F-score are trivial because the score of each feature is computed independently [11] and sometimes the characteristics of features when they are combined are significant. So the relevance between the features is significant when it is tried to select informative genes. The proposed hybrid algorithm considers not only the information within the features (within classes) and also the relevance/information between any two features.

D. Algorithm for the Proposed Method

```
1.Input the Gene Expression Samples $S_{i., i=1...N}$
2.For Each Gene G in sample $S_i$ Compute
  2.1 Discretize expression profiles for all the genes which
  suits all the levels of data values present in the
  data set.
  2.2 Compute probability for each state occurs in the
  profile for the two genes that suits all the levels
  of discretization.
  2.3 Compute Initial Entropy using the Eq 1
  2.4 Conditional Entropy using the Eq. 2
  2.5 MI using the Eq. 3
3.Rank the Genes according to highest MI values
4.Compute F-Score for the set of selected genes (from
  step-3) as follows
  4.1 Compute μ0 and μ1 for class 0 and class 1 of
  sample Si
  4.2 Compute the absolute difference between the means
  |μi0-μi1|
  4.3 Compute SNR for all the genes as follows F= (|μi0-
  μi1|)/(σi0+σi1)
  4.4 Rank all the selected genes according to highest F-
  Score
5.The top ranked genes are highly informative and are
  input to the classification algorithm.
```

The performance of gene selection is improved by multilevel discretization and considering the information within the features and also between the features with hybrid method of gene selection.

E. Block Diagram

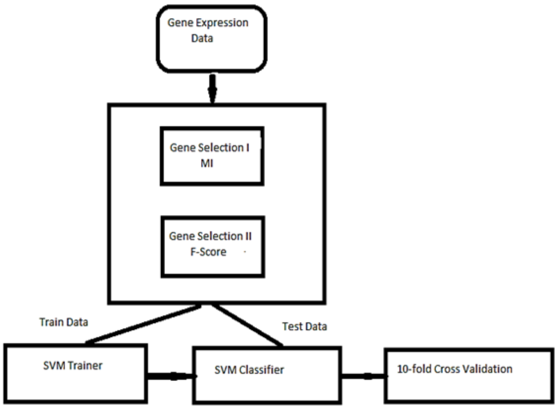

Fig. 1 Block Diagram of Proposed Method

F. Classification

SVMs have been widely used in the recent years in the field of computational biology due to their high accuracy and their flexibility in modelling diverse sources of data. They are mainly used in binary classification and regression. They are very suitable for classifying microarray gene expression data [12].

IV. Experimental Results

A. Data Sets

The Datasets are taken from Kent Ridge Biomedical Data Repository. Below are the descriptions of the Datasets used.

TABLE II
DATA SETS

Data Set Name	Number of Genes	Class	Total Samples
DLBCL Harvard	7129	DLBCL, FL	77(58/19)
AML-ALL	7129	AML, ALL	72(47/25)
Lung Harvard2	12533	ADCA, Mesothelioma	181(150/31)

TABLE III
PERFORMANCE OF CLASSIFICATION WITH SVM-LINEAR

Data Set Name	Sensitivity	Specificity	Classification Accuracy
AML-ALL	96.8	100	97.78
Lung Harvard2	99.5	100	99.5
DLBCL Harvard	94.3	95.4	94.7

TABLE IV
PERFORMANCE OF CLASSIFICATION WITH SVM-RBF

Data Set Name	Sensitivity	Specificity	Classification Accuracy
AML-ALL	100	95.2	96.7
Lung Harvard2	100	99.4	99.6
DLBCL Harvard	79.4	97.3	92.1

TABLE V
GENES SELECTED - AML/ALL DATA SET

Gene No	Gene ID	Gene Description	F-Score
4847	X95735_at	zyxin, ZYX	1.31
1882	M27891_at	cystatin C (amyloid angiopathy and cerebral hemorrhage), CST3	1.14
2354	M92287_at	cyclin D3, CCND3	1.12
2642	U05259_rna1_at	MB-1 gene	1.11
4328	X59417_at	PROTEASOME IOTA CHAIN	1.1
1685	M11722_at	Deoxynucleotidyltransferase, terminal, DNTT	1.07
4196	X17042_at	PRG1 Proteoglycan 1, secretory granule	1.06
1745	M16038_at	LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog	1

B. Results

In order to evaluate the efficiency of the method, performance measures like sensitivity, specificity are considered. The measures are computed using the following formulae.

