Third International Conference on Bioinformatics (InCoB'04). September 5-8, 2004. Auckland, New Zealand.

GENETIC ALGORITHMS WRAPPER APPROACH TO SELECT INFORMATIVE GENES FOR GENE EXPRESSION MICROARRAY CLASSIFICATION USING SUPPORT VECTOR MACHINES [*]

MOHD SABERI MOHAMAD, SAFAAI DERIS, MUHAMMAD RAZIB OTHMAN

Artificial Intelligent and Bioinformatic Laboratory, Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia.

Constantly improving gene expression technology offer the ability to measure the expression levels of thousand of genes in parallel. Gene expression data is expected to significantly aid in the development of efficient cancer diagnosis and classification platforms. Key issues that need to be addressed under such circumstances are the efficient selection of minimum number of genes that contribute to a disease from the thousands of genes measured on microarrays that are inherently noisy. This work deals with finding the minimum number of informative genes from gene expression microarray data which maximum the classification accuracy. In this work, we apply genetic algorithm wrapper to search out and identify the minimum number of potential informative genes combinations for classification and then use the classification accuracy from the support vector machine classifier to determine the fitness in genetic algorithm for each of the combinations. Experimental results using benchmark dataset produced the proposed approach achieves better classification accuracies by using minimum informative genes than other published methods on the same datasets. The genes from the outcomes are explored for biological plausibility.

1. Introduction

Due to recent advances in biotechnology, gene expression can now be quantitatively monitored on a global scale. Gene expression data is created by process known as microarraying that yields a set of floating point and absolute values [3]. These values represent the activation level of every gene within an organism at a particular point in time and a typical dataset can often consist of thousands of genes. If these microarrays are taken from several individuals with disease and also from those who are normal, a database of gene expression records that fall into separate classes can be created.

Recent studies on molecular level classification of tissue have produced remarkable results and indicated that gene expression microarray could significantly aid in the development of efficient cancer diagnosis and classification platform. However classification based on the microarray data confronts with more challenges. One of the major challenges is the overwhelming number of genes relative to the number of training samples in the datasets. Many of the genes are not relevant to the distinction between different tissue types and introduce noise in the classification process, and thus potentially drown out the contribution of the relevant ones [1].

[*] This work is supported by National Science Fellowship research program sponsored by Malaysian Ministry of Science, Technology and Environments (MOSTE).

In the gene expression microarray domain research, the gene refers to the feature. The research on feature selection in field of data mining is usually focused on small scale (5—60 features) or middle scale (60—hundred features). Even though there are some works on feature selection of too big scale (over 1000 features), they only introduce the filtering methods before classification and do not proposed clear way for it [6]. The application of genetic algorithm (GA) for feature selection has grown in recent years as the data has become more readily available [2], [8], [10]. But the previous works only was supporting the data ranging from small to medium features. Liu *et al.*,[7] combined the parallel genetic algorithm with classification method proposed by Golub *et al.*,[14] and Slonim *et al.*,[4] for gene expression classification. However the experiment requires much run time of the hybrid component. Then, the best subset of the runs will be selected as the final optimal subset. Nevertheless, the testing accuracy is still less than other research in the same domain [7].

From the literatures, all previous work that applied GA for feature subset selection used the same model of chromosome representation in GA. Unfortunately the chromosome representation cannot select a minimum number of genes. The number of genes selection depends solely on initial population in GA that produced by random. If the model used for the high dimension data, it requires many run time that will affect this manner and not efficient for selection of optimal subset that has minimum number of genes. It happens because the model is unsuitable to search in high features space and more dependent on initial population.

Since the data used in this work have thousand of features, the conventional approach is hard to be applied. So this work will modify the model of chromosome representation in GA to identify the minimum number of the genes combination for improving the classification accuracy. The combinations of genes are used for classification and then classification accuracy obtained from SVM classifier used to determine the fitness function.

2 Methodology

The overall classification strategy consists of two main components. The two main components are GA for features subset selection and SVM as classifier. The gene expression data usually have many thousand of features. The thousand of features can possible to cause the over fitting which learning a decision surface that performs well on the training data but bad on testing data. So in this work, we try to scale them. The data was pre-processed scaling between –1 to 1 as formula below to generate a new small features or values of dataset

$$x' = 2\frac{(x - m_i)}{M_i - m_i}. \tag{1}$$

where x and x' are the original value and new value after scaled respectively. Whereas M_i and m_i are maximum and minimum values of the i-th attribute respectively. After this process, the computational time may be reduced because scaling factor scales the large values to new small values.

