cDNA Microarray Data Based Classification of Cancers Using Neural Networks and Genetic Algorithms

Hyun Sung Cho*, Tae Seon Kim**, Jae Woo Wee*, Sung Mo Jeon*, and Chong Ho Lee*

*Department of Information Technology and Telecommunications, Inha University, Korea
**School of Computer Science and Electronic Engineering, Catholic University of Korea, Korea,
tkim@catholic.ac.kr

ABSTRACT

In this paper, intelligent cancer classification method using cDNA microarray data was developed using neural networks and genetic algorithms. For classification, selection of gene expression data from microarray data are performed as a first step to find highly related genes to disease among microarray data. The fitness values for selection results are determined based on neural network model prediction results. To overcome the limitation of pre-determined number of gene selection method, variable-length chromosome based genetic algorithms are applied. For performance evaluation, pre-tested tumor data are used for classification model and evaluated through blind test. Experimental results are compared with principal component analysis (PCA) based statistical methods, and the test results showed that proposed method has superior classification results (96% accuracy) compare to statistical method (92% accuracy).

Keywords: microarray, neural networks, genetic algorithms, gene selection, cancer classification

1 INTRODUCTION

Recently, as a new and advanced way, gene expression data from cDNA microarrays were widely applied to many areas, including medical, environmental, and agricultural areas, and several research results showed their innovative technical progressive on various application areas such as medical diagnosis and bioinformatics. Among them, classification of cancers using cDNA microarrays data is considered as one of most difficult, but essential research area, since the way of medical treatment is varies through cancer types. Also, the time and cost of medical diagnosis are the other difficulties for prompt medical treatment. To solve these difficulties, cDNA microarray data based disease diagnosis and classification approaches are under studied from many researchers. Ben-Dor et al. developed hierarchical clustering method for analysis of gene expression data [1]. They define an appropriate stochastic error model on the input, and prove that under the conditions of the model, the algorithm recovers the cluster structure with high probability. Slonim et al. developed the classification of patient samples, and applied to the problem of classifying two types of acute leukemia, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Including above two research examples, most of conventional approaches are based on statistical theory based data clustering and data mining schemes were applied to solve this problem. However, for proper analysis, statistical way requires various assumptions and formats on gene expression data, which is not possible to find in practical data set. One of alternative way to resolve these problems is application of intelligent algorithms. Khan et al. used neural networks on classification and diagnostic prediction of cancers based on gene expression profiling [3]. Blind test results showed potential applications of their method for tumor diagnosis. Li et al. also developed the gene selection method using genetic algorithm and k-nearest neighbor (KNN) model [4]. Using KNN model, they classified the DNA microarray data from the myeloid samples of the patients who have either acute lymphoblastic leukemia (ALL) or the acute myeloid leukemia (AML). However, they are still based on statistical theory, therefore, they require large volume of patient samples for high statistical confidence. In this paper, we proposed intelligent gene selection and cancer classification method using genetic algorithms and neural networks. For selection of gene expression data from microarrays, genetic algorithms are applied. The fitness values for selection results are determined based on neural network model prediction results. And also, instead of fixed-length chromosome based genetic algorithms, variable-length chromosome based genetic algorithms are used, and it can increase the robustness on biased test data. After selection of gene expression data, neural networks are used to classify the cancer types. To show the performance, equal experimental data set as Khan et al. [3] are used and experimental results are compared with principal component analysis (PCA) based statistical methods. Blind test results showed that proposed method has superior classification results compare to statistical method.

2 INTELLIGENT CANCER CLASSIFICATION ALGORITHM

The proposed cDNA microarray data based cancer classification algorithm is consisting of two parts; gene selection from microarray data and disease type classification using selected gene expression data.

2.1 Gene Selection

As a first step to cancer classification, gene selection scheme is required to reduce the dimension of microarray data. Generally, compare to number of genes on microarrays, acquired number of patient sample is very limited and biased. In other words, the number of patient sample is not enough for statistical analysis, and even, they are not equally distributed for each disease types. Also, the size of gene expression data from microarrays is too large to calculate. For that reason, it is not desirable to use them directly in terms of performance and computational time. Therefore, from gene expression data set, valid genes that have clues for cancer need to be selected for analysis. For this, in this paper, genetic algorithms were applied to select valid genes and neural networks were used as classification model to support fitness values to genetic algorithms. Each chromosome has the information of index number for selected gene, and selected genes are fed into input of neural network for fitness evaluation.

