GA-Based Classifier with SNR Weighted Features for Cancer Microarray Data Classification

Supoj Hengpraprohm
Faculty of Science and Technology, Nakhon Pathom Rajabhat University
Nakhon Pathom, Thailand
Email: supojn@yahoo.com

***Abstract*—This work presents the method to classify the gene expression cancer data –Microarray data. The proposed method combines two techniques: classification and feature selection. The classification technique used in this work is Genetic Algorithm (GA) and the feature selection technique is Signal-to-Noise Ratio (SNR). Lymphoma and Leukemia datasets are used to test the performance of the proposed method and 10-Folds cross validation technique is applied to report the experimental results in term of classification accuracy. The results show that the proposed method yields the best result comparing with the simple GA-based classifier in both classification accuracy and the number of generations to found the solutions. Additionally, the results are compared to the other classification and feature selection techniques reported in the literature and it is found that the proposed method achieves a good result, especially, in the Lymphoma dataset the proposed method is the best.**

***Index Terms*—microarray data; genetic algorithm; signal-to-noise ratio; classification; feature selection**

I. Introduction

Microarray is a popular technique to study the mechanism of living cells in molecular level. This technique makes it possible to study gene expression of tens to hundred thousand genes simultaneously. The microarray dataset comprises of a small number of samples with very high features. Therefore, the effectiveness of data analysis with the techniques of data mining, machine learning or statistics will be decreased because these techniques require sufficient samples with a few features.

Machine learning techniques [1]-[4] were applied to microarray data to enhance the efficiency of microarray data analysis including evolutionary computation approach [5]. Advances in computer performance enable evolutionary computation approach to solve difficult real-world optimization problems. In our previous work [6], we found that GA-Based classifier can achieve a good result in terms of classification accuracy comparing with other machine learning methods. In [7], the results show that GA-Based classifier is more efficient to classify microarray data in terms of the accuracy and the number of generations comparing with other evolutionary computation method namely Genetic Programming.

In microarray data classification, the learning process usually comprises of two parts. The first one is to find a subset of features which suitable to the next part. The other part is to build the classifier with the subset of features getting from the first part. This work presents the efficiency of the GA-based classifier with the feature selection technique namely SNR (Signal-to-Noise Ratio) to weight the feature for the classifier.

The paper is organized as follows: section The paper is organized as follows: Section II presents background knowledge. Section III describes the data and method implemented in this research. Section IV shows the result of the experiment. Conclusions are presented in Section V.

II. Background Knowledge

A. Microarray Data

Microarray is a technique that presents thousands of expression level of genes simultaneously. This technique makes it possible to analyze and observe a complex organism in details. Microarray data is generated by hybridization of sample DNA labeled with red-fluorescent (dye Cy5) and DNA library labeled with green-fluorescent (dye Cy3) in equal quantities. Then, the slide of hybridization of DNA is imaged by a scanner that measured each dye. The process of microarray technique is shown in Fig. 1. The expression level of genes is defined as follows:

$$gene\_expression = \log_2 \frac{Int(Cy5)}{Int(Cy3)} \qquad (1)$$

where Int(Cy5) and Int(Cy3) are the intensities of red and green colors which scanned after the hybridization of the samples with the arrayed DNA probes.

Manuscript received Feb. 15, 2013; revised May 11, 2013

©2013 Engineering and Technology Publishing

doi: 10.12720/ijsps.1.1.29-33

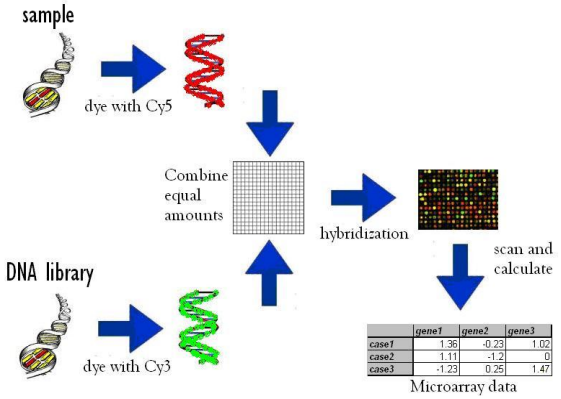

Figure 1. The process of microarray technique.

B. Genetic Algorithm

Genetic Algorithm (GA) [8] is a search method that imitates natural evolution and selection. The representation of the solution is a chromosome which is represented by fixed length binary string. The algorithm of GA is shown in fig. 2 and details of each step are as follows:

1) Generate an initial population of solutions: The initial solutions are created to full the population. There will be a large variation of solution structures through the process of this random generation.