2.1. Genetic Algorithms

A genetic is a global optimization procedure that uses an analogy of the genetic evolution of biological organisms [16]. It is a heuristic search procedure will modify function values of individuals coded as binary or real string by using GA operators in a stochastic manner. The string referred as a chromosome is divided into individual section called genes.

The individuals represent candidate solutions to the optimization problem being solved. In the feature subset selection problem each individual would represent a subset of features [7]. It is assumed that the quality of each candidate solution can be evaluated using a fitness function.

2.3. Support Vector Machine (SVM) Classifier

As one of machine learning algorithms, SVM is a good method suggested by Vapnik,[15] to get a high performance from real world problem, which have too big scale data.. SVM builds up a hyperplane as the decision surface in such a way to maximize the margin of separation between positive and negative examples. SVM achieves this by the structural risk minimization principal that the error rate of a learning machine on the test data is bounded by the sum of the training error rate and a term that depends on the Vapnik-Chervonenkis (VC) dimension.

Given a labeled set of M training samples $(x_1, y_1) \ldots (x_n, y_n)$, where $x_i \in R^N$ and y_i is the associated label, $y_i \in \{+1, -1\}$, the discriminate hyperplane is defined by

$$f(X) = \sum_{i=1}^{M} Y_i \alpha_i k(X, X_i) + b. \tag{2}$$

where $k(.)$ is a kernel function and the sign of $f(x)$ determines the membership of X. Constructing an optimal hyperplane is equivalent to finding the entire nonzero support vector α_i and a bias b. In this work we use the Radial Basis Function kernel defined as

$$K(x_i, x_j) = e^{-(x_i \square x_j)^2 / \sigma^2}. \tag{3}$$

where σ is gamma. The rational for using this kernel, is that early results from other work are readily found with excellent generalization performance in non-linear separable and low computational cost [9].

3. Proposed Method

We have used SVM to classify the data and consider the characteristics caused by features combination. In order to select a set of features used in the classification, GA has been adopted.

3.1. *GA Wrapper Manner to Feature Subset Selection for SVM Classifier*

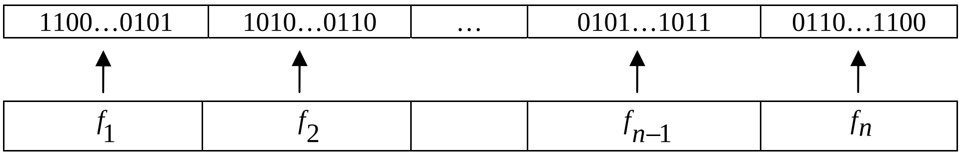
The representation of chromosome used in the early beginning of the works [2], [8], [10] is a structure which has number of feature bits and determines the usage of them by their values as shown in Figure 1 below. Bit value of 1 mean that the corresponding feature is selected. A value of 0 indicates that the corresponding feature is not selected. The total number of bit in the chromosome represents the totality of the features. This kind of structure only valid for the number of features is small or medium.

01010101000010000101…0111001011001001

Figure 1. The representation of chromosome in GA used in previous works.

Our project is the extension of this author's work by applying the general technique to a substantially different problem domain. The primary difference between gene expression data and all of the datasets used in the publications is that our data roughly have 100 times more available features with much fewer sample cases. We also modified the chromosome representation for features subset selection to support the very big scale data and change the fitness function, which is motivated to get the best accuracy in test data. We replaced the neural network with the SVM to classify the subset of features.

Each individual of the current GA population represents a competing subset of features that must be evaluated to provide fitness feedback to the evolutionary process. In this work, we have modified the representation of chromosome suited to gene expression microarray data that has huge-scale features as shown in Figure 2 below.



f_i = index of feature

n = number of total features

s = number of selected features

Figure 2. The representation of chromosome for huge-scale data.

It includes the real index f_i which indicates a selected feature of the i-th features among total features. The original binary value of index f_i will be converted to real value and actually the real values refer to the index of selected feature. This structure is not much affected by the total number of features and is able to represent chromosome in relatively small size. Its length can vary according to the size of the number of total features n and the number of selected features s. The length of chromosome is same size for each chromosome.