One of dilemma on gene selection is determination of optimal number of genes for classification. Too many selections of gene data may cause undesirable noise addition to the valid data and also, it increase the computational complexity. In contrast, too small selections of gene data cause lack of information for classification. Also, the number of valid genes are very various dependent to disease type. However, fixed-length chromosome based genetic algorithms based gene selection methods need to pre-define the number of genes to select, and sometimes it made undesirable effects on classification model [5].

For variable-length chromosome based genetic algorithm, initial number of selected genes is set to 1. And then, the number of selected genes is increased to converge the fitness value to target value. Therefore, the proposed variable length chromosome scheme can determine the number of valid genes by itself, and it can increase the accuracy with reduced computational complexity. For genetic operations, tournament selection is used for reproduction. Also, to reserve the best chromosome set per each generation, elitism with n=1 scheme was applied. For mutation and crossover, simple mutation and uniform crossover method are used. And also, add-on-delete method was applied.

For gene selection, the fitness values of genetic algorithms are defined as the number of correctly classified patterns among input patterns as shown in equation (1),

$$f = 3 \times \sum_{i=1}^{n} \sum_{j=1}^{m} g_{ij} \tag{1}$$

$$g_{ij} = \begin{cases} 1, y_{ij} = t_{ij} \\ 0, y_{ij} \neq t_{ij} \end{cases} \tag{2}$$

where, f, y_{ij}, and t_{ij} represent fitness value, output value of neural prediction model, and target value for training input, respectively.

For fitness evaluation, 3-fold cross-validation scheme is used. This method randomly divides training sample data to three groups, and then uses two groups of data for training, and rest one group of data for evaluation. For full evaluation, three cross evaluation is required in this case.

2.2 Classification

After gene selection, selected genes are considered as inputs of neural classification model. It has been shown that the ability of neural networks to discover input/output relationships from limited data is useful in many areas, where numerous highly nonlinear characteristics are exist and experimental data for modeling is expensive to obtain [6]. The each of output neuron of network represents the types of disease. Therefore, in ideal case, only one neuron is excited for each input. Network is single layered perceptron neural network with sigmoid nonlinear transfer function.

3 RESULTS AND DISCUSSION

3.1 Microarray Experimental Data

For classification, “the small, round blue cell tumors” (SRBCTs) of childhood, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL), and Ewing family of tumors (EWS) are considered. These cancers are difficult to distinguish by light microscopy, and currently no single test can precisely distinguish these cancers [3]. Microarray experimental data set is obtained from National Human Genome Research Institute [7]. The total number of genes is 2308 and total number of data set is 88. Among 88 sample data, 63 samples are used for training pattern and rest 25 data are used for blind test. As described, 63 training patterns are divided to three groups for 3-fold cross-validation scheme.

3.2 Gene Selection Results

For gene selection, fixed-length chromosome based genetic algorithm and variable-length chromosome based genetic algorithm were performed and compared with principal component analysis (PCA) method results as reference [3]. Through 3-fold cross-validation scheme, 100 most significant genes are ranked among 2308. Table 1 shows the comparison of selected gene from proposed method and PCA method. Among 20 genes, only five genes are selected for both methods. However, all of selected gene from PCA method can be listed on top 100 among 2308 genes in fixed-length chromosome based genetic algorithm. The result of variable-length chromosome based genetic algorithm are not compared because it doe not use the ranking information for selection.

Table 1: Comparison of top 20 selected genes for fixed-length chromosome based genetic algorithm with neural networks (GA/ANN) and PCA method.

