2010 International Conference on Intelligent Computing and Cognitive Informatics

A GA-Based Classifier for Microarray Data Classification

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

Suvimol Mukviboonchai
Faculty of Science and Technology
Nakhon Pathom Rajabhat University
Nakhon Pathom, Thailand
suvimol@npru.ac.th

Rujirawadee Thammasang
Faculty of Science and Technology
Nakhon Pathom Rajabhat University
Nakhon Pathom, Thailand
rujith01@msn.com

Prabhas Chongstitvatana
Department of Computer Engineering
Chulalongkorn University
Bangkok, Thailand
prabhas@chula.ac.th

***Abstract*— This work presents an algorithm for generating the GA-based (Genetic Algorithm) classifier for microarray data classification. The microarray dataset comprises of a small number of samples with very high features. In order to construct the GA-based classifier, a number of informative features (genes) are selected. These features are divided into 2 groups (10 features or less in each group). The summation of gene expression values selected by GA in each group is then 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. In the experiment, 3 microarray benchmark datasets for the 2-class problem are used. There are Lymphoma, Leukemia and Colon datasets. 10-Folds cross validation is used to test the performance of the proposed method. The experimental results show that the proposed GA-based classifier yields a good effectiveness in the 2-class microarray data classification comparing with the other methods.**

Keywords- Microarray, Data Classification, Genetic Algorithm, Feature Selection, Learning Algorithm

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 hundreds 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.

Recently, there are many researches that study in microarray data classification with various techniques such as Artificial Neural Network (ANN) [1], Support Vector Machine (SVM) [2], Decision Tree [3] and Genetic Programming (GP) [4]. The efficient of the data classification comes from two parts: feature selection and learning algorithms to build the classifier.

In microarray data classification, the learning process usually comprises of two parts. The first 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. Many researchers reported that Genetic Algorithm (GA) yields a good result in finding subset of features part to improve the accuracy of microarray data classification [5-6]. Furthermore, GA is widely used to solve various problems including data classification [7-8].

This work presents an algorithm for generating the GA-based classifier for microarray data classification. The proposed algorithm is able to find a subset of features and act as a classifier itself by dividing selected features into two groups and comparing the summation of gene expression value in each group.

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 Figure 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.

B. Genetic Algorithm

Genetic Algorithm (GA) [9] is a search method that imitates natural evolution and selection. The representation of the solution is a chromosome which is represented by

978-0-7695-4014-6/10 $26.00 © 2010 IEEE
DOI 10.1109/ICICCI.2010.62

fixed length binary string. The algorithm of GA is shown in figure 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.

- 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 invert value (0 and 1).

III. Data and Method Implemented

A. Datasets

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

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) [10].

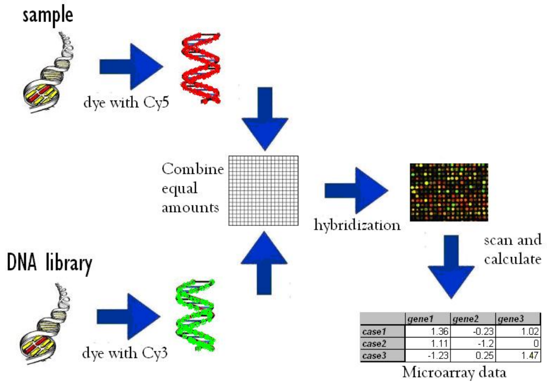
Leukemia dataset: comprises of 72 samples with 7,129 features. It is classified as 47 ALLs and 25 AMLs [11].

Colon cancer dataset: comprises of 62 samples with 2,000 features. It is classified as 40 cancers and 22 normals [12].

B. The proposed 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 figure 3. The chromosome is divided into 2 groups equally (n/2 genes) as shown in figure 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.



	gene1	gene2	gene3
case1	1.36	-0.23	1.02
case2	1.11	-1.2	0
case3	-1.23	0.25	1.47

Figure 1. The process of microarray technique.

