**Feature Selection for Classification of Gene Expression Data**

Zuraini Ali Shah  
Laboratory of Computational Intelligence and Biology  
Universiti Teknologi Malaysia, 81310  
UTM Skudai, MALAYSIA  
+607-5599230  
aszuraini@utm.my

Puteh Saad  
Department of Software Engineering  
Faculty of Computer Science and Technology  
Universiti Teknologi Malaysia, 81310  
UTM Skudai, MALAYSIA  
+607-5599230  
puteh@utm.my

Razib M. Othman  
Laboratory of Computational Intelligence and Biology  
Universiti Teknologi Malaysia, 81310  
UTM Skudai, MALAYSIA  
+607-55992303  
razib@utm.my

**ABSTRACT**

This paper presents a gene selection method for classification of gene expression data. First, a feature selection technique based on t-test statistic is used applied in order to select the n top-ranked significant genes. Then, a combination of Genetic Algorithm (GA) and Support Vector Machines (SVM) is used to further select significant genes to the resulting set of genes obtained from t-test selection. The selection of t-test improves the selection of t-test. SVM classifier is utilized implemented to evaluate performance of the feature selection approaches. The final gene. Results of comparative studies are provided, demonstrating that effective feature selection approaches are essential to the development of classifiers intended for to be used in for high-dimensionality data problems. The result of SVM classifier with all genes compared with t-test/SVM and t-test/GA/SVM procedure are reported. This research also shows that feature selection helps increase computational efficiency while improving classification accuracy.

**Keywords**

Gene expression data, Feature selection, Classification, t-test, GA, SVM.

1. **INTRODUCTION**

In recent years, microarray technology allows the biological and medical research to analyze thousands or even more of gene simultaneously. Microarray have has opened the possibility of creating data sets of molecular information to represents many systems of biological or clinical interest [1,2,19]. [1] Recently, the application of whole-genome technologies such as gene expression profiling has opened up new avenues for biomedical field for identifying complex disease genes, finding biomarkers for disease and support clinical disease diagnosis. In traditional clinical practice, doctors use morphological appearance of tumours or the parameters derived from clinical observations for prediction in the individual patients. Although clinical predictors are useful, they are not accurate enough to predict the stage or type of certain disease in an individual patient [210]. Therefore, using gene expression profile and computational methods is an important step towards integration of complex genomic data customised for individual patient. In order to produce a reliable prediction result, it requires an appropriate approach that can produce high classification accuracy subjected to proper and systematic validation method.

The main procedures of gene expression data classification task includes: feature selection and classification. When dealing gene selection problems, the researchers often show interest in one of the following objectives:

1. To identify relevant genes for subsequent research; this involves obtaining a (probably large) set of genes that are related to the outcome of interest, and this set should include genes even if they perform similar functions and are highly correlated.

2. To identify small sets of genes that could be used for diagnostic purpose in clinical practice; this involves obtaining the smallest possible set of genes that can still achieve good predictive performance (the redundant genes should not be selected).

This paper will focus on the second objective. Recently, a new framework of feature (gene) selection has been used. Yu and Liu [211] used a fast correlation-based filter algorithm (FCBF) which used correlation measure to obtain relevant genes and to remove redundancy. Recursive feature elimination (RFE) is a proposed feature selection algorithm described by Guyon et al. [89]. Ding and Peng [75] have used mutual information for gene selection that has maximum relevance with minimal redundancy by solving a simple two-objective optimization. Xing et al. [209] proposed a hybrid of filter and wrapper approaches to feature selection. Shutao et al. [18] proposed a gene selection method based on genetic algorithm (GA) and support vector machines (SVM) for cancer classification. The Wilcoxon rank sum test is used to filter noisy and redundant genes. Shen et al. [16] proposed a gene selection method based on a modified PSO and used support vector machines (SVM) for cancer classification. The t-test is used to filter noisy and reduce the number of genes in high-dimensional microarray data. Shen and Shutao used LOOCV for feature selection in order to avoid using a fixed set of features selected with the whole training data set, because this induces a bias in the results. Ambroise and McLachlan [326] recommend to use using 10-fold rather than leave-one-out cross-validation, because the last one can be highly variable. Therefore, based on Shutao proposed method, this research improved the implementation of GA/SVM by modifying the cross-validation implementation and show the interesting aspect for biologist especially for colon cancer research.

This paper is organized as follows. Section 2 explains the motivation of the research. Section 3 describes the
methodology of this research. Section 4 presents the experimental results from using the proposed method. Section 5 summarizes the results and last section draws a general conclusion.