The fitness function of an individual is determined by evaluating the SVM classifier. From previous works [4], [7], [14] it is obvious that the best accuracy for cross-validation from training set is not necessary as the result will be best in testing set. It caused by over fitting the data during the training phase when learning a decision surface in SVM that performs well on the training data but badly on testing data. So in this paper, we used 1-criteria fitness function containing only accuracy for testing data as shown below

$$fitness(x) = accuracy(x). \tag{4}$$

where $accuracy(x)$ is the test accuracy for testing data of the classifier built with the feature subset selection of training data which is represented by x. The classification accuracy is $accuracy(x) = (C_T / A) x100$ where C_T and A are the numbers of true classified samples in the testing data and number of total samples in testing data respectively.

4. Experiments

In the previous section we have discussed on the proposed approach for feature subset selection. In this section we examine their performance on experimental datasets.

4.1. *Leukemia Cancer Dataset*

Leukemia dataset consists of 72 samples, 25 samples of *acute myeloid leukemia* (AML) and 47 samples of *acute lymphoblastic leukemia* (ALL). 38 out of 72 samples were used as training data and the remaining were used as testing data. It contains a training set composed of 27 samples of ALL and 11 samples of AML, and an independent testing set composed of 20 ALL and 14 AML samples. Gene expression levels in these 72 samples were measured by using high-density oligonucleotide microarrays [1], [11], [14]. Each sample contains 7129 gene expression levels.

4.2. *Experiment Environment*

Our experiments were run using Steady-State GA and roulette wheel selection strategy [16]. In these experiments, we assessed uniform crossover and also applied Gaussian mutation operated based upon the probability on each of the offspring strings produced from crossover [16]. The parameters setting in Table 1 and Table 2 were chosen based on results of several preliminary run.

Table 1. Parameters of the GA for Leukemia Dataset

Genetic Algorithm Parameter	Value
Size of population	100
Number of generation	2000
Replacement rate (Roulette Wheel)	0.8
Crossover rate	0.7
Mutation rate	0.01

Table 2. Parameters of the SVM

Support Vector Machine Parameter	Value
Regularization cost, C	100
Gamma, g	1/k,[a]

First experiment in this work selects the whole genes in the datasets for classification process. Second experiment uses the proposed approach on number of selected genes that ranged 1 to 10. The classification accuracy is compared with previous works. Genes in the best subset, which have been used ranging from 1 to 10 genes, will be evaluated as the identical biological significant with previous works to examine the informative genes.

4.3. *Analysis of Results*

Table 3 below shows the results using SVM classifier for the whole genes in the dataset. The accuracy classifier is first tested using leave one out cross validation (LOOCV) procedure. A classifier will be built up using the whole training set. Then the accuracy on the testing set of samples is taken.

Table 3. Result of experiments using SVM classifier

Dataset	Number of Genes	LOOCV Accuracy	Test Accuracy
Leukemia	All (7129)	94.7369	85.2941 (29/34)

Table 4. Results of experiments using proposed method

Dataset	Number of Genes Selected	Test Accuracy (%)
Leukemia	1	97.0588 (33/34)
	2	97.0588 (33/34)
	3	100 (34/34)
	4	100 (34/34)
	5	100 (34/34)
	6	100 (34/34)
	7	100 (34/34)
	8	100 (34/34)
	9	100 (34/34)
	10	100 (34/34)

As shown in Table 4, the accuracy rate has increased when the numbers of selected genes become increased. At least three selected genes are complete enough to classify all 34 samples of leukemia cancer either AML or ALL. However, the first works [14],[4]

[a] The k in the gamma option means the number of attributes in the input data.

that bas been carried out before, required about 50 genes are required to correctly classify only 29 of the 34 samples. Furey *et al.,*[13] proposed feature selection method is almost similar to one that used in Golub *et al.,*[14] and Slonim *et al.,*[4]. The work classifies the testing set and correctly produced results between 30 to 32 out of the 34 samples. Mukherjee *et al.,*[1] has suggested neighborhood analysis method for genes selection and using SVM as a classifier. The accuracy of this work was reported to be 100% using LOOCV procedure on training data and also 100% on testing data. But Mukherjee *et al.,*[12] required 49 genes to make perfect classification in all 34 samples. Liu *et al.,*[7] incorporated genetic algorithm into weight voting classifier that selects 29 genes to correctly classifies 30 of the 34 samples.

4.4. Biological Plausibility for Informative Genes in Cancer Datasets

A few informative genes used in the leukemia dataset have the potential to be marked as biological plausibility. In this work, the genes found in ALL specifically *CD33 antigen* (M23197) and MB-1 encoded cell surface proteins for *monoclonal antibodies* (U05259) have been demonstrated to be useful in distinguishing *lymphoid* from *myeloid lineage cells* [1], [14]. In additional, Golub *et al.,*[14] reported that one of the informative genes encoded *topoisomerase II* (J04088) that is the principal target of the *antileukaemic drug etoposide*. *Topoisomerase II* (J04088) also reported as an informative gene in this work. In particular, *adipsin gene* (M84526) of leukemia that has been reported as informative [11] also found as informative gene in this work.