2) Evaluate each solution by a fitness function: Each solution is evaluated to determine its fitness. The evaluation function, called "fitness function", is an important element in Genetic Algorithm. The fitness function is problem specific. Each solution will have a measure of goodness associated with it.

3) Create a new population by genetic operators: Genetic operations on the population have the goal of generating a new population that has better quality solutions. There are three genetic operators: reproduction, crossover, and mutation.

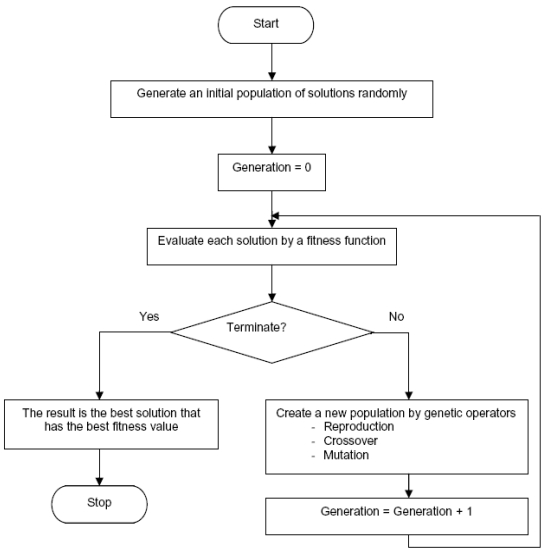

Figure 2. The algorithm of Genetic Algorithm.

Reproduction: A number of good solutions are selected based on their fitness value to be reproduced to the next generation. This process conserves good solutions.

Crossover: This operator recombines parts from two good solutions, called "parents", to create new solutions, called "offspring". Two good solutions are selected. The probability of a solution being selected is proportional to its fitness. The crossover points, which determine the location to exchange parts, are randomly selected. The strings after the crossover point from parents are exchanged. This process creates two new offspring.

Mutation: To maintain diversity in the population and to encourage exploration of different solutions, the mutation operator changes some part of a solution randomly. A solution is selected randomly and a location to be changed is selected. A value is mutated by changing it with inverted value (0 and 1).

C. Signal-to-Noise Ratio Feature Selection

There are two major feature selection approaches: filter and wrapper. Filter approach selects informative features regardless of classification algorithms according to some scoring metric, while the wrapper approach selects features with regard to a particular learning algorithm. Because the wrapper approach uses the target learning algorithm to find the best subset of features, it takes a longer computation time in the process than the filter approach.

The filter approach is simpler and it is fast enough to obtain a good performance regardless of classification algorithms. There are many metrics to measure the importance of features, for example, Pearson Correlation Coefficient (PC), Spearman Correlation Coefficient (SC), Euclidean Distance (ED), Cosine Coefficient (CC), Information Gain (IG), Mutual Information (MI) and Signal-to-Noise Ratio (SNR) (see [9], [10] for more details).

Many researchers reported that SNR metric provided the best result for classification [11]-[13]. We used this approach in the experiment. SNR is a statistical metric that measures effectiveness of a feature in identifying a class out of another class. The signal-to-noise ratio of a feature is defined as follows:

$$F = \left| \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \right| \quad (2)$$

where μ_1 and μ_2 denote the mean expression level for the samples in class 1 and class 2 respectively. σ_1 and σ_2 denote the standard deviation for the samples in each class.

III. The Experimental Setting

A. The Datasets

Two datasets of benchmark cancer microarray data are used to test the proposed method. There are Lymphoma and Leukemia datasets. The details of each dataset are as follows:

©2013 Engineering and Technology Publishing

1) Lymphoma dataset: comprises of 47 samples with 4,026 features. It is classified as 24 germinal centre B-likes (GCs) and 23 activated B-likes (ACs) [14].

2) Leukemia dataset: comprises of 72 samples with 7,129 features. It is classified as 47 ALLs and 25 AMLs [15].

B. The GA-based Classifier

In order to construct the GA-based classifier, the chromosome length is fixed to *n*. Each gene in the chromosome is the position gene (feature) in the microarray data as shown in fig. 3. The chromosome is divided into 2 groups equally (n/2 genes) as shown in fig. 4.

To classify the data, the summation of gene expression values selected by GA in each group is calculated and compared between groups. If the summation of the first group is greater than the other, it is classified as class 1; otherwise, it is classified as class 2. The GA parameters used in this work are shown in Table I.