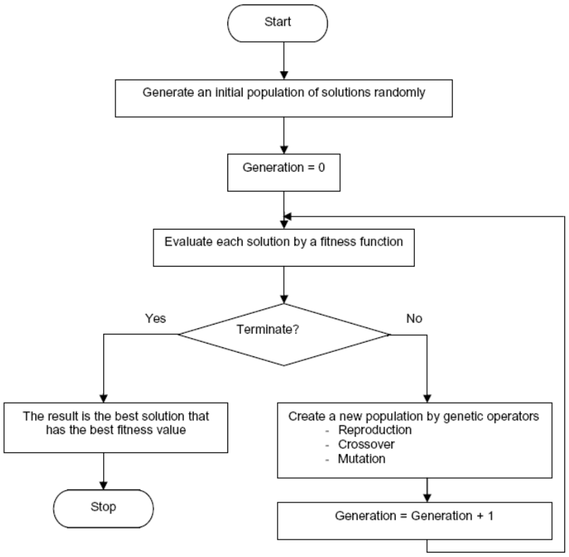

Start
Generate an initial population of solutions randomly
Generation = 0
Evaluate each solution by a fitness function
Terminate?
Yes
No
The result is the best solution that has the best fitness value
Stop
Create a new population by genetic operators
- Reproduction
- Crossover
- Mutation
Generation = Generation + 1

Figure 2. The algorithm of Genetic Algorithm.

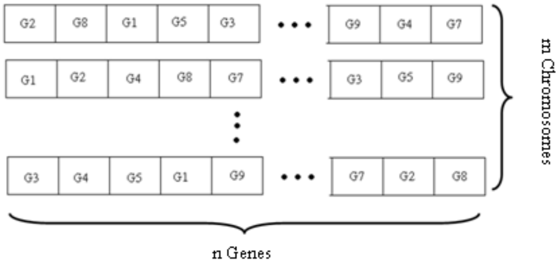

G2 G8 G1 G5 G3 • • • G9 G4 G7
G1 G2 G4 G8 G7 • • • G3 G5 G9
G3 G4 G5 G1 G9 • • • G7 G2 G8
m Chromosomes
n Genes

Figure 3. The Chromosome used in this work.

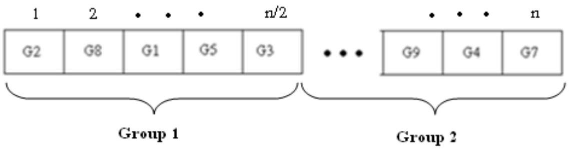


Figure 4. The representation of GA-based classifier.

TABLE I. THE GA PARAMETERS USED IN THIS WORK

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

IV. EXPERIMENTAL RESULTS

To evaluate the performance of a classifier, we used a method known 10-Folds cross validation. N records of data are divided into 10 groups equally, 9 groups are used as training set and one group is used as a test. We exchange a test data through 10 groups and evaluate an expression in terms of its accuracy defined as follows:

$$Accuracy = \frac{TP+TN}{N} \qquad (2)$$

where N is the total number of test 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.

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).

We compare the experimental results with many feature selections and classifiers reported in [13-14] in 3 datasets (as shown in Table 2). 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 2, the values with highlight are better than our method. The comparison shows that, the proposed method gives the better performance than other methods about 69.05%, 83.33% and 88.10% in the Leukemia, Colon and Lymphoma dataset respectively.

V. CONCLUSIONS

The experimental results suggested that the proposed GA-based classifier can achieve a good result in term of classification accuracy comparing with other methods. Due to the proposed algorithm is able to find a subset of features and act as a classifier itself by dividing selected features into two groups and comparing the summation of gene expression value in each group, the feature set selected is very suitable for the classifier itself. Whereas the other methods separate between the feature selection and creation of a classifier, the feature set may be not suitable for the classifier creation step.