2. Motivation

The vast amount of gene expression data pose great challenges for successful application such as medical diagnosis application. Therefore, biological experiment requires an approach that can extract informative genes that is related to cancer disease. It is impractical to extract manually without computational effort. This is a key factor to select the meaningful genes that can contribute to cancer research and also for other disease such as Alzheimer [15]. By finding genes contributed to a particular disease or a type of disease can aid a doctor in further analysis. From computational perspective, feature selection is a crucial task to select groups of genes that will help to increase the biological knowledge between genes and disease. In addition, this study is also important for cancer diagnostic purposes.

One important element in developing a class predictor is the identification of most significant genes. The accuracy and cost of an automated classifier is dependent on the appropriate selection method chosen. By considering the different cross validation (CV) and resampling small samples domains, this paper proposed a computational method that aim to identify genes that can lead to more accurate classification. Thus, it can be foreseen that a reliable, robust, fast and efficient feature selection method is a great contribution to biological, medical and computational field.

3. Methodology

This section will describe the proposed computational method including data sources, instrumentation and analysis of the results.

3.1 The Proposed Computational Method

The proposed computational method contains five phases namely the pre-processing phase, pre-selection phase, feature selection phase, classification phase and validation phase. The proposed computational method is shown in Figure 1. In pre-processing, the colon dataset is pre-processed by normalized to a scaled into (0-1) scale for further analysis. Then the dataset are firstly divided intro training and testing samples. Out Among of a total 62 colon samples, 42 samples selected randomly selected are treated as samples were used as training dataset. And the remaining 20 samples are utilized as the testing dataset.

Figure 1 : The proposed computational method

are will be selected using a by statistical pre-selection technique. The pre-selection phase aim to evaluate the distribution of the dataset, rank the features and sort them according to the top-rank statistically informative features. In this process, the parametric pre-selection method, t-test filtering algorithm is used recommended. The t-test is work on each feature and compares p-value for each feature as a measure of how effective it is at separating groups. The empirical cumulative distribution function (CDF) of the p-values will be used to visualize and tends to decide how many features will be selected for further processes studies. Base on Figure 1, the output from pre-selection phase represent the number of pre-selected features. In feature selection phase, this output is evaluated using wrapper feature selection method, known as GA. GA is will be used to evaluate the subset of features (gene) by deploying classification based fitness function. The fitness function is the 10-CV that evaluates the error rate of the classification. Thus, each subset of features is evaluated using a classification algorithm in order to generate new and better subsets. In classification phase, n features and class label of testing dataset will be evaluated by a classification algorithm, called SVM. In classification phase, a 10-CV is used to evaluates the classification accuracy.

3.1.1 Pre-selection using t-test

t-test is a classical and very simple statistical pre-selection approach. The implementation of t-test can be summarized as follows:

- Each sample is labelled with {1,-1}. The value of 1 represents the sample that is normal and -1 for is abnormal sample.
- For each feature (gene) \( f_j \) the mean \( \mu_j \) and standard deviation \( \delta_j \) are calculated using only the samples labelled 1 (resp. -1).
- Then a score \( t(f_j) \) can be obtained by the following equation defined as:
During making a selection process, those genes features with the highest scores are selected, considered as the most discriminatory features. In this paper the top-400 features given by t-test are selected to be classified. chosen for further analysis. Generally, the aim of pre-selection is to reduce the dimensionality of the data.

3.1.2 FGA as eature selection using GA

GA will select subset of genes based on the value of fitness function. and fitness function will return error rate from classifier task. This study utilized identify it as classification-based fitness function. As known, GA is a search algorithm inspired by the principle of natural selection. The basic idea is to evolve a population of individuals, where each individual is candidate solution to a given problem. Each individual is evaluated by a fitness function, which measures the quality of its corresponding solution. At each generation (iteration) the fittest (the best) individuals of the current population survive and produce offspring resembling them, so that the population gradually contains mostly a set of fit fitter and fitter individuals (better and better candidate solutions to the underlying problem). There are three important components of GA namely; which are chromosome representation, classification-based fitness function and reproduction.

(a) Chromosome representation

Any The candidate subset of selected features is represented as an individual which consists of $N$ genes (equating to the total number of all features in the population). In this work, a number of 400 genes of colon cancer dataset is has been used as chromosomes length. In the chromosome representation, it uses the binary coding system and when the value is ‘1’ it means the gene is selected. On the other hand, when the value is ‘0’ it means the gene is not selected. Figure 2 depicts a is chromosomes representation for the colon dataset.