	GA/ANN		PCA	
Rank	Image Id.	Gene Label	Image Id.	Gene Label
1	784224	FGFR4	296448	IGF2
2	**770394**	FCGRT	**207274**	IGF2
3	**377461**	CAV1	841641	CCND1
4	1435862	MIC2	365826	GAS1
5	814260	FVT1	486787	CNN3
6	740801	BCKAD	**770394**	FCGRT
7	812105	AF1Q	**244618**	EST
8	898219	MSTH	233721	IGFBP2
9	**866702**	PTP	43733	GYG2
10	**244618**	EST	295985	EST
11	491565	CIT	629896	MAP1B
12	81518	apelin	840942	HLA-DPB1
13	839552	NRC1	80109	HLA-DQA1
14	814526	HSRNASEB	41591	MN1
15	796258	SGCA	**866702**	PTPN13
16	769716	NF2	357031	TNFAIP6
17	68950	cyclin E1	782503	EST
18	1473131	TLE2	**377461**	CAV1
19	143306	LSP1	52076	NOE1
20	**207274**	IGF2	811000	LGALS3BP

3.3 Classification Results

As described on microarray experimental data, 25 experimental data among 88 data are used for blind test and classification results are shown in Table 2. These data have the historical diagnosis results. To verify the robustness of classification algorithms, several data came from the patient who is not affected by SRBCTs. The first column represents the randomly labeled blind sample ID, and the second to the fifth column shows the output value of neural classifier. Ideally, value of "1" means that input patterns are classified to excited neuron corresponding to disease type with 100% of confidence. As shown in Tablw2, some of results are not converged perfectly into "1", but there's no result that excite two neurons simultaneously. Therefore, and for diagnosis, if the output value is bigger than 0.9, then it's considered as value "1". The sixth column and seventh column represents the classification results of GA/ANN and PCA method. N/A means that test sample is not from the SRBCTs affected patients (non-SRBCTs). Among 25 test samples, 20 samples are from SRBCTs and the rest 5 samples are from non-SRBCTs. Compare to historical diagnosis results, GA/ANN method perfectly classified for 20 SRBCTs test patterns. For non-SRBCTs, GA/ANN method missed only one same (sample id: 20), and the final classification accuracy is 96%. PCA method also missed for same sample, and also it missed for one of SRBCTs sample (sample id: 12). Therefore, final accuracy of PCA method shows 92%.

Table 2: Comparison of cancer classification results between GA/ANN method and PCA method

Sample ID	Types of SRBCTs				GA/ANN	PCA	Histological Diagnosis
	EWS	BL	NB	RMS			
1	0	0	**0.98**	0	NB	NB	NB-C
2	**1**	0.01	0	0	EWS	EWS	EWS-C
3	0	0	0	0	N/A	N/A	***C***
4	0	0	0	**1**	RMS	RMS	ARMS-T
5	0	0	0	0	N/A	N/A	***Sarcoma-C***
6	**1**	0	0	0	EWS	EWS	EWS-T
7	0	**1**	0	0	BL	BL	BL-C
8	0	0	**1**	0	NB	NB	NB-C
9	0	0.01	0	0.15	N/A	N/A	***Sk.Muscle***
10	0	0	0	**0.99**	RMS	RMS	ERMS-T
11	0	0	0	0	N/A	N/A	***Prostate Ca-C***
12	**1**	0	0	0	**EWS**	N/A	EWS-T
13	0	0	0	0.58	N/A	N/A	***Sk.Muscle***
14	0	0	**0.99**	0	NB	NB	NB-T
15	0	**0.99**	0	0	BL	BL	BL-C
16	0	0	**0.99**	0	NB	NB	NB-T
17	0	0	0	**1**	RMS	RMS	ARMS-T
18	0	**0.99**	0	0	BL	BL	BL-C
19	**1**	0	0	0	EWS	EWS	EWS-T
20	0	0	0	0	N/A	N/A	EWS-T
21	**1**	0	0	0	EWS	EWS	EWS-T
22	0	0	0	**1**	RMS	RMS	ERMS-T
23	0	0	**0.94**	0	NB	NB	NB-T
24	0	0	0	**1**	RMS	RMS	ERMS-T
25	0	0	**1**	0	NB	NB	NB-T

To show the variation of classification performance with variation of number of selected gene for classification, genes are selected from 1 to 100 and the results are shown in Figure 1. As shown in Figure 1, bigger selection number can't guarantee the classification capability in both GA/ANN and statistical method. Also, it shows the importance of the selection of optimum number of genes for classifications.