C. The Method Implemented

The GA-Based classifier described in Section 3.2 is used in the experiment. The features of data are weighted by SNR score (eq. 2). These weighted features are used in the gene in chromosome of the solutions of GA-Based classifier. To evaluate the performance of a classifier, we used a method known as 10-Fold cross validation. The records of dataset are divided into 10 subgroups with randomly chosen records (without replacement). Nine subgroups are used as training set and the rest subgroup is used as a test set. We exchange a test set of data through all subgroups and evaluate an expression in terms of its accuracy, sensitivity and specificity which are defined as follows:

$$Accuracy = \frac{(TP + TN)}{N} \tag{3}$$

where *N* is a total number of cases, *TP* is a total number of affected subjects correctly classified, *TN* is a total number of normal subjects correctly classified, and TP+TN is the total number of subjects correctly classified.

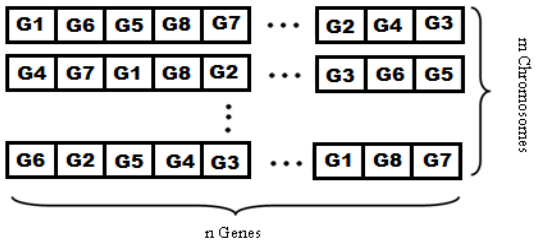

Figure 3. The Chromosome used in this work.

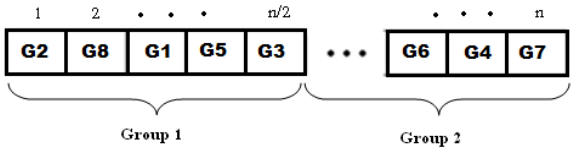

Figure 4. The representation of GA-based classifier.

TABLE I. THE GA PARAMETERS USED IN THIS WORK

Population size (number of chromosomes in each generation)	100
Chromosome length (n value in figure 3 and figure 4)	20
Generation	1000
Reproduction Rate	10%
Crossover Rate (Single point crossover)	80%
Mutation Rate	10%
Selection Method	Tournament (size = 5)

IV. THE EXPERIMENTAL RESULTS

Due to the GA is a randomize algorithm, the experiment is repeated and the result is reported from the average of 10 runs (using 10-Folds cross validation method, the total number of experiment in each data set is 100).

The experimental results are compared with the simple GA-Based classifier (the result shows in Table II). The results show that the GA-Based classifier with SNR weighted features yields the best performance in both classification accuracy and the number of generations to found the solution.

TABLE II. COMPARISON OF THE EFFICIENCY OF GA-BASE WITH SNR WEIGHTED FEATURE AND SIMPLE GA-BASED CLASSIFIERS

Classifier	**Data sets**			
	Lymphoma		***Leukemia***	
	Accuracy	***#gen***	***Accuracy***	***#gen***
GA-based with SNR weighted feature	92.33 ±2.50	4	91.94 ±2.15	247
Simple GA-based	82.55 ±4.46	69	89.58 ±1.88	438

In addition, the paper compares the results with many feature selections and classifiers reported in [9], [10] in 2 datasets (as shown in Table III). The feature selection methods are Pearson's and Spearman's correlation coefficients (PC, SC), Euclidean distance (ED), cosine coefficient (CC), information gain (IG), mutual information (MI) and signal to noise ratio (SNR). The classifiers are Multi-layer perceptron (MLP), K-nearest neighbour (KNN), support vector machine (SVM) and structure adaptive self–organizing map (SASOM).

In Table III, the values with highlight are better than the proposed method. The comparison shows that the proposed method gives better performance than other methods about 76.19%, and 100.00% in the Leukemia and Lymphoma dataset respectively.

©2013 Engineering and Technology Publishing

TABLE III. COMPARISON OF THE ACCURACY OF THE PROPOSED METHOD WITH OTHER METHODS

Classifier	Feature Selection	Dataset	
		Leukemia	Lymphoma
MLP	PC	***97.1***	64.0
	SC	82.4	60.0
	ED	91.2	56.0
	CC	***94.1***	68.0
	IG	***97.1***	92.0
	MI	58.8	72.0
	SN	76.5	76.0
SASOM	PC	76.5	48.0
	SC	61.8	68.0
	ED	73.5	52.0
	CC	88.2	52.0
	IG	91.2	84.0
	MI	58.8	64.0
	SN	67.7	76.0
SVM (linear)	PC	79.4	56.0
	SC	58.8	44.0
	ED	70.6	56.0
	CC	85.3	56.0
	IG	***97.1***	92.0
	MI	58.8	64.0
	SN	58.8	72.0
SVM (RBF)	PC	79.4	60.0
	SC	58.8	44.0
	ED	70.6	56.0
	CC	85.3	56.0
	IG	***97.1***	92.0
	MI	58.8	64.0
	SN	58.8	76.0
KNN (Cosine)	PC	***97.1***	60.0
	SC	76.5	60.0
	ED	85.3	56.0
	CC	91.2	60.0
	IG	***94.1***	92.0
	MI	73.5	80.0
	SN	73.5	76.0
KNN (Pearson)	PC	***94.1***	76.0
	SC	82.4	60.0
	ED	82.4	68.0
	CC	***94.1***	72.0
	IG	***97.1***	92.0
	MI	73.5	64.0
	SN	73.5	80.0
GA-based classifierwith SNR weighted features		91.9	92.3

