Gene Selection for SRBCTs Subtype Classification Using Fuzzy Neural Network

Xue W. Tian and Joon S. Lim
I.T. College, Gachon University, Seongnam, South Korea
Email: tianxuemao@gmail.com, jslim@gachon.ac.kr

***Abstract*—An approach for cancer molecular classification based on their gene expression profiles is proposed. Four subtypes of the small, round blue-cell tumors (SRBCTs) were classified in this research. The Bhattacharyya distance of each gene was used as the gene selection method to select the twelve preliminary good genes for the SRBCTs subtypes classification. We then developed a classification method based on a fuzzy neural network (FNN) and a three-level classification model. Using the twelve preliminary good genes we did 100,000 iterations for each experiment on each level by using the FNN. After the experiments we got the number of bad cases (BC) for each gene. By the number of BC we deleted the bad genes one by one by using the FNN. Finally, we selected four genes for the SRBCTs subtype classification with 100% classification accuracy.**

***Index Terms*— SRBCTs, Bhattacharyya distance, gene expression profiles, fuzzy neural networks, gene selection**

I. Introduction

With the successful completion of the Human Genome Project (HGP), we are entering the post genomic era. Facing mass amounts of data, traditional biological experiments and data analysis techniques encounter great challenges. In this situation, cDNA microarrays and high-density oligonucleotide chips are novel biotechnologies as global (genome-wide or system-wide) experimental approaches that are effectively used in systematical analysis of large-scale genome data. In recent years, with its ability to measure simultaneously the activities and interactions of thousands of genes, microarray promises new insights into the mechanisms of living systems and is attracting more and more interest for solving scientific problems and in industrial applications. Meanwhile, further biological and medical research also promoted the development and application of microarray.

Typical issues addressed by microarray experiments include two main aspects: finding co-regulated genes for classification based on different cell-type [1], stage-specific [2], [3], disease-related [4]-[6], or treatment-related [6]-[8] patterns of gene expression and understanding gene regulatory networks by analyzing functional roles of genes in cellular processes [9], [10]. Here we focus on the former, especially on tumor classification using gene expression data, which is a hot topic in recent years and has received general attention by many biological and medical researchers [11]-[18]. A reliable and precise classification of tumors based on gene expression data may lead to a more complete understanding of molecular variations among tumors, and hence, to better diagnosis and treatment strategies.

Microarray experiments usually generate large datasets with expression values for thousands of genes (2000~20 000) but not more than a few dozens samples (20~80). Thus, very accurate classification of tissue samples in such high-dimensional problems is difficult, but often crucial, for successful diagnosis and treatment. Several comprehensively comparative and improved methods have been proposed recently [18-20]. In this paper, we introduce a combinational feature selection method using neural fuzzy networks to remarkably decrease the number of differently expressed gene (DEG) for the sample classification. In recent years, several researchers have used ensemble neural networks for tumor classification based on gene expression data [12], [21]. Khan et al. [12] used neural networks to classify 4 subcategories of small round blue-cell tumors. By using 3750 networks generated by three fold cross-validation 1250 times and using the list of 96 most influential genes as the inputs, they reported very excellent results based on their dataset. Also O'Neill and Song [21] used neural networks to analyze lymphoma microarray data and can predict the long-term survival of individual patients with 100% accuracy based on the datasets published by Alizadeh et al [18]. Both of them are very good work in microarray data analysis using neural networks.

In this paper, we applied a fuzzy neural network (FNN) which is named NEWFM (neu-ro-fuzzy network with a weighted fuzzy membership function) [22] to the SRBCTs classification. The small, round blue cell tumors (SRBCTs) of childhood, which include neurob-lastoma (NB), rhabdomyosarcoma (RMS), burkitt lymphomas (BL), and the Ewing family of tumors (EWS), are so named because of their similar appearance on routine histology [23]. NEWFM is a kind of FNN which is modeled on the structure and behavior of neurons in the human brain and can be trained to recognize and categorize complex patterns [22]. The problem of feature selection is very important in pattern recognition. Feature selection in this paper is gene selection. We used Bhattacharyya distance [23] for the preliminary good gene selection method. The Bhattacharyya distance of