(b) Classification-based fitness function

In this study, SVM machine learning act as classifier or predictor in GA fitness function. The classifier is used to assess the quality of a gene subset. Using the output of this classifier, the class label will be used to evaluate the error rate or misclassification error. For a chromosome $x$ that represents a gene subset, a 10-CV method is applied in order to calculate the average error rate of a GA trained with this gene subset. The 10-CV procedure proceeds as follows: divide the dataset is divided into 10 parts equally, and then iteratively use one part is used for testing and the rests other parts for training. Finally, compute the average error is computed. This process is will repeated 10 times. and Eeach of time a the different fold of sample will be serve as a test and a train data. Thus, for each chromosome $x$, $f(x) = \sum_{i=1}^{n} e(x) / n$ (1)

where $f(x)$ indicates fitness function for each chromosome $x$, $\Sigma e(x)$ indicates sum of the classification error and $n$ indicates number of samples.

Figure 2 Chromosomes representation for colon dataset.

3.1.2 SVM as classification method

In order to For classifying microarray data, one can use the classical linear discriminant analysis, naive bayes, decision tree, artificial neural networks, k-nearest neighbor (KNN), as well as some more sophisticated machine learning methodologies including bagging, boosting and kernel methods. Among them, SVM is one of the most powerful supervised learning algorithms in gene expression analysis [14]. SVM has a been found generalization ability and useful in handling classification tasks for in case of the high dimensionality and sparsity of data. A points. Thus, this study proposes SVM as classification method at two phases, feature selection and classification. In feature selection, GA is used as the search strategy, while SVM is used as classifier in GA fitness function. Next, after the feature selection task is complete, once again SVM is used as classifier to evaluate the selected features for test dataset.

3.2 Data Sources

A benchmark dataset by Alon [2] is chosen to measure the performance of the proposed method. The colon dataset consists of 62 samples of colon epithelial cells. There are 40 samples of epithelial cells and 22 normal cells and. The dataset can be downloaded from http://www.microarray.princeton.edu/oncology/. Originally, the raw data consists of three files, the matrix $I_{2000}$, the tissues data and the gene name. However, The dataset is available quiet popular and user can found the formatted in a matrix form weather in .arff format (for WEKA) or .mat format (for Matlab). Since the raw expression values have been pre-processed, this study kept all 2,000 genes and then scaled into (0,1.0, 1.0). for further analysis.

3.3 Results Validation Method

Result analysis

The proposed computational method is developed and tested on a standalone Personal Computer (PC). The experiment was conducted using the HP PC and the specification of the PC includes, Intel Pentium IV CPU 3.00GHz with 1.00 GB of RAM and 80GB of HDD. The experiment is conducted in a Matlab environment using a LibSVM package that can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

The development of the computational complete method will be developed using Matlab program. The SVM implementation is based on LibSVM using C language and can be integrated by customicustomizing the LibSVM package in Matlab program.
The source code can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

From computational perspective, a few measurements are considered to be described here. First, is about the limited number of samples. In order to use a particular classifier designed, a testing that represents an independent set of samples of known classification as well must be provided. Normally, using the conventional but reliable estimation method, the training and testing dataset will be divided dividing the dataset. The training and testing partition, is randomly picked from complete samples of each dataset. The testing dataset is separated from feature selection process in order to avoid superior result in prediction task. Meanwhile, cross-validation is a technique that helps one to estimate the error precisely in terms of statistics. In order to verify and validate the results, Two common metrics are utilized; types are LOOCV and $k$-fold cross validation ($k$-CV). LOOCV used only observation for testing, and the other for training. LOOCV use in data set with sparse data. Second type is $k$-CV (often use $k = 10$), divide the dataset into $k$ parts equally, and then iteratively use one part for testing and the other parts for training. Finally, compute the average error. This study implement $10$-CV to evaluate the classification performance at to place: classification based fitness function (training) and classification (testing).

4. Result and discussion

Initially At the beginning, 2000 genes of colon dataset are classified using SVM classifier alone. with all 2000 genes was carried out for colon cancer dataset. Next, Second, using a t-test filtering algorithm, 400 top-ranked informative genes is selected and then classified using SVM. Lastly, GA and SVM is used t-test/GA/SVM is implemented to evaluate the 400 top-ranked genes. As depicted shown in Table 1, 80% classification performance is obtained without any feature selection method for colon dataset. Using t-test with /SVM, 91% accuracy is obtained to classify with 400 genes. With a combination of t-test, GA and SVM and achieved t-test/GA/SVM can give the best result of 96% accuracy based using only on three genes. The reliability of a classification model is an essential issue in microarray data analysis [8,16 ]. In order To evaluate accurately the predictive ability and reliability of models, the total tissue samples were randomly partitioned into training and testing sets 10 times and then averaged the classification accuracy for each selected of genes. This procedure is run three times (R1-R3) and the maximum classification accuracy is reported here. From Table 1, as the number of genes become small, the higher classification accuracy obtained. This is due to may be because most of the more irrelevant genes are were removed that improves and able to increase the classification performance. Hence, the results of these experiments It indicates the necessity to exclude irrelevant genes from in microarray data during data analysis processess. analysis.

