INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Dimensionality reduction and classification in gene expression data using hybrid approach

K. Iswarya
aiswaryakumar96@gmail.com
Anna University BIT-Campus, Tiruchirappalli, Tamil Nadu

C. Kanimozhi
kanimozhi.durai@gmail.com
Anna University BIT-Campus, Tiruchirappalli, Tamil Nadu

ABSTRACT

The microarray technology has modernized the approach of biological research in such a way that scientists can now measure the expression levels of thousands of genes simultaneously in a single experiment. With this large quantity of gene expression data, experts have started to discover the possibilities of disease classification using gene expression data. Quite a large number of methods have been planned in recent years with hopeful results. But there are still a set of issues which need to be addressed and understood. In order to gain insight into the disease classification difficulty, it is necessary to get a closer look at the problem, the proposed solutions, and the associated issues altogether. In this paper, we present a dimensionality reduction methods and classification method such as Mutual information, Pearson's correlation, search algorithms (Recursive feature elimination, Genetic Algorithm, simulated annealing) with Support Vector Machine (SVM) classification algorithm and estimate them based on their evaluation time, classification accuracy and ability to reveal biologically meaningful gene expression. Our experimental results shows that classifier performance through graphs with improved accuracy in disease prediction.

Keywords— *Feature selection, Mutual information, Pearson's correlation, Genetic algorithm, Simulated annealing, SVM*

1. INTRODUCTION

The high dimensional data has brought unusual challenges to machine learning researchers, making the learning task more difficult and computationally challenging. The term high dimensionality is applied to a database that presents one of the following characteristics: (a) the number of samples is very high; (b) the number of features is very high, or (c) both the number of samples and features are very high.

DNA microarray technology is one of the fastest-growing new technologies that has empowered the study of gene expression in such a way that scientists can now measure the expression levels of large numbers of genes in a single experiment rather than performing experiments and gathering data for a single gene at a time. Gene expression microarray data comprises up to hundreds of thousands of features with relatively small sample size, so selecting a small subset of informative genes from microarray data continues to be a challenge. Gene Expression information provides the gene expression level contributing to a specific action.

With the "curse of dimensionality" of gene expression microarray data, it is common that a large number of genes are not informative for classification because they are either irrelevant or redundant. Currently, dimensionality reduction and feature selection have become essential tools for data mining tasks, especially for handling high-dimensional data such as gene expression microarray data. Feature selection can be defined as the process of detecting the relevant features and discarding the irrelevant and redundant ones with the goal of obtaining a subset of features. Feature selection techniques are classified into two major groups: filter and wrapper methods.

Filter methods are used to filter the features based on performance evaluation. Filtering is generally based on the characteristics of data and independent of any classification algorithm. The idea of the wrapper methods to select a feature subset using a learning algorithm as part of the evaluation function. Wrappers commonly select feature subsets on the basis of how well a learning machine algorithm performs.

In this paper, we propose two-stage filter methods and wrapper method for effective gene classification. In the first stage a new filter based Feature Selection model, Mutual Information (MI) is presented to estimate the dependency between features and output classes. Mutual information (MI) between the genes and the class label is used for finding the most informative genes. From, this most relevant features subsets are retained. In the second stage, Pearson's Correlation between the genes is used for removing redundant data from the relevant feature subsets. Finally, the wrapper-based feature selection method, Support Vector Machine (SVM) is presented. The SVM performs well with a simple kernel when analyzing microarray expression data for genes. The selected features are classified and it is validated by using fivefold Cross Validation approach. This approach is to partition the training data into five different sets.

If one set is taken as testing means, remaining sets are considered as training based on this classification can be done.

2. RELATED WORK

As specified earlier, the high dimensionality of gene expression data remains challenging. Among thousands of genes, some genes are not informative for classification. So, there is a need to select some genes that are highly related to particular classes for classification, known as informative genes. This process is known as gene selection, which is also referred to as feature selection in machine learning. Many works have been done related to feature selection by using numerous data mining techniques and algorithms. The aim of all work is to achieve better accuracy and to make the system more efficient and effective. Some of the most widely applied methods in the literature are briefly described as follows

