International Journal of Bio-Science and Bio-Technology
Vol.7, No.4 (2015), pp.213-222
http://dx.doi.org/10.14257/ijbsbt.2015.7.4.20

Lung Cancer Classification using Fuzzy Interactive Naïve Bayesian Network

Zhenxing Zhang[1], Letao Qu[2] and Joon S.Lim[3]*

[1] *School of Information and Electrical Engineering, Ludong University, Shandong, China*
[2] *IT College, Gachon University, Seongnam, South Korea*
[3]*IT College, Gachon University, Seongnam, South Korea*
billzhenxing@gmail.com,quletaog@gmail.com, jslim@gachon.ac.kr

Abstract

Lung cancer is the most serious disease in the world and millions of people die of it every year. Because of the limitations of current treatment processes, it is difficult to cure lung cancer if the patient is no longer in the early stages. Therefore, it is necessary to diagnose lung cancer as early as possible, thereby increasing the chances to cure it. The Fuzzy Interactive Naïve Bayesian (FINB) network is a new Bayes network that can be used to classify lung cancer by using microarray data sets. The FINB network is an interactive network and every attribution has an interactive parent and with a weight on the relationship that shows the interaction of the attribution in the data set. In our experiments, we use the gene expression profiles from the Affymetrix Human Genome U133 Plus 2.0 microarray. We use the Neural Network with a Weighted Fuzzy Membership Function (NEWFM) to train the data set and reconstruct the Fuzzy Interactive Naïve Bayesian network. Then we compare the results with Tree augment naïve Bayesian (TAN) network. We conclude that the FINB network performs better than the TAN network.

Keywords: *Lung cancer, gene microarray data, Naïve Bayesian, FINB, NEWFM, TAN.*

1. Introduction

Cancer has the highest morbidity and mortality rates among diseases, and lung cancer is one of most serious types of cancer. Every year, millions of people die of lung cancer. In recent years air pollution has become much worse and the number of smokers has increased. In conjunction with these factors, the number of lung cancer patients is also increasing. Therefore, diagnosing lung cancer in its early stages has become a very important concern in the medical field. Much effort has been put towards the analysis of microarray data [1, 2].

The Naïve Bayesian (NB) network is a universal way to classify lung cancer using gene expression data. As we can see in Fig. 1, in the NB network, the attributions are not related to any other attributions [3, 4, 5].

* Corresponding author: Joon S.Lim, IT College, Gachon University, Seongnam, South Korea, jslim@gachon.ac.kr

ISSN: 2233-7849 IJBSBT

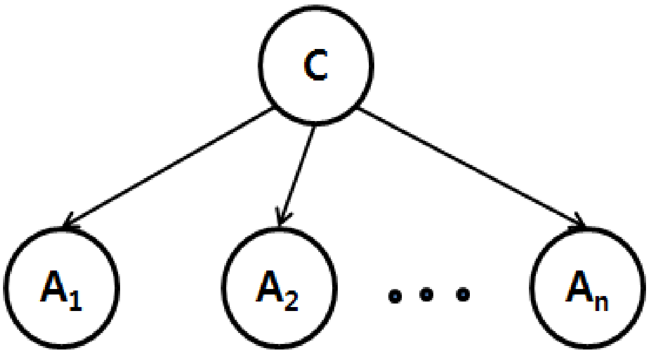

Figure 1. The Construction of NB Network

The Tree augmented Naïve Bayesian (TAN) network is an advanced naïve Bayesian network. In Fig. 2, we can see that each attribution in a TAN network depends on the class and the other attributions, but there is no weight on the relationship between attributions that indicates the interaction of the relationship. Some research shows that TAN's performance is often better than that of an NB network [6].

