DNA Microarray Data Analysis Using TNoM Score and NEWFM

Sang-Hong Lee and Xue-Wei Tian

***Abstract*—**An important aspect in microarray data analysis is the selection of an appropriate number of the most relevant genes among a large population of genes. In this study, we have proposed a gene selection using the threshold number of misclassification (TNoM) score. In the TNoM score, we selected an appropriate number of the top-ranked genes for microarray data analysis. In this study, from a colon cancer dataset and a leukemia dataset, we selected the top-ranked 93 colon cancer and 143 leukemia genes with ≤14 (colon cancer) and ≤13 (leukemia) TNoM scores from a total of 2000 colon cancer and 7129 leukemia genes. When the minimal 93 colon cancer and 143 leukemia genes were used as inputs for a neural network with weighted fuzzy membership functions (NEWFM), the performance accuracies were 96.77% and 100% for colon cancer and leukemia, respectively.

***Keywords*—**Gene selection, TNoM score, microarray, neural network.

I. Introduction

MANY recent studies using microarray technology have produced remarkable results, strongly indicating the usability of gene expression data as diagnostic tools. One of the major challenges of handling gene expression data is the large number of genes in the data sets. Although thousands of genes are evaluated simultaneously, most of them are irrelevant or insignificant to a clinical diagnosis [1][2][4], and it is important therefore to identify those genes that are relevant to a diagnosis. Statistical methods, such as the t-test [12], mutual information [2], and threshold number of misclassifications (TNoM) score [2] have been widely used in finding relevant genes. Feature selection is among the techniques that have contributed to an increase in an important methodology for pattern recognition and machine learning [6]. Moreover, by identifying the influence of each selected gene, feature selection in gene expression data can increase the comprehensibility of the generated results. In pattern recognition and machine learning, genetic algorithms [7][8] are typical search methods used in gene selection. As specific classifiers, support vector machine (SVM) [2][5][7][8], and rough set [11] have all been used to verify the efficiency of the selected genes.

In this study, we have proposed a gene selection that is the TNoM score among the statistical methods and a neural network with weighted fuzzy membership functions (NEWFM) [9][10] among the classifiers to classify tumor biopsies and normal biopsies from a colon cancer data set, and acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) from a leukemia data set. We selected from a colon cancer dataset and a leukemia dataset the top-ranked 93 colon cancer and 143 leukemia genes that have ≤ 14 (colon cancer) and ≤ 13 (leukemia) TNoM scores from a total of 2000 colon cancer) and 7129 leukemia genes.

Corresponding Author, Sang-Hong Lee is with Anyang University (corresponding author to provide e-mail: shleedosa@anyang.ac.kr).

Xue-Wei Tian is with Gachon University (e-mail: aitianxuemao @gmail.com).

II. Materials

Descriptions of the two data sets studied are as follows. The colon cancer data set involves comparing tumor and normal samples of the same tissue, and the leukemia data set involves comparing samples of acute myeloid leukemia and samples of acute lymphoblastic leukemia of the same tissue.

A. Colon Cancer Data Set

The colon cancer data set is a collection of expression measurements from colon biopsy samples reported by Alon et al. [1]. The data set consists of 62 samples of colon epithelial cells. These samples are divided into two variants of colon tissue: 40 colon tumor samples and 22 normal colon samples. Both sets of samples were collected from colon cancer patients. The tumor biopsies were collected from tumors, and the normal biopsies were collected from healthy parts of the colons of the same patients. The final assignments of the status of biopsy samples were made by pathological examination. Gene expression levels in these 62 samples were measured using high density microarrays. Two thousand genes were selected based on the confidence in the measured expression levels. The data set, representing 2000 genes across 62 samples, is available at http://genomics-pubs.princeton.edu/oncology/.

B. Leukemia Data Set

The leukemia data set is a collection of expression measurements reported by Golub et al. [4], and contains 72 samples. These samples are divided into two variants of leukemia: 25 samples of acute myeloid leukemia (AML) and 47 samples of acute lymphoblastic leukemia (ALL). The source of the gene expression measurements was taken from 63 bone marrow samples and nine peripheral blood samples. Gene expression levels in these 72 samples were measured using high density microarrays reporting the expression levels of 7129

genes. The data set, representing 7129 genes across 72 samples, is available at http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43.

III. Methods

A. Threshold Number of Misclassifications Score for Feature Selection

When gene expression data is analyzed and classified, it is of interest to find genes that have different expression patterns in different cell types. These genes are referred to as informative genes in that they are potentially relevant to the distinction between classes of cells, e.g., normal and tumor cells of the same tissue, as well as AML cells and samples of ALL cells of the same tissue. Most of the genes, however, are irrelevant for this distinction, and only a small fraction may play a major role in the biochemical pathways that underlie these distinctions.

A TNoM score is obtained for classifiers of two classes, e.g., normal and tumor or AML and ALL. The score is the minimum number of errors that can be achieved using one separator that assigns all samples on its left to one class label and the second class label to all samples on its right [2].

$$l(x \mid t,d,g) = \begin{cases} d & x_i[g] > t \\ -d & x_i[g] < t \end{cases}, \qquad (1)$$

where t is some threshold value and is a direction parameter of gene g in the ith sample. Rewriting it, the prediction class is simply . The error of such a predictor gene is the sum of prediction errors over all samples.

$$Err(d,t \mid g,l) = \sum 1\{l_i \neq sign(d \cdot (x_i[g] - t))\}, \qquad (2)$$

where is the expression value of gene g in the ith sample and is the label of the ith sample. The TNoM score of a gene is simply defined as:

$$TNoM(j,l) = \min_{d,t} Err(d,t \mid g,l), \qquad (3)$$

the number of errors made by the best rule. The intuition is that this number reflects the quality of decisions made based solely on the expression levels of this gene [2].

