Paired Neural Network with Negatively Correlated Features for Cancer Classification in DNA Gene Expression Profiles

Hong-Hee Won
Yonsei University
Department of Computer Science
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
cool@candy.yonsei.ac.kr

Sung-Bae Cho
Yonsei University
Department of Computer Science
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
sbcho@cs.yonsei.ac.kr

Abstract-**While several conventional techniques for diagnosis of cancer in clinical practice can be often incomplete or misleading, molecular level diagnostics with gene expression profiles can offer the methodology of precise, objective, and systematic cancer classification. Moreover, since accurate classification of cancer is very important issue for treatment of cancer, it is desirable to make a decision by combining the results of various basis classifiers rather than by deciding the result with only one classifier. Generally combining classifiers gives high performance and high confidence. In spite of many advantages of ensemble classifiers, ensemble with mutually error-correlated classifiers has a limit in the performance. In this paper, we propose the ensemble of neural network classifiers learned from negatively correlated features to precisely classify cancer, and systematically evaluate the performances of the proposed method using three benchmark datasets. Experimental results show that the ensemble classifier with negatively correlated features produces the best recognition rate on the three benchmark datasets.**

I. Introduction

DNA microarray technology has advanced so much that we can simultaneously measure the expression levels of thousands of genes under particular experimental environments and conditions [1]. DNA microarray technology makes it possible