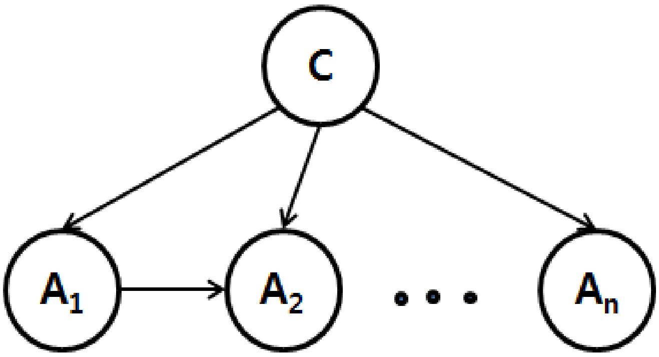

Figure 2. The Construction of TAN Network

In this paper, we propose to use another Bayesian network, the Fuzzy Interactive Naïve Bayesian (FINB) network [7], to perform the classification. The construction of an FINB network is shown in Fig. 3. As we can see from the figure, as opposed to NB network and the TAN network, the FINB network is an interactive Bayesian network in which every attribution has an interactive parent that has the biggest influence with it from all the attributions. We set a weight between each attribute and its interactive parent in order to represent their relationship. The weights are constructed by a neural network with weighted membership functions (NEWFM) [8, 9,10].

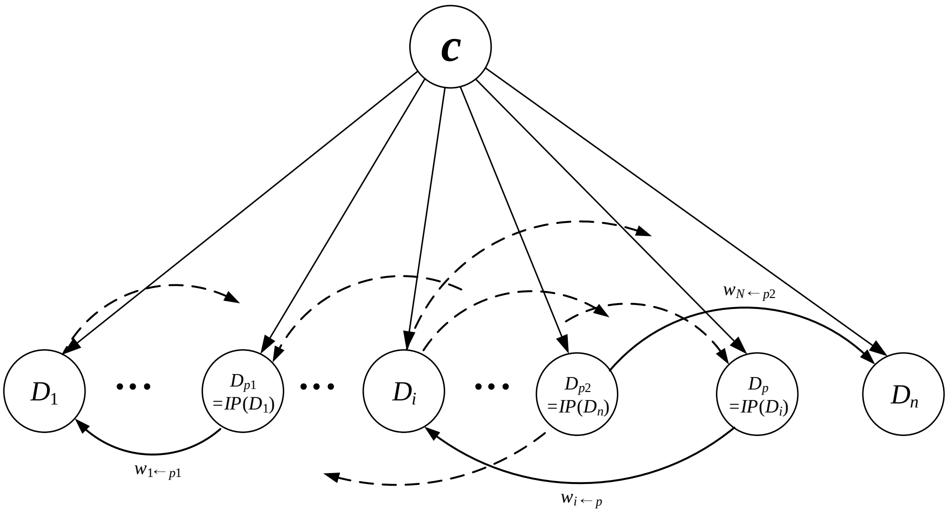

Figure 3. The Construction of FINB Network

2. Materials and Method

In this section, we talk about the materials used in the experiment and the method of the FINB network construction

2.1. Materials

Our experiment data consists of sixty pairs of tumor specimens and adjacent normal lung tissue specimens, analyzed by Affymetrix Human Genome U133 Plus 2.0 expression arrays [11]. We use all 120 samples for both training and testing.

2.2. Process of FINB Network Construction

The process of FINB network construction is shown in Fig. 4.

As we can see from the figure, the process of consists of four parts: preprocessing the data set (feature selection and normalization), finding interactive parents, weight calculation, and the final FINB network construction.

2.3. Preprocessing the Data Set

In this section, we discuss how we perform the feature selection and data normalization. Feature selection is very important because appropriate feature selection methods can improve the performance by reducing the influence of unwanted features [12]. In feature selection stage, we use the Bhattacharyya distance (BD) method to find the Bhattacharyya distance between all the genes. Then, we select the genes that have the greatest distance. For data normalization, we use the sigmoid method to normalize the selected genes.