Manuscript received February 13, 2013; revised April 23, 2013.

doi: 10.12720/joig.1.3.134-137

each gene was used as the criterion for ranking genes in the training dataset. In statistics, the Bhattacharyya distance measures the similarity of two discrete or continuous probability distributions. It is a measurement of the amount of overlap between two statistical samples or populations. We calculated the Bhattacharyya distance between each subtype and other three subtypes, and then applied the Bhattacharyya distance to rank genes. Because the Bhattacharyya distance discriminate only two classes, we constructed a three-level classification model for the four SRBCTs subtypes classification. Then, we selected the twelve top ranked genes as the preliminary good genes for the further gene selection. In the further gene selection, we did 100,000 iterations for each experiment on each level by using the FNN. After the experiments we got the number of BC (the number of the bad effect in the 100,000 times classification) for each gene. We summed all the BC on the three levels and gave an order of the sum. The gene with bigger sum is less differently expressed gene (DEG). Therefore, we deleted the bad genes one by one by the order of the sum using the FNN. Finally, we selected four best genes for the SRBCTs subtype classification with 100% classification accuracy.

II. Materials and Method

A. Materials

Khan [12] filtered the 6567 cDNA gene expression profiles by requiring a minimal intensity of expression, which reduced the number of genes to 2308 [12]. In this research, we used the 2308 genes with 63 training samples (23 EWS, 8 BL, 12 NB, and 20 RMS samples) and 20 test samples (6 EWS, 3 BL, 6 NB, and 5 RMS). Each sample was a feature vector of 2308 natural log-normalized gene expression values.

B. Gene Selection Method and Multiclass Classification Model

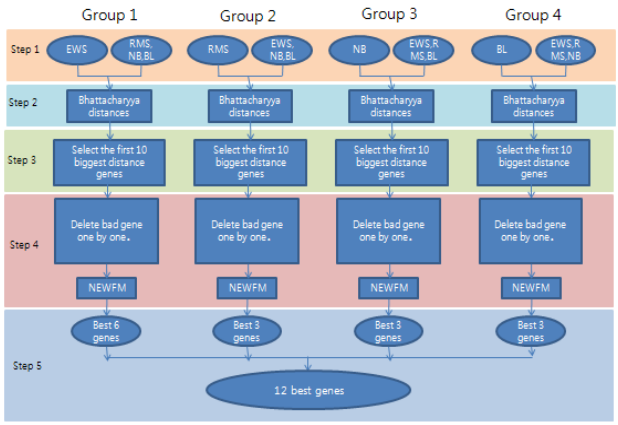

Figure 1. Structure of Preliminary Gene Selection.

Preliminary Gene Selection: The preliminary gene selection method includes six steps for four groups which were grouped before the gene selection process, as shown in Fig. 1. In step 1, we divided the samples into two classes (class one and class two) in each group. For example, in group 1, the four subtypes are divided into EWS and other three subtypes. There are the same in group 2, 3, and 4. In step 2, we calculated the Bhattacharyya distances between class one and class two in each group. And then we listed the genes by Bhattacharyya distances in descending order. The gene with bigger distance is more DEG. In step 3, we selected the first ten biggest distance genes from each group. In step 4, we deleted the bad gene one by one from bottom to top of the ten biggest distance genes in each group by NEWFM. After step 4, we selected 6, 3, 3, and 3 best genes from group 1, 2, 3, and 4, respectively. Because there have several duplicated genes in the four groups, finally we got twelve best genes for the SRBCTs subtypes classification

TABLE I. Ordering the 4 Groups for the Three-Level Classification

Image Id.	ranking in group 1	ranking in group 2	ranking in group 3	ranking in group 4
770394	1	51	3	96
866702	2	9	30	67
814260	3	16	54	297
377461	4	53	69	63
245330	5	3	28	56
295985	6	764	787	1289
784224	15	1	12	165
461425	10	2	2	44
812105	203	167	1	238
236282	168	222	180	1
839736	826	183	6	2
43733	9	14	25	3
sum of ranking	1252	1485	1197	2321
order of sum	2	3	1	4