5. Conclusion

In this paper, we have investigated and solved the problem of efficient selection of good small subset of genes from thousands of genes measured on microarray. A major goal of this work is to find the minimum number of informative genes from gene expression microarray data, which maximize the classification accuracy. We have introduced new approach by applying GA wrapper to select genes combinations for classification and then uses the accuracy from SVM classifier for the fitness in GA. Finally a leukemia benchmark dataset are used to test the approach.

As our results the whole amount of features can contribute negative impact on classification performance because most of features in the data have much noise. We have shown that the selection of a few features using the proposed approach can lead to significant improvements in classification accuracy. So we might be able to understand the biological significance of genes because the approach can distinguish classes of samples with only a few genes. Moreover, the model of chromosome representation in the approach has reduced the combinations number of feature subsets with fitting of the chromosome length. Besides, the model has decreased the complexity searching on features space.

We are currently studying more on principled design of fitness using domain knowledge as well as mathematically well-founded tools of multi-attribute utility theory. In future new domain related genetic crossover operator would be undertaken.

Acknowledgements

This work was supported by National Science Fellowship research program sponsored by Malaysian Ministry of Science, Technology and Environments (MOSTE).

References

1. A. Ben-Dor, L. Bruhn, N. Friedman, I. M. Schummer and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of Computational Biology,* 7: 559—584, 2000.
2. D. Eads, D. Hill, S. Davis, S. Perkins, J. Ma, R. Porter and J. Theiler. Genetic algorithms and support vector machines for time series classification. *5th Conference on the Application and Science of Neural Networks, Fuzzy Systems and Evolutionary Computation,* 74—85, 2002.
3. D. J. Lockhart and E. Winzeler. Genomics, gene expression and DNA arrays. *Nature Insight,* 405:827—836, 2000.
4. D. K. Slonim, P. Tamayo, J. P. Mesirov, T. Golub and E. Lander. Class prediction and discovery using gene expression data. *Proceedings of the 13th Annual Conference on Computational Molecular Biology*, 263—272, 2000.
5. F. Naef, D. A. Lim, N. Patil and M. O. Magnasco. From features to expression: High-density oligonucleotide array analysis revisited. *Proceedings of the DIMACS Workshop on Analysis of Gene Expression Data,* 2001.
6. J. Bins and B. A. Draper. Feature selection from huge feature sets. *Proceeding of International Conference on Computer Vision,* 2:159—165, 2001.
7. J. Liu, H. Iba and M. Ishizuka. Selecting informative genes with parallel genetic algorithms in tissue classification. *Genome Informatics.* 12:14—23, 2001.
8. J. Sepulveda-Sanchis, G. Camps-Valls, E. Soria-Olivas, S. Salcedo-Sanz, C. Bousono-Calzon, G. Sanz-Romero and J. Marrugat. Support vector machines and genetic algorithms for detecting unstable angina. *Computers in Cardiology, IEEE Computer Society Press, Menphis (USA),* 2002.
9. J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters,* 9(3):293—300, 1999.
10. J. Yang and V. Hanovar. Feature subset selection using genetic algorithm. *Journal of IEEE Intelligent Systems.* 13:44—49.
11. S. B. Cho and H. H. Won. Machine learning in DNA microarray analysis for cancer classification. *1st Asia-Pacific Bioinformatics Conference.* 19, 2003.
12. S. Mukherjee, P. Tomayo, D. K. Slonim, A. Verri, T.Golub, J. P. Mesirov and T. Poggio. Support vector machine classification of microarray data. *AI Memo*. 1667. Massachusetts Institute of Technology. 1999.

13. S. T. Furey, N. Cristianini, N. Duffy, M. Schummer, D. W. Bednarski and D. Haussler. Support vector machine classification and validation of cancer tissue sample using microarray expression data. *Bioinformatics,* 16(10):906—914, 2000.
14. T. R. Golub, D. K. Slonim, P. Tomayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A Caligiuri, C. D. Bloofield and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science,* 286:531—537, 1999.
15. V. Vapnik. *The nature of Statistical Learning Theory*. New York, USA, Springer, 1995.
16. Z. Michalewicz. *Genetic Algorithms + Data Structures = Evolution Programs.* Springer-Verlag, New York, Third edition, 1996.