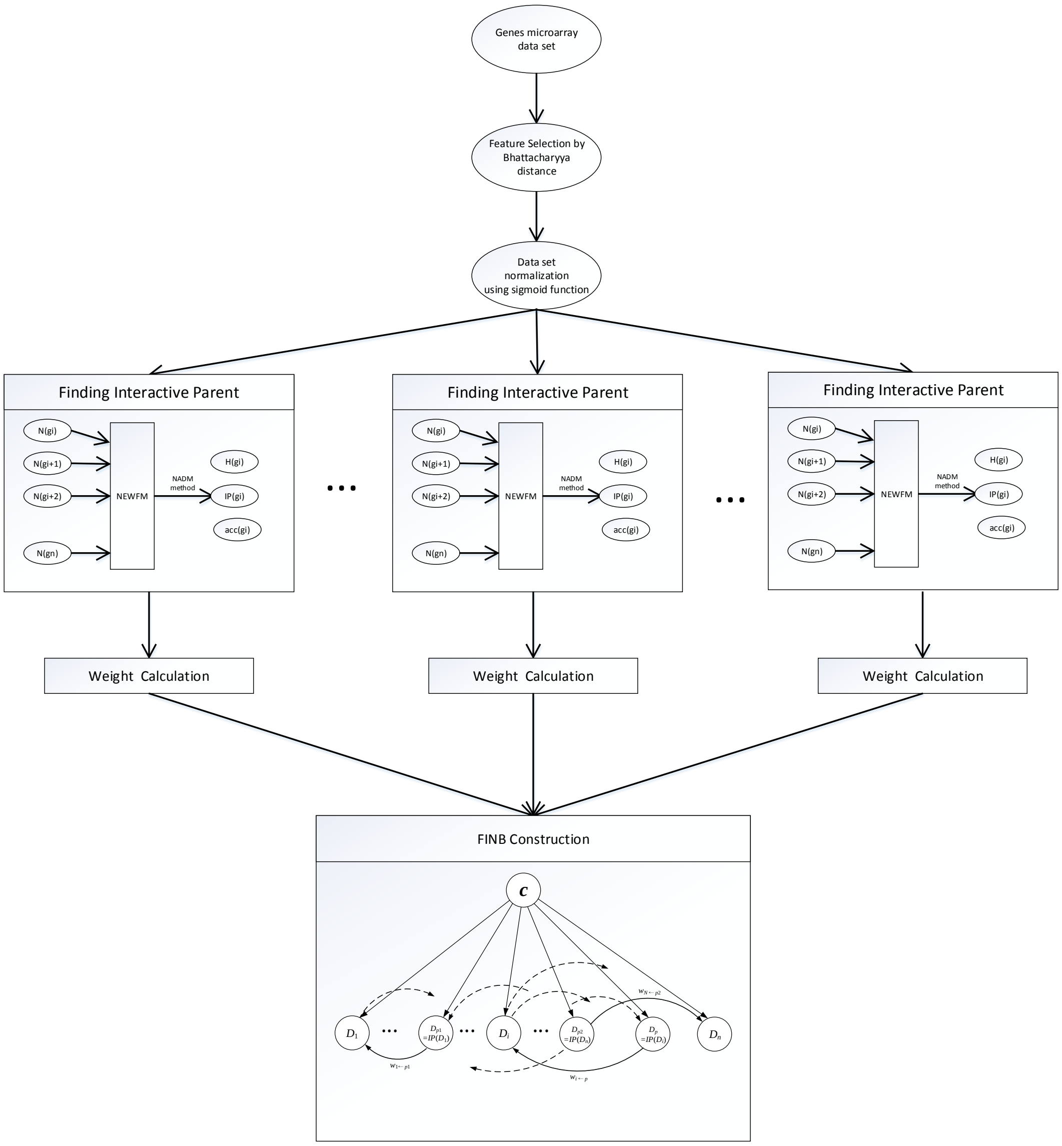

Figure 4. The Process of FINB Construction

2.3.1. Feature Selection

Feature selection is an integral part of the process because using an appropriate feature selection method can improve the performance by reducing the influence of unwanted features.