In this study, we selected from a colon cancer dataset and a leukemia dataset the top-ranked 93 (colon cancer) and 143 (leukemia) genes that have ≤ 14 (colon cancer) and ≤ 13 (leukemia) TNoM scores from a total of 2000 (colon cancer) and 7129 (leukemia) genes.

B. Neural Network with Weighted Fuzzy Membership Function

This study used a NEWFM for classifying tumor biopsies and normal biopsies from the colon cancer data set, and AML and ALL from the leukemia data set. The NEWFM is a supervised classification neuro-fuzzy system. Fig. 1 shows that the NEWFM contained three layers, i.e., the input, hyperbox, and class layers. The top-ranked 93 genes from the colon cancer data set and the top-ranked 143 genes from the leukemia data set were used as inputs for the NEWFM, as shown in Fig. 1.

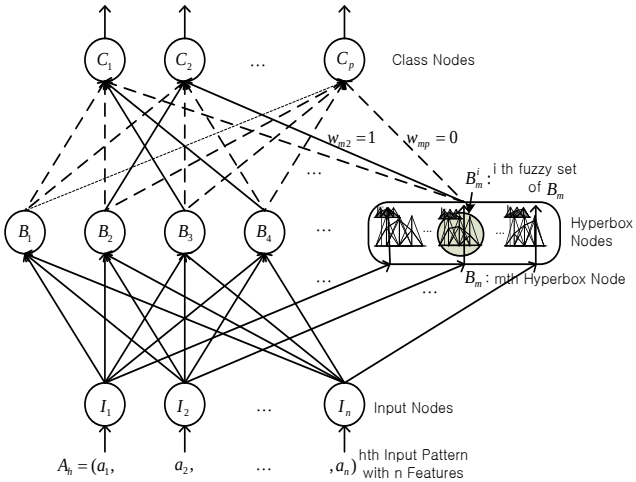

Fig. 1 Structure of the NEWFM [9][10]

IV. Experimental Results

In this study, tumor biopsies and normal biopsies, and AML and ALL were classified from the colon cancer and leukemia data sets, respectively. Classification accuracies are shown in Table I. The accuracy of NEWFM was compared with that determined by Cho [3], Guyon [5], and Wang [13].

TABLE I
THE CLASSIFICATION ACCURACY OF THE EXISTING METHODS

Data set	Colon	Leukemia
Cho et al. [3]	82.08 (10)	94.12 (17)
Guyon et al. [5]	90.32 (8)	100 (4)
Wang et al. [13]	91.9 (3)	100 (5)
Our study	96.77 (93)	100 (143)

V. Discussion

This study used two gene selection steps for microarray data analysis. We used the TNoM score in gene selection. In gene selection step, the top-ranked 93 colon cancer and 143 leukemia genes having ≤ 14 (colon cancer) and ≤ 13 (leukemia) TNoM scores were selected. The fuzzy membership functions of the NEWFM were used to classify tumor biopsies and normal biopsies from the colon cancer data set, and acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) from the leukemia data set, respectively

Acknowledgment

This study was supported by a grant of the Korean Health Technology R&D Project, Ministry of Health & Welfare, and Republic of Korea. (A112020)

This research was supported by MSIP (the Ministry of Science, ICT and Future Planning), Korea, under the IT-CRSP(IT Convergence Research Support Program) (NIPA-2013-H0401-13-1001) supervised by the NIPA (National IT Industry Promotion Agency).

REFERENCES

[1] Alon U, Barkai N, Notterman D, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Nat. Acad. Sci. USA 96, 6745-6750.

[2] Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z (2000) Tissue Classification with Gene Expression Profiles. J. Computational Biology 7:559-584.

[3] Cho JH, Lee D, Park JH, Lee IB (2004) Gene selection and classification from microarray data using kernel machine. FEBS Letters 571:93-98.

[4] Golub T, Slonim D, Tamayo P, Huard C, Caasenbeek JM, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, Lander E (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531-537.

[5] Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Machine Learning 46:389-422.

[6] Hong Y, Kwong S, Chang Y, Ren Q (2008) Unsupervised feature selection using clustering ensembles and population based incremental learning algorithm. Pattern Recognition 41:2742-2756.

[7] Huang HL, Chang FL (2007) ESVM: evolutionary support vector machine for automatic feature selection and classification of microarray data. BioSystems 90:516-528.

[8] Lee CP, Leu Y (2011) A novel hybrid feature selection method for microarray data analysis. Applied Soft Computing 11:208-213.

[9] Lee SH, Lim JS (2011) Forecasting KOSPI based on a neural network with weighted fuzzy membership functions. Expert Systems with Applications 38:4259-4263.

[10] Lee SH, Lim JS (2012) Parkinson's disease classification using gait characteristics and wavelet-based feature extraction. Expert Systems with Application 39:7338-7344.

[11] Maji P, Paul S (2011) Rough set based maximum relevance-maximum significance criterion and gene selection from microarray data. International Journal of Approximate Reasoning 52:408-426.

[12] Wang S, Li D, Song X, Wei Y, Li H (2011) A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. Expert Systems with Applications 38:8696-8702.

[13] Wang Y, Makedon FS, Ford JC, Pearlman J (2005) HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics 21:1530-1537.