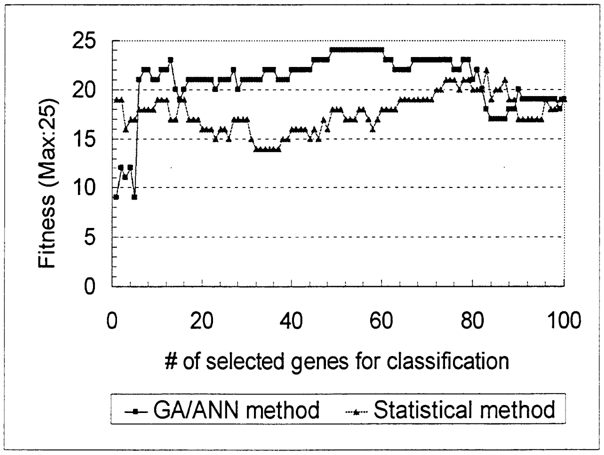

Figure 1. Classification capability with variation to number of selected gene for classification

Above GA/ANN method based on fixed-length chromosome based genetic algorithm with neural networks. However, to find optimum number of selected gene, all of genes are ranked and tested. It requires long computational times, and it can't applicable to practical analysis if the number of gene is increased. Therefore, variable-length chromosome based genetic algorithm is desirable. Figure 2 shows the Classification capability for variable-length chromosome based genetic algorithm with neural networks. By increasing the required selected genes for classification, proposed algorithm can remove the undesirable effect from unrelated gene expression data.

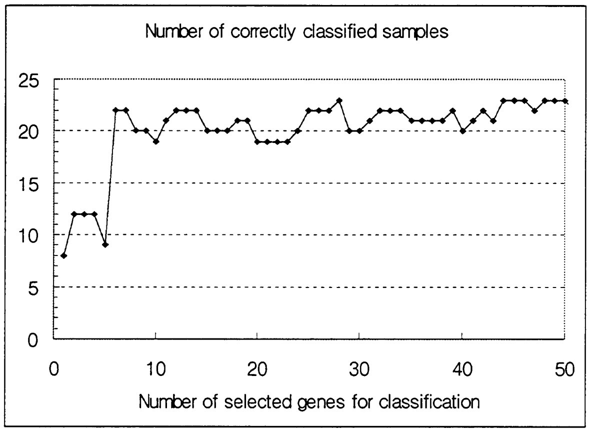

Figure 2. Classification capability for variable -length chromosome based genetic algorithm with neural networks

4 CONCLUSION

In this paper, cDNA microarray data based cancer classification algorithm using neural networks and genetic algorithms was. For classification, selection of gene expression data from microarray data are performed as a first step to find highly related genes to disease among microarray data, and the fitness values for selection results are determined based on neural network model prediction results. To overcome the limitation of pre-determined number of gene selection method, variable-length chromosome based genetic algorithms are applied. For performance evaluation, microarray experimental data from SRBCTs are used for classification model and evaluated through blind test. Experimental results are compared with principal component analysis (PCA) based statistical methods, and the test results showed that proposed method has superior classification results (96% accuracy) compare to statistical method (92% accuracy). With the successful implementation of proposed method, proposed method take a role of catalyst on practical applications of microarray data on disease diagnosis.

REFERENCES

[1] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns," *Journal of computational biology*, vol. 6, pp. 281-297, 1999.

[2] D. Slonim, P. Tamayo, J. Mesirov, T. Golub, and E. Lander, "Class prediction and discovery using gene expression data," *Proc. international conference on computational molecular biology*, Tokyo, Japan, pp. 263-272, 2000.

[3] J. Khan, J. Wei, M. Ringner, 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. 673-697, 2001

[4] L. Li, C. Weinberg, T. Darden, and L. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitive to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, pp. 1131-1142, 2001

[5] H. Cho, T. Kim, S. Jeon, J. Wee, and C. Lee, " Gene selection method using neural networks and genetic algorithm and its applications to classification of cancers," *Proc. KIEE*, July, 2002

[6] G. May, "Manufacturing ICs The Neural Way", *IEEE Spectrum*, vol. 31, no. 9, pp. 47-51, September, 1994.

[7] http://www.genome.gov/