Bhattacharyya distance [13] is a method used to measure the distribution of two probabilities. The value of a Bhattacharyya distance shows how close two probability distributions are to one another. The greater the value is, the closer the two probability distributions are. It is an efficient way to perform feature selection [14, 15].

Therefore, in our work, we use the Bhattacharyya distance to perform feature selection. The Bhattacharyya distance equation is given by

$$D_B(D_i) = \frac{1}{4}\ln(\frac{1}{4}\left(\frac{\sigma_1(D_i)^2}{\sigma_2(D_i)^2} + \frac{\sigma_2(D_i)^2}{\sigma_1(D_i)^2} + 2\right)) + \frac{1}{4}\frac{(\mu_1(D_i) - \mu_2(D_i))^2}{\sigma_1(D_i)^2 + \sigma_2(D_i)^2} \qquad (1),$$

where: $D_B(D_i)$ is the Bhattacharyya distance of ith gene in the data set and D_i is the value of ith gene in the data set. Moreover, $\sigma_1(\mathrm{D}_i)$ is the standard deviation of class 1 of D_i, $\sigma_2(\mathrm{D}_i)$ is the standard deviation of class 2 of D_i. In addition $\mu_1(\mathrm{D}_i)$ is the mean value of class1 of D_i and $\mu_2(\mathrm{D}_i)$ is the mean value of class 2 ofD_i.

After calculating the Bhattacharyya distances of all of the genes, we select the five genes that have the greatest distances [14, 16].

2.3.2. Normalization

After the feature selection, we use the sigmoid method to normalize the selected gene data set [17]. Then, we use the sigmoid method to normalize the data set between 0 and 1[18]. The equation of the sigmoid method is given by

$$\mathrm{N}_i = \left\{ n_{i,j} \middle| \, n_{i,j} = \frac{1}{1+e^{-(d_{i,j}-\min(\mathrm{D}_i))/(\max(\mathrm{D}_i)-\min(\mathrm{D}_i))}}, \forall\, j = 1,2,\ldots,s \right\} \quad (2),$$

where D_i is the set of value of ith gene in the data set, $\min(\mathrm{D}_i)$ is the minimum value of D_i, $\max(\mathrm{D}_i)$ is the maximum value of D_i, $d_{i,j}$ is the value of jth sample in D_i, $n_{i,j}$ is the normalized value of $d_{i,j}$ and N_i is the normalized set of values of D_i.

2.4. Finding Interactive Parents and Calculating Weight

After preprocessing the data set, we find the interactive parent of every attribution by using the Non-overlap Area Distribution Method (NADM) [8, 9,10].

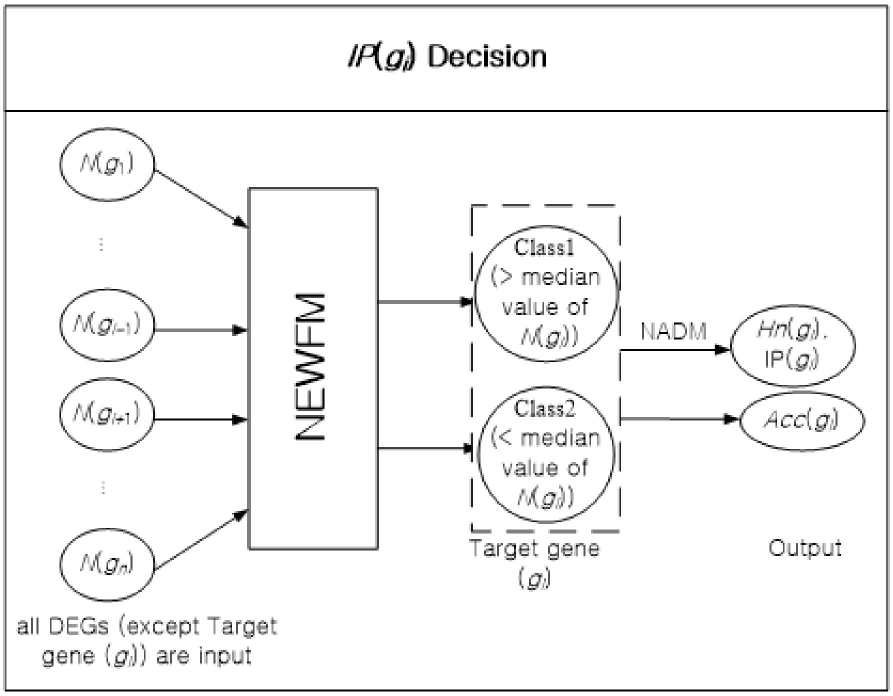
IP(gi) Decision
N(g1)
N(gi-1)
N(gi+1)
N(gn)
all DEGs (except Target gene (gi)) are input
NEWFM
Class1 (> median value of N(gi))
Class2 (< median value of N(gi))
Target gene (gi)
NADM
Hn(gi), IP(gi)
Acc(gi)
Output