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

Classifier	**Feature Selection**	**Dataset**		
		Leukemia	**Colon**	**Lymphoma**
MLP	PC	***97.1***	74.2	64.0
	SC	82.4	58.1	60.0
	ED	***91.2***	67.8	56.0
	CC	***94.1***	***83.9***	68.0
	IG	***97.1***	71.0	***92.0***
	MI	58.8	71.0	72.0
	SN	76.5	64.5	76.0
SASOM	PC	76.5	74.2	48.0
	SC	61.8	45.2	68.0
	ED	73.5	67.6	52.0
	CC	88.2	64.5	52.0
	IG	***91.2***	71.0	84.0
	MI	58.8	71.0	64.0
	SN	67.7	45.2	76.0
SVM (linear)	PC	79.4	64.5	56.0
	SC	58.8	64.5	44.0
	ED	70.6	64.5	56.0
	CC	85.3	64.5	56.0
	IG	***97.1***	71.0	***92.0***
	MI	58.8	71.0	64.0
	SN	58.8	64.5	72.0
SVM (RBF)	PC	79.4	64.5	60.0
	SC	58.8	64.5	44.0
	ED	70.6	64.5	56.0
	CC	85.3	64.5	56.0
	IG	***97.1***	71.0	***92.0***
	MI	58.8	71.0	64.0
	SN	58.8	64.5	76.0
KNN (Cosine)	PC	***97.1***	71.0	60.0
	SC	76.5	61.3	60.0
	ED	85.3	***83.9***	56.0
	CC	***91.2***	***80.7***	60.0
	IG	***94.1***	74.2	***92.0***
	MI	73.5	74.2	80.0
	SN	73.5	64.5	76.0
KNN (Pearson)	PC	***94.1***	77.4	76.0
	SC	82.4	67.7	60.0
	ED	82.4	***83.9***	68.0
	CC	***94.1***	***80.7***	72.0
	IG	***97.1***	***80.7***	***92.0***
	MI	73.5	***80.7***	64.0
	SN	73.5	71.0	80.0
GA-based classifier		89.5	80.3	82.5

ACKNOWLEDGMENT

This research has been supported by Research and Development Institute, Nakhon Pathom Rajabhat University (NPRU), Thailand.

REFERENCES

[1] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," NATURE MEDICINE, Vol. 7(6), 2001, pp. 673 – 679.

[2] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," BIOINFORMATICS, Vol. 16(10), 2000, pp. 906 – 914.

[3] A. L. Boulesteix, G. Tutz and K. Strimmer, "A CART-based approach to discover emerging patterns in microarray data," BIOIFORMATICS, Vol. 19(18), 2003, pp. 2465 – 2472.

[4] J. H. Hong and S. B. Cho, "The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming," Artificial Intelligence in Medicine, Vol. 36, 2006, pp. 43-58.

[5] T. C. Lin, "Breast Cancer Classification and Biomarker Discovery on Microarray Data Using Genetic Algorithms and Bayesian Classifier," http://hdl.handle.net/2377/11177, Feng Chia University, 11-Feb-2009.

[6] C. H. Ooi and P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," BIOIFORMATICS, Vol. 19(1), 2003, pp. 37 – 44.

[7] S. Bandyopadhyay, C. A. Murthy and S. K. Pal, "Pattern classification with genetic algorithms," Pattern Recognition Letters, Vol. 16, 1995, pp. 801-808.

[8] S. Bandyopadhyay, C. A. Murthy and S. K. Pal, "VGA-Classifier: Design and Applications," IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART B: CYBERNETICS, Vol. 30(6), 2000, pp. 890-895.

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

[10] 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) : 503-511.

[11] 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) : 531-537.

[12] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack and A. J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," Proceedings of National Academy of Sciences of the United States of American, 96, 1999, pp. 6745-6750.

[13] S. B. Cho and H. H. Won, "Machine Learning in DNA Microarray Analysis for Cancer Classification," the Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003, 19: 189 – 198, 2003.

[14] J. H. Hong and S. B. Cho, "The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming"., Artificial Intelligence in Medicine, Vol. 36, 2006, pp. 43-58.