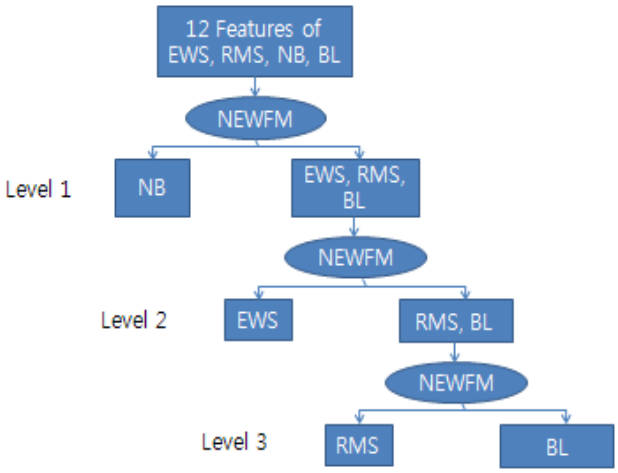

Figure 2. Structure of preliminary gene selection.

Multiclass Classification Model: Because the Bhattacharyya distance only can calculate the distance between two classes, we made a three-level classification model for the four SRBCTs subtypes classification. For making the three-level classification model, we gave the sum of the rankings of the twelve best genes in each group as shown in Table I. The group 3 has the smallest sum, that means these twelve best genes have significant strengths in classify NB and other three subtypes. So, in

the first level of our model is to classify NB from the four subtypes. The result of level one is classifying the four subtypes into class one (NB) and class two (EWS, RMS, and BL). In the same way, the second level is to classify EWS from the class two which is classed in level one, the third level is to classify RMS and BL. The three-level classification model is shown on Fig. 2.

Further Gene Selection: We did 100,000 iterations for each experiment on each level by using the FNN. After the experiments we got the number of BC for each gene, the number of BC are shown on Table II. Then we summed all the BC on the three levels and gave an order of the sum. The gene with bigger sum is less DEG. Therefore, we deleted the bad genes one by one by the order of the sum using the FNN. Finally, we selected four genes for the SRBCTs subtype classification with 100% classification accuracy.

TABLE II. ORDERING THE SUM OF BC FOR THE PRIMARY TWELVE GOOD GENES

Image Id.	BC on LEVEL1	BC on LEVEL2	BC on LEVEL3	SUM of BC	order of sum
236282	7844	9253	3258	20355	1
784224	7421	7408	6093	20922	2
770394	8678	3914	8536	21128	3
812105	26	14146	7687	21859	4
43733	9005	8929	5891	23825	5
377461	10839	6450	7492	24781	6
245330	8417	8918	8610	25945	7
839736	7552	10458	8296	26306	8
295985	8032	7086	11331	26449	9
461425	9949	9681	8974	28604	10
814260	10956	5339	12501	28796	11
866702	11281	8418	11331	31030	12

III. EXPERIMENTAL RESULTS AND CONCLUDING REMARKS

The performance results of this study can achieved at 100% recognition accuracy with only four best genes. In Khan's research, they classified the four SRBCTs subtypes with 96 best genes. The best genes in our study are greatly less than in Khan's research.

In the future research, we will apply this method on other gene expression profiles analysis. We will work on how to use the standard deviation or other mathematical method to get more efficiently microarray data analysis method.

ACKNOWLEDGMENT

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the IT-CRSP (IT Convergence Research Support Program) (NIPA-2013-H0401-13-1001) supervised by the NIPA (National IT Industry Promotion Agency).

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology. (2012R1A1A2044134).

REFERENCES

[1] W. David and Galbraith, "Global analysis of cell type-specific gene expression," *Comparative Functional Genomics*, vol. 4, pp. 208-215, 2003.