V. CONCLUSIONS

This paper presents the efficiency of GA-based classifier with SNR weighted features comparing with simple GA-based classifier in microarray binary classification. The tested datasets were Lymphoma and Leukemia datasets. The experiment took 10-folds cross validation method. The results show that GA-based classifier with SNR weighted features is more efficient to classify microarray in terms of classification accuracy and the number of generations. Furthermore, the proposed method yields a good result comparing with other classifier and feature selection methods.

REFERENCES

[1] J. Khan, J. Wei, M. Ringnér, L. Saal, M. Ladanyi, F. Westermann., F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673-679, 2001.

[2] T. Furey, N. Cristianini, N. Duffy, D. 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.

[3] A. Boulesteix, G. Tutz, and K. Strimmer, "A cart-based approach to discover emerging patterns in microarray data," *Bioiformatics*, vol. 19, no. 18, pp. 2465-2472, 2003.

[4] S. Hengpraprohm and P. Chongstitvatana, "Diffuse large B-Cell lymphoma classification using genetic programming classifier," in *Proc. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, La Jolla, CA, USA, November 14-15, 2005, pp. 333-338.

[5] T. Back, U. Hammel, and H. Schwefel, "Evolutionary computation: comments on the history and current state," *IEEE Transactions on Evolutionary Computation*, vol. 1, pp. 3-17, 1997.

[6] S. Hengpraprohm, S. Mukviboonchai, R. Thammasang, and P. Chongstitvatana, "A GA-Based classifier for microarray data classification," in *Proc. International Conference on Intelligent Computing and Cognitive Informatics*, Kuala Lumpur, Malaysia, June 22-23, 2010, pp.199-202.

[7] S. Mukviboonchai and S. Hengpraprohm, "Evolutionary computation approach to cancer microarray data classification: a comparative study of GA- and GP-based classifier," in *Proc. 2nd International Conference on Intellectual Technology in Industrial Practice*, Changsha, China, September 8-9, 2010.

[8] J. H. Holland, *Adaptation in Natural and Artificial System*, University of Michigan Press, 1975.

[9] C. Park and S. B. Cho, "Evolutionary ensemble classifier for lymphoma and colon cancer classification," in *Proc. Congress on Evolutionary Computation*, 2003, pp. 2378-2392.

[10] S. B. Cho and H. H. Won, "Machine learning in DNA microarray analysis for cancer classification," *in Proc. the First Asia-Pacific bioinformatics conference on Bioinformatics*, 2003, pp.189 -198.

[11] D.K. Slonim *et al*., "Class prediction and discovery using gene expression data," in *Proc. the 4th Annual Int. Conf. on Computational Molecular Biology*, 2000, pp.263-272.

[12] J. Ryu and S. B. Cho, "Gene expression classification using optimal feature/classifier ensemble with negative correlation," in *Proc. Int. Joint Conf. on Neural Network*, 2002, pp. 198-203.

[13] C. J. Huang and W. C. Liao. "A comparative study of feature selection methods for probabilistic neural networks in cancer classification," in *Proc. the 15th IEEE Int. Conf. on Tools with Artificial Intelligence*, 2003.

[14] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. JR. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, "Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature 403(2000)*, pp. 503-511.

©2013 Engineering and Technology Publishing

[15] 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 286(1999)*, pp. 531-537.

Supoj Hengpraprohm received B.Sc. degree in Computer Science from the Rajabhat Institute Nakorn Pathom, Thailand in 1999, M.Sc. degree in Computer Science from the Chulalongkorn University, Thailand in 2002 and Ph.D. degree in Computer Engineering from the Chulalongkorn University, Thailand in 2009. Presently, he is a lecturer with the faculty of Science and Technology, Nakhon Pathon Rajabhat University, Thailand. His research interests include Evolutionary Computation, Bioinformatics, Machine Learning and Data Mining.

©2013 Engineering and Technology Publishing