[1] Raid Alzubi et al. presented a framework using conditional mutual information maximization (CMIM) and Support Vector Machine as Recursive Feature Elimination (SVM-RFE). This framework consists of three stages: Preprocessing, Feature selection, Classification. In the preprocessing stage, they remove redundant data. In the feature selection stage, CMI to calculate the amount of relevancy and redundancy. It works by selecting features that maximize their mutual information with the class to predict, conditionally to the response of any feature already selected (S). During the selection process, which calculates CMI only for the features that carry more information and are not redundant. In stage 3, SVM-RFE finds a subset of features that lead to the margin maximization of class separation. SVM-RFE provides a ranked feature list from which a group of top-ranked features can be selected in order to select the optimal features subset. They conducted an experiment with 5 different datasets obtained from NCBI and they conclude their model outperformed the compared methods (mRMR, CMIM, FCBF, ReliefF) in terms of accuracies.

[2] Osama Ahmad Alomari et al. developed a hybrid filter-wrapper gene selection method using Minimum Redundancy Maximum Relevancy (MRMR) as the filter approach and flower pollination algorithm (FPA) as the wrapper approach. First, they used the MRMR filter method to find the most important genes from all genes in the gene expression data. In the filter stage, the MRMR is employed to filter out noisy and redundant genes. Relevancy and Redundancy were two main components, MRMR attempts to select the genes that have the minimum redundancy for input genes and the maximum relevancy to the target class. In the wrapper approach, FPA algorithm is used to perform a subset generation, while SVM classifier is adopted for the evaluation of each candidate gene subset produced by FPA. FPA algorithm starts to generate the initial population or group of pollens. Each pollen composed of a series of 0's and 1's bits, where bit value 1 denotes that this gene is chosen and bit value 0 imply that this gene is ignored. They use three cancer datasets to test the performance and they conducted a comparison with the MRMR-GA, MRMRFPA and they conclude MRMRFPA achieved a classification accuracy with a less number of genes.

[3] Edmundo Bonilla Huerta et al. designed a hybrid framework for gene selection and classification of DNA microarray data. They use multiple fusion filter by combining five statistical approaches and each filtering method generates a different list of relevant genes. Then different relevant gene subsets are selected by using an embedded method that uses a Genetic Algorithm (GA), with a Tabu Search (TS) and Support Vector Machine (SVM). In the first stage, they combine five filters to achieve an initial gene subset that could be improved by a classifier or wrapper approach in a second stage. The main idea of this combination is to select a useful small subset of informative genes with high classification accuracy that can serve biologists to understand a certain disease or cancer tumor. GA performs a global search and Tabu Search performs a local search to improve the quality of the candidate gene subset by finding a better gene subset. Finally, SVM was presented to evaluate the fitness of a candidate gene subset in terms of classification and size of genes subset by using 10-Fold Cross-Validation as a validation method. They can conclude that more filters are used more possibilities that not leave out any relevant gene or potential biomarker at the first stage.

[4] Osama Ahmad Alomari et al. developed a hybrid gene selection algorithm for cancer classification. In this paper, they proposed a filter method (mRMR) and wrapper method (Bat algorithm, BA) for gene selection in the microarray dataset. They use mRMR (Minimum Redundancy Maximum Relevance) filter method to figure out the best genes and then reduced the set of genes generated from Mrmr method were evaluated by using SVM classifier. Three microarray datasets were used (colon, Breast, and Ovarian) to test the performance of the approach. By the experimental results, they conclude that MRMR-BA approach achieves high accuracy with less number of genes compared with MRMR-GA.

[5] Huijuan Lu et al. presented a hybrid feature selection algorithm for gene expression data classification. In this paper, they combine the mutual information maximization (MIM) and the adaptive genetic algorithm (AGA) for robust feature selection. Six gene expression data and four different classifiers (Backpropagation neural network (BP), Support vector machine (SVM), ELM and Regularized extreme learning machine (RELM)) were used in their experiments and they show that the developed MIMAGA-Selection method reduces the dimension of gene expression data and removes the redundancies for classification and they conclude that RELM was the most suitable classifier for the MIMAGA-Selection algorithm based on classification accuracy.