<table>
<thead>
<tr>
<th>Method</th>
<th>Genes</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>Max acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>2000</td>
<td>80%</td>
<td>79%</td>
<td>78%</td>
<td>80%</td>
</tr>
<tr>
<td>t-test/SVM</td>
<td>400</td>
<td>83%</td>
<td>81%</td>
<td>91%</td>
<td>91%</td>
</tr>
<tr>
<td>t-test/GA/SVM</td>
<td>3</td>
<td>90%</td>
<td>95%</td>
<td>96%</td>
<td>96%</td>
</tr>
</tbody>
</table>

Table 1. Classification accuracy of colon dataset

Based on Table 1, it is observed that Analyzing the results of all features for SVM classification, the performance is low compare with the combination of t-test pre-selection with SVM. Meanwhile, the two stage of selection the features that combine t-test and GA presented the best result. It can be noticed that the combination of t-test and GA is an advisable step for gene expression data since the number of irrelevant features has been reduced and can increase the accuracy of SVM classification.

The comparison between the recently published methods and our proposed method are shown in Table 2. Based on the table 1, the abbreviations are as follows: parallel genetic algorithm (PGA), nonparametric scoring (NPS) method, local linear embedding (LLE), support vector machine (SVM) normalized mutual information (NMI), K nearest neighbors (KNN), optimized between-group classification (OBC), mutual information quotient (MIQ), Naïve Bayes (NB) classifier, joint classifier and feature optimization (JCFO), Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). The numbers shown in the table is the classification accuracy and selected genes is are the number of genes selected by the methods. As shown in Table 2, this paper have has the least informative numbers of genes that produce 96% classification accuracy.

<table>
<thead>
<tr>
<th>Method</th>
<th>Genes</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>Max acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KNN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OBC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MIQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JCFO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Classification accuracies (%) obtained by the various methods as reported in the literature
5. Conclusion

This paper proposed three different approaches to classify colon dataset. The first approach is using SVM alone for classification of the colon data. Next, data is pre-processed using t-test and used SVM as classifier using GA and SVM to perform the classification. The findings indicate that a combination of t-test, GA and SVM achieves the best performance when compared to a combination of t-test with GA and SVM or SVM alone.

Teni, kita report yg bagus dulu baru yg tak bagus.. ambil cara Allah dlm surah al-waqiaah; cerita syurga dulu.. then baru neraka.. begitu juga.. if ada berita tidak enak nak sampaikan pd suami.. cerita banyak kabar baik 1 hingga 9.. yg kesepuluh baru cerita yg tak baik.—drp talk ust zahazan mohamad.

analysises on colon dataset in order to classify the data. Comparing the results obtained by using all the features, SVM classification presents the worst result followed by doing a pre-selection. The results show that the best prediction is obtained using t-test/GA/SVM. This paper investigates that the features selection task is essential to improve the classifier performance by the previous literature that used various methods and compare with the proposed method. Therefore, based on the experiments conducted, we can be concluded that it is important to note that a combination of t-test, GA and SVM t-test/GA/SVM can achieve the highest classification performance when compared to SVM alone of a combination of t-test and GA. select the optimal features that really contribute to high classification accuracy.

6. Future Works

As for For future works, similar the pre-processing, feature selection and classification method will be conducted for other dataset that are reported by [10]. This task is essential important in order to measure the stability of the proposed methods. on other dataset.

7. REFERENCES


<table>
<thead>
<tr>
<th>Method (t-test+GA+SVM)</th>
<th>Selected Genes</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGAL+Gloub’ classifier</td>
<td>30</td>
<td>92</td>
</tr>
<tr>
<td>NPS+LogitBoost</td>
<td>2000</td>
<td>87.1</td>
</tr>
<tr>
<td>JCF0 (linear kernel)</td>
<td>25</td>
<td>96.8</td>
</tr>
<tr>
<td>LLE+SVM</td>
<td>N/A</td>
<td>91.0</td>
</tr>
<tr>
<td>NMI+KNN</td>
<td>N/A</td>
<td>91.9</td>
</tr>
<tr>
<td>Jacknife+OBC</td>
<td>20</td>
<td>94.0</td>
</tr>
<tr>
<td>MIQ+NB</td>
<td>10</td>
<td>91.9</td>
</tr>
<tr>
<td>Wilcoxon+GA+SVM</td>
<td>15</td>
<td>93.6</td>
</tr>
<tr>
<td>t-test+PSO+SVM</td>
<td>3</td>
<td>94</td>
</tr>
<tr>
<td>Proposed method (t-test+GA+SVM)</td>
<td>3</td>
<td>96</td>
</tr>
</tbody>
</table>