Figure 5. The Process of Finding Interactive Parents

The process of finding interactive parents is shown in Fig. 5. In this process, we select gene as the target gene for which we need to find the interactive parent. We test all of the other genes as potential parent candidates by using NEWFM. We select the gene with the highest accuracy as the target gene's interactive parent [18, 19, 20]. We repeat this process for every gene, so that each one is assigned a parent.

After finding an interactive parent, we can calculate the weight of its relationship with the target gene. The weight equation is given by:

$$w = \mathrm{H}(\mathrm{g}_i) * \mathrm{acc}(\mathrm{g}_i)/\mathrm{t} \qquad (3),$$

where: g_i is the target gene, $\mathrm{H}(\mathrm{g}_i)$ is the greatest number of selected genes, $\mathrm{acc}(\mathrm{g}_i)$ is the highest accuracy, and t is the time taken for training using NEWFM.

2.5. Construction of an FINB Network

After finding the interactive parents and calculating the weights, we can use this information to construct an FINB network. The construction of an FINB network is shown in Fig. 6.

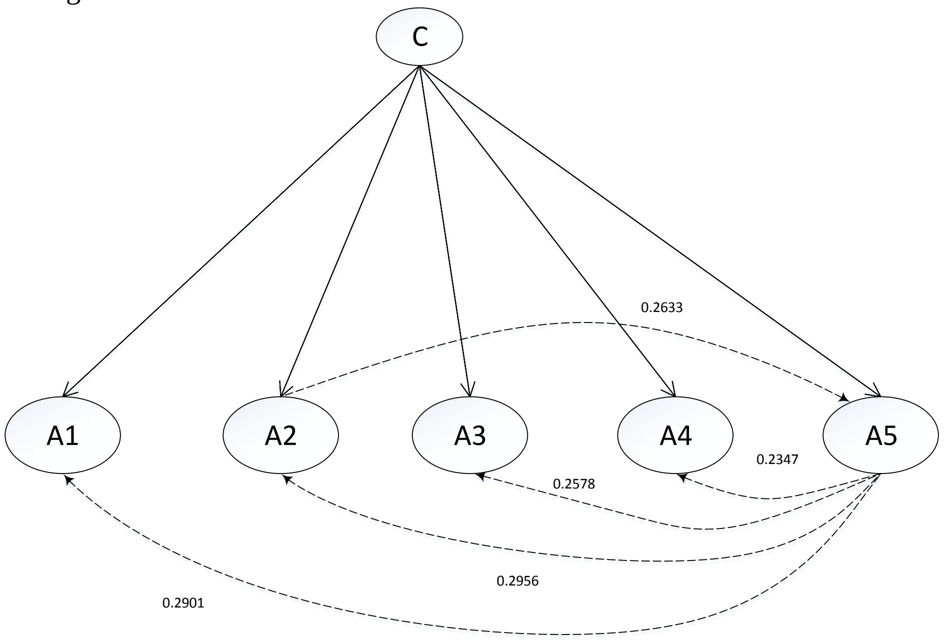

Figure 6. The Construction of an FINB Network

A1-A5 represent the genes we select from the experiment, the numbers are the weights of the relationship.