[2] P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, and K. Anders, *et al.*, "Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273-3297, 1998.

[3] T. L. Nicholson, L. Olinger, K. Chong, G. Schoolnik, and R. S. Stephens, "Global stage-specific gene regulation during the developmental cycle of chlamydia trachomatis," *Bacteriology*, vol. 185, pp. 3179-3189, 2003.

[4] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, and J. Manola, *et al.*, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, pp. 203-209, 2002.

[5] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, and V. A. Fusaro, *et al.*, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet*, vol. 359, pp. 572-577, 2002.

[6] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, and M. Angelo, *et al.*, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, pp. 436-442, 2002.

[7] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, and J. L. Kutok, *et al.*, "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, pp. 68-74, 2002.

[8] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *National Academy of Sciences of the USA*, vol. 95, pp. 14863-14868, 1998.

[9] V. Cherepinsky, J. Feng, M. Rejali, and B. Mishra, "Shrinkage-based similarity metric for cluster analysis of microarray data," *National Academy of Sciences of the USA*, vol. 100, pp. 9668-9673, 2003.

[10] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, and M. Gaasenbeek, *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[11] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, and M. Ladanyi, *et al.*, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, pp. 673-679, 2001.

[12] L. J. Veer, H. Dai, M. J. Vijver, Y. D. He, and A. A. Hart, *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, pp. 530-536, 2002.

[13] U. Alon, N. Barkai, D. A. Notterman, K. Gish, and S. Ybarra, *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by Oligonucleotide arrays," *National Academy of Sciences of the USA*, vol. 96, pp. 6745-6750, 1999.

[14] E. J. Yeoh, M. E. Ross, S. A. Shurtleff, W. K. Williams, and D. Patel, *et al.*, "Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling," *Cancer Cell*, vol. 1, pp. 133-143, 2002.

[15] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, and J. E. Blumenstock, *et al.*, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Research*, vol. 62, pp. 4963-4967, 2002.

[16] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, and I. S. Lossos, *et al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503-511, 2000.

[17] J. DeRisi, Penland L, P. O. Brown, M. L. Bittner, and P. S. Meltzer, *et al.*, "Use of a cDNA microarray to analyse gene expression patterns in human cancer," *Nature Genetics*, vol. 14, pp. 457-460, 1996.

[18] S. J. Dudoit, J. Fridlyand, and T. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, pp. 77-87, 2002.

[19] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic Patterns," *Genome Informatics*, vol. 13, pp. 51-60, 2002.

[20] J. Jaeger, R. Sengupta, and W. L. Ruzzo, "Improved gene selection for classification of microarrays," in *Pacific Symposium on Biocomputing*, 2003, pp. 53-64.
[21] M. C. O'Neill and L. Song, "Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect," *BMC Bioinformatics*, vol. 4, no. 13, 2003.
[22] J. S. Lim, "Finding features for Real-time premature ventricular contraction detecti-on using a fuzzy neural network system," *IEEE Transactions on Neural Networks*, pp. 522-527, 2009.
[23] P. A. Pizzo, *Principles and Practice of Pediatric Oncology*, Lippincott Williams & Wilkins, Philadelphia, 1997.

Xue W. Tian received the B.S. and M.S. degrees in computer science from Shandong University of Technology, China in 2008, Kyungwon University, Korea in 2010.

She is in doctor's course in computer science from Gachon University, Korea. Her research focuses on neuro-fuzzy systems, biomedical prediction systems, and signal process.

Joon S. Lim received the B.S. and M.S. degrees in computer science from Inha University, Korea, University of Alabama at Birmingham, and Ph.D. degree from Louisiana State University, Baton Rouge, Louisiana, in 1986, 1989, and 1994, respectively.

He is currently a Professor of Division of Software at Gachon University, Korea. His research focuses on neuro-fuzzy systems, biomedical prediction systems, and human-centered systems. He has authored three textbooks Artificial Intelligence Programming (Green Press, 2000), Javaquest (Green Press, 2003) and C# Quest (Green Press, 2006).