[6] Jaison Bennet et al. developed a hybrid approach for gene selection and classification using support vector machine. Redundancy in gene expression data leads to poor classification accuracy and also acts badly on multiclass classification. This paper presented an ensemble feature selection technique which is a combination of Recursive Feature Elimination (RFE) and Based Bayes error Filter (BBF) for gene selection and Support Vector Machine (SVM) algorithm for classification. In RFE method, nested subsets of features are selected in a sequential backward elimination manner, which starts with all the feature variables and removes one feature variable at a time. Based Bayes Error Filter method can effectively perform gene selection with reasonably low classification error rates and a small number of selected genes. After proper ranking by SVM-RFE which reduces computational complexity, the resultant is given to BBF which eliminates redundancy. Based on the experimental results on leukemia dataset it is found that the performance of SVM-RFE and BBF combined with SVM for classification was superior to the other related works in terms of gene selection and classification but it consumes more time.

[7] Atiyeh Mortazavi et al. Presented a robust feature selection from microarray data based on cooperative game theory and qualitative mutual information. In this paper, the authors developed a multiphase cooperative game theoretic feature selection approach for microarray data classification. In the first

phase, due to the high dimension of microarray data sets, the features are reduced using two filter-based feature selection methods (mutual information and Fisher ratio). In the second phase, Shapley index was used to evaluate the weight of each feature considering the complex and inherent relationships among features and Qualitative Mutual Information (QMI) measure was used for more robust feature selection. The results show that the proposed method was a stable method for reducing the dimensions of data and was able to reach relative improvement as compared to other feature selection methods.

3. PROPOSED SYSTEM

This part describes the proposed system. The overall System Design is shown in Fig.1. The proposed system consists of two main stages: 1) a hybrid feature selection stage (Filter and wrapper methods), and 2) a classification stage.

Datasets: The data set is collected from publically accessible website https://www.ncbi.nlm.nih.gov/sites/GDSbrowser/

Table 1: Details of Gene Expression Datasets

Dataset	Total samples	No of Genes	Class label	Class wise samples
GDS3257	107	22217	Tumor, cancer	59,48
GDS2771	187	22215	No cancer, Cancer	91,96

3.1 Hybrid feature selection

3.1.1 Filter method

(a) Mutual information

The relevance of a feature is the most significant selection criterion because using highly relevant feature improves the accuracy of the system and also the equal attention should be specified for the selected features need to be non-redundant. Two-stage filters are used to filter noisy and redundant genes in high-dimensional microarray data. In the first stage, a theoretical analysis of Mutual Information (MI) is introduced to evaluate the dependence between features and output classes. Mutual information refers to the dependent information of one random sample (x) on the other random sample (y). This feature selection approach that tends to select features with a high correlation with the class (output). The most relevant features are retained to perform further exploration on these genes and thus identify a subset of informative genes.

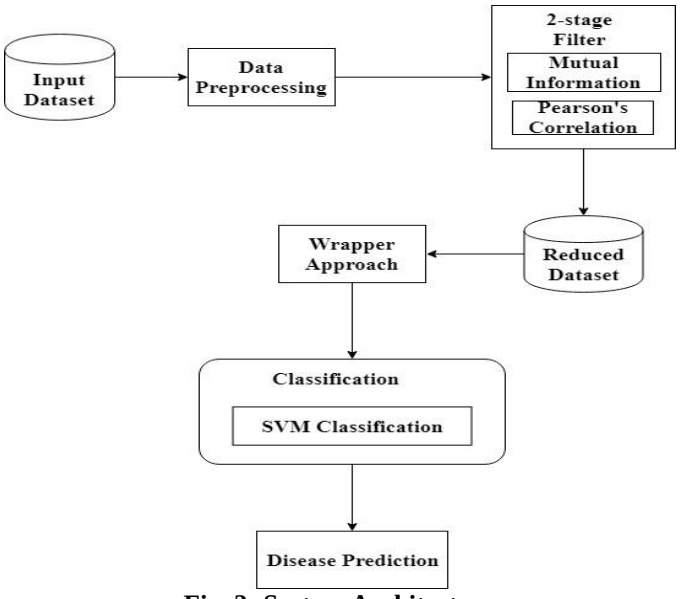

Fig. 2: System Architecture

Procedure

Input: Feature Set F= { i ,i=1, …..,n}

Output: Select feature subset

Step1: Initialization set S empty

Step2: Calculate the amount of information for every feature i with regard to class

Step3: Select the feature i which has the maximum mutual information.