2.6. Classification Using an FINB Network

After constructing the FINB network, we can calculate the distribution of each attribution. Using the distribution and weight of each attribution, we can find the classification of the data set.

First, we can find the probability of each sample in every class C. The equation is given by

$$P(\mathrm{D}_1, \ldots, D_n, C) = P(C) \prod_{i=1}^{n} P(D_i | D_b, C) \qquad (4),$$

and

$$P(D_i | D_b, C) = P(D_\mathrm{i} | C) + \left(P(D_b | C) - P(D_\mathrm{i} | C)\right) * \mathrm{w}_i * P(D_i | D_1, \ldots, D_{i-1}, D_{i+1}, \ldots, D_n, C) \qquad (5)$$

Next, we can find the highest probability of the sample about every class $c(S)$, and then the classification of this sample is the class that $c(S)$ represents.

$$c(S) = \arg\max_{c \in C} P(c)\prod_{i=1}^{n} P(\mathrm{D}_i | D_b, c) \quad (6).$$

3. Results

We analyze the results of our experiment, to determine the accuracy of both the FINB network and the TAN network. Our results are shown in Table 1, the values of Acc (accuracy), TP (true positive), FP (false positive), TN (true negative) and FN (false negative) are listed in the table. We can see that the classification accuracy of FINB is 96.67% which is better than the accuracy of TAN and the value of true positive is also better than TAN.

Table 2. Experimental Results Comparing FINB and TAN

Method	*TP*	*FP*	*TN*	*FN*	*Acc*
FINB	59	3	57	1	96.67%
TAN	58	3	57	2	95.83%

From this table, we can see in this experiment the classification performance of FINB is better than the performance of TAN.

4. Conclusion

In this experiment, we used a gene expression data set with 120 samples, of which 60 had lung cancer and 60 did not. We constructed an FINB network and use a TAN network and use both to classify the data set. The results show the performance of FINB is better than TAN.

In the Naïve Bayesian network, attributions only have a relationship with the class node and do not have any relationships between each other. In the Tree augmented Naïve Bayesian network, attributions depend not only on the class node, but also on another attribution. However, there are no weights on the relationships to represent the interaction of the connections.

In the Fuzzy Interactive Naïve Bayesian network, every attribution has a parent class node and also has an interactive parent, to which it has a stronger connection than with any other attribution. In addition, there is a weight assigned to the relationship to describe the interaction.

It is clear that FINB is more advantageous than TAN in terms of classification. When attributions in the data set have close connections, it especially good to use the FINB network to perform the classification

Acknowledgements

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).

The Research Foundation for Doctoral Program of Ludong University(Grant No.210-28350301)

References

[1] M. Q. Zhang, “Large-scale gene expression data analysis: a new challenge to computational biologists”, Genome Research, vol.9, **(1999)**, pp.681-688.

[2] P. Baldi and G. W. Hatfield, DNA microarrays and gene expression [M]. Cambridge University Press, **(2002)**.

[3] M. Schena, D. Shalon and R. W. Davis, “Quantitative monitoring of gene expression patterns with a complementary microarray”, Science in China Series C-Life Sciences, vol.270, **(1995)**, pp.467-470.

[4] M. Grunsein and D. Hogness, “Colony hybridization: a method for the isolation of cloned DNA that contain a specific gene”, PNAS, vol.72, **(1975)**, pp.3961-3965.

[5] S. P. Fodor, R. P. Rava and X. C. Huang, “Multiplexed biochemical assays with biological chips”, Nature, vol.364, **(1993)**, pp.555-556.

[6] N. Friedman, D. Geiger and M. Goldszmidt, “Bayesian network classifiers”, Machine Learning, vol.29, no.9, **(1997)**, pp.131-163.