Classification Accuracy (%): (TP + TN) / (TP + FP + FN + TN)

Sensitivity (%) = TP / TP + FN×100

Specificity (%) = TN / FP +TN ×100

The Proposed Techniques are implemented using R version 3.2.2.

TABLE VI
GENES SELECTED - LUNG HARVARD2 DATA SET

Gene No	F-Score
5301	1.39
7249	1.32
3764	1.27
9824	1.26
7046	1.2
3389	1.11
3508	1.06
5847	1

TABLE VII
GENES SELECTED - DLBCL DATA SET

Gene No	F-Score
1818	0.91
4372	0.89
2988	0.89
4183	0.87
4463	0.85
1373	0.85

V. CONCLUSIONS

In this paper, an efficient hybrid feature selection method is presented by embedding the Improved MI and the F-score statistics. The proposed hybrid method effectively reduces the dimension of the samples in capturing the features relevant to classes. The results of the 10-fold Cross Validation test using the standard datasets shows the potential of the proposed method with the advantage of reduced computational complexity. Hence, it can be used as an efficient approach for class prediction of microarray gene expression samples.

REFERENCES

[1] P. Ganesh Kumar and T. Aruldoss Albert Victorie, Multistage Mutual Information for Informative Gene Selection, Journal of Biological Systems, Vol. 19, No.4, pp 1–221,2011.

[2] P. E. H. R. O. Duda and D. G. Stork. Pattern Classification. Wiley-Interscience Publication, 2001.

[3] Saeys Y, Inza I, Larranaga P, A review of feature selection techniques in bioinformatics, Bioinformatics 23, pp 2507–2517, 2007.

[4] Sun-Yuan Kung, Man-Wai Mak, Feature Selection for Self-Supervised Classification with Applications to Microarray and Sequence Data, IEEE Journal of Selected Topics in Signal Processing, Vol. 2, No. 3, June 2008.

[5] Zhou X, Wang X, Dougherty ER, Nonlinear probit gene classification using mutual information and wavelet-based feature selection, J Biol Syst, Vol. 2, No. 3, pp.371–386, 2004.

[6] Fleuret F, Fast binary feature selection with conditional mutual information, J Mach Learn Res 5, pp 1531–1555, 2004.

[7] Xiaosheng Wang and Osamu gotoh, A Robust Gene selection Method for Microarray-based cancer Classification, Cancer Informatics, pp. 15–30,2010.

[8] Xiaosheng Wang and Osamu gotoh, Accurate Molecular Classification of Cancer using simple Rules, BMC Medical Genomics, pp 2:64,2009.

[9] Laiwan Chan, "Informative Gene Discovery for Cancer Classification from Microarray Expression Data", IEEE Workshop on Machine Learning for Signal Processing, vol.28, pp.393-398, Sept. 2005.

[10] 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, E.S. Lander, “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring” Science, 286, pp.531–537,1999.

[11] Quanquan Gu, Zhenhui Li and Jiawei Han, Generalized Fisher Score for Feature Selection, In Proc. of the 27th Conference on Uncertainty in Artificial Intelligence (UAI), Barcelona, Spain, 2011.

[12] Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Michel Schummer, David Haussler, Support Vector Machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics, Vol. 16, No. 10, pp.906–914, 2000.

[13] Dina A. Salem, Rania Ahmed A. A. Abul Seoud, and Hesham A. Ali,” A New Gene Selection Technique Based on Hybrid Methods for Cancer Classification Using Microarrays”, International Journal of Bioscience, Biochemistry and Bioinformatics, Vol. 1, No. 4, November 2011.

[14] Ghaffari, Noushin, and Hisham Al-Mubaid. "A New Gene Selection Technique Using Feature Selection Methodology." In Computers and Their Applications, pp. 217-222. 2006.

[15] Hala Alshamlan, Ghada Badr, and Yousef Alohali1, mRMR-ABC: A Hybrid Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling, Hindawi Publishing Corporation BioMed Research International Volume ,2015.

[16] Li-Yeh Chuang, Cheng-Huei Yang, Jung-Chike Li and Cheng-Hong Yang, A Hybrid BPSO-CGA Approach for Gene Selection and Classification of Microarray Data, Journal of Computational Biology Volume 19, No 1, pp. 68–82,2012.

[17] Han F, Sun W, Ling Q-H, A Novel Strategy for Gene Selection of Microarray Data Based on Gene-to-Class Sensitivity Information. PLoS ONE, 2014