Step4: Add the feature i to the set S

(b) Pearson's correlation

In the second stage, to remove redundancy from the relevant feature subset, Pearson's correlation is used. Pearson correlation is commonly used in statistics to measure the degree of the relationship between linear related variables. The relevant feature subset obtained from mutual information is further processed by using Pearson's correlation to remove redundancies in the subsets. This reduces the dimensionality of the data and may permit machine learning algorithms to work faster and more effectively.

3.1.2 Wrapper method

Wrapper methods search through the space of possible features (n feature sets) by using a search procedure and it runs a model on the subset to evaluate it. The top-ranked genes selected from filter methods are passed to the wrapper approach to perform further exploration on these genes and thus identify a subset of informative genes. In the wrapper approach, RFE (Recursive Feature Elimination) algorithm is used to perform a subset generation, while SVM classifier is adopted for the evaluation of each candidate gene subset produced by RFE. Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. RFE is a greedy optimization algorithm which aims to find the best performing feature subset. First, the algorithm fits the model to all predictors. Each predictor is ranked using its importance to the model. This method implements a backward selection of predictors based on predictor importance ranking. The predictors are ranked and the less important ones are sequentially eliminated prior to modeling. The main goal is to find a subset of predictors that can be used to produce an accurate model. Recursive feature elimination method used to find the importance of feature by ranking them. The features which have low rank are considered less important to the model. By this, we ranking the features and find the best feature subset in order to improve the accuracy of the model.

To evaluate the performance of feature selection methods by comparing with other wrapper approaches, Genetic Algorithm and Simulated Annealing are used. Genetic algorithm is a heuristic search technique used for finding optimized solutions to problems based on the theory of natural selection and evolutionary biology. There are five phases Initial population, Fitness Function, Selection, Cross over, Mutation.

GA algorithm starts to generate the initial population. Each population composed of a series of 0's and 1's bits, where bit value 1 denotes that this gene is chosen and bit value 0 imply that this gene is ignored. The genetic algorithm begins with a population that is produced randomly. All chromosomes are calculated using a fitness function to determine their fitness values. The higher fitness chromosomes are kept, and the fewer fitness ones are discarded in generating a new population through crossover and mutation. At the end of the GA, gene subsets are compared among them, and the highest fitness value corresponding to the gene subset is selected.

Simulated annealing is a global search algorithm that allows a suboptimal solution to be accepted in the hope that a better solution will show up eventually. It works by making little unsystematic changes to an initial solution. It requires a larger number of model fits and has a greater chance of overfitting the features to the training set.

The basic idea here consists of using a wrapper approach to discover good subsets of genes, the goodness of a subset being evaluated by the classification accuracy for a given classifier.

3.2 Classification

Support vector machines (SVM) is a powerful classification algorithm that has shown state-of-the-art performance in a variety of biological classification tasks. The SVM classifier is well suited to work with high-dimensional data, such as microarray gene expression data. The first generation of SVMs was only designed for binary classification.

SVM classifies data in large data sets by identifying a linear or non-linear separating surface in the input space of a data set. The separating surface depends only on the subset of the original data known as a set of support vectors. An SVM classifies data by placing one or more planes on data such that it achieves good classification results. A good result of separation is achieved by a plane that has the largest distance to the nearest data points of any class, called functional margin. If this functional margin is large, then the generalization error of the classifier will be small and vice versa.

4. PERFORMANCE EVALUATION

This section presents the results of the proposed approach using two cancer microarray datasets. The original gene expression data are continuous values. We need to preprocess the data for effective classification so, we discretize the data. First, a discretized method is applied over the input data, with the aim of solving problems of unbalanced values and preparing the attributes of the sample to be processed by the feature selection algorithm of the next step. After discretization, feature selection is carried out using filters and wrappers. Finally, a classier is applied. 10-fold cross-validation is used. In each fold, feature selection is applied to the training set and the selected features are tested on the testing set.

Table 2: Performance comparison of our Feature selection methods with SVM

Datasets	Two-stage Filter and RFE with SVM	Two-stage Filter and GA with SVM	Two-stage Filter and Simulated Annealing with SVM
GDS3257	94.29	91.43	80.14
GDS2771	77.42	51.61	67.74

In this study, we tested the proposed method (two-stage filter with RFE) by comparing it with the Genetic algorithm. Mutual Information was carried out to filter-out the relevant genes. To remove redundant genes from relevant subset Pearson's correlation was used. The top 59 genes selected from the two-stage filter form a search space to RFE. After that SVM is applied to improve the classification accuracy. RBF kernel was assigned for SVM classifier. Furthermore, the grid search algorithm was running to tuning the parameter of SVM classifier.