[7] X. W. Tian, Constructing Fuzzy Interactive Naïve Bayesian Network for DNA microarray Data Classification, PhD dissertation, **(2014)**.

[8] J. S. Lim, D. Wang, Y.-S. Kim and S. Gupta, “A neuro-Fuzzy Approach for Diagnosis of Antibody Deficiency Syndrome”, Neurocomputing, vol.69, no.9, **(2006)**, pp.969-974.

[9] J. S. Lim, “Finding Features for Real-Time Premature Ventricular Contraction Detecti-on Using a Fuzzy Neural Network System”, IEEE Transactions on Neural Networks, **(2009)**, pp.522-527.

[10] Z. X. Zhang, S. H. Lee and J. S. Lim, " Detecting ventricular arrhythmias by NEWFM", Granular Computing, IEEE International Conference on, **(2008)**. [11] T. P. Lu, M. H. Tsai, J. M. Lee and C. P. Hsu, “Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women”, Cancer Epidemiol Biomarkers Prev **(2010)**.

[12] J. M. Sotoca and F. Pla, “Supervised feature selection by clustering using conditional mutual information-baseddistances”, Pattern Recognition, vol.43, **(2010)**, pp.2068-2081.

[13] G. Xuan, “Bhattacharyya distance feature selection”, Pattern Recognition, Proceedings of the 13th International Conference, **(1996)**. [14] T. R. Golub, D. K. Slonim and P. Tamayo, “Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring”, Science, **(1999)**, pp.531-537.

[15] K. Y. Yeung, R. E. Bumgarner and A. E. Raftery, “Bayesian Model Averaging: Development of an Improved Multi-class, Gene Selection and Classification Tool for Microarray Data”, Bioinformatics, vol.21, **(2005)**, pp.2394-2402.

[16] G. B. Coleman and H. C. Andrews, "Image Segmentation by Clustering", Proc IEEE, **(1979)**.

[17] H. Jun and M. Claudio, “The influence of the sigmoid function parameters on the speed of backpropagation learning”, From Natural to Artificial Neural Computation, **(1995)**, pp.195–201. [18] Y. Wang, F. S. Makedon, J. C. Ford and J. Pearlman, “HykGene: A Hybrid Approach for Selecting Marker Genes for Phenotype Classification Using Microarray Gene Expression Data”, Bioinformatics, vol.21, **(2005)**, pp.1530-1537.

[19] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, “Gene Selection for Cancer Classification Using Support Vector Machines”, Machine Learning, **(2002)**, pp.389-422.

[20] J. H. Cho, D. Lee, J. H. Park and I. B. Lee, “Gene Selection and Classification from Microarray Data Using Kernel Machine”, FEBS Letters, vol.571, **(2004)**, pp.93-98.

Authors

Zhen-Xing Zhang, he received the B.S. and M.S. degrees in computer science from Shandong University of Technology, China in 2005, Kyungwon University, Korea in 2008, and Ph.D. degree from Gachon University, Korea in 2012. He is currently a teacher in the department of computer science and technology at Ludong University, China. His research focuses on neuro-fuzzy systems, biomedical prediction systems, and signal process.

Letao Qu, he received the B.S. degrees in computer science from Ludong University, China in 2013. He is in master’s course in computer science from department of computer software at Gachon University, Korea. His research focuses on neuro-fuzzy systems, biomedical prediction systems, and signal process.

Joon S. Lim, he received his B.S. and M.S. degrees in computer science from Inha University, Korea, The University of Alabama at Birmingham, and Ph.D. degree was from Louisiana State University, Baton Rouge, Louisiana, in 1986, 1989, and 1994, respectively. He is currently a professor in the department of computer software at Gachon University, Korea. His research focuses on neuro-fuzzy systems, biomedical prediction systems, and human-centered systems. He has authored three textbooks on Artificial Intelligence Programming (Green Press, 2000), Javaquest (Green Press, 2003), and C# Quest (Green Press, 2006).