Figure 3 shows the performance comparison of our Feature selection methods with SVM and other methods. It is very obvious that two-stage filter and RFE with SVM achieves higher classification accuracy than the other methods.

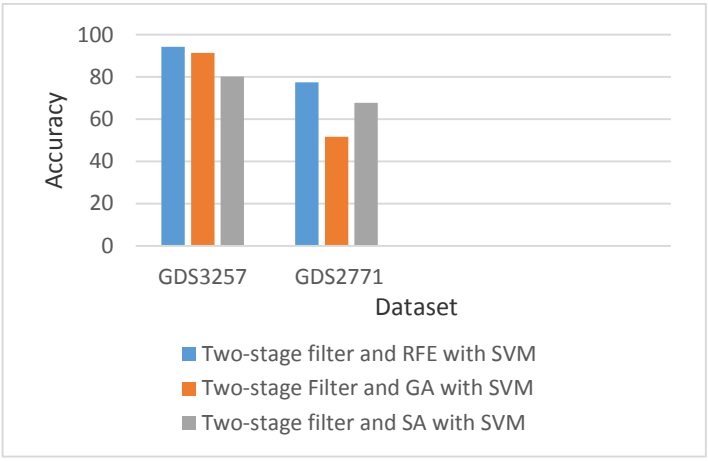

Fig. 3: Comparison Table

5. CONCLUSION

In this paper, a new approach proposed to solve the gene selection problem which combined Mutual Information, Pearson's Correlation and Search algorithms with SVM classifier. This approach is a hybrid filter-wrapper approach. The two-stage filter is used to filter irrelevant features (genes) and the RFE is adapted to perform the subset generation process in the wrapper approach. Initially, mutual information used to select the relevant subset after Pearson's correlation used to remove redundant the data from the relevant subset. Then top selected genes form as a gene search space to the wrapper approach. Two cancer datasets were used to test the performance of the proposed approach. Furthermore, a comparison with GA, SA shows that RFE achieved higher classification accuracy with less number of genes. The results exhibit that the RFE method is a promising approach for solving the gene selection problem.

6. REFERENCES

[1] M. Waddell, D. Page, and J. Shaughnessy Jr, "Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma," in Proceedings of the 5th international workshop on Bioinformatics. ACM, 2005, pp. 21–28.

[2] H. He, W. S. Oetting, M. J. Brott, and S. Basu, "Power of multifactor dimensionality reduction and penalized logistic regression for detecting gene-gene interaction in a case-control study," BMC medical genetics, vol. 10, no. 1, p. 127, 2009

[3] H. Schwender and K. Ickstadt, "Identification of snp interactions using logic regression," Biostatistics, vol. 9, no. 1, pp. 187–198, 2008.

[4] D. Evans, "An snp microarray analysis pipeline using machine learning techniques," Biol. Med., 2010.

[5] N. Batnyam, A. Gantulga, and S. Oh, "An efficient classification for single nucleotide polymorphism (snp) dataset," in Computer and Information Science. Springer, 2013, pp. 171–185.

[6] K. Anekboon, C. Lursinsap, S.Phimoltares, S.Fucharoen, and S.Tongsima, "Extracting predictive snps in crohnâA˘ Zs disease using ´ a vacillating genetic algorithm and a neural classifier in case-control association studies," Doctoral dissertation, Ohio University, 2014.

[7] Y. Saeys, I. Inza, and P. LarraÃsaga, "A review of feature selection ´ techniques in bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507– 2517, 2007.

[8] V. BolÃ¸sn-Canedo, N. SÃ ¸anchez-MaroÃso, A. Alonso-Betanzos, ´ J. BenÃ tez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," Information Sciences, vol. 282, pp. 111 – 135, 2014.

[9] T. Cover and J. Thomas, "elements of information."," Theory Wiley, 1991.

[10] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A survey on filter techniques for feature selection in gene expression microarray analysis," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 4, pp. 1106–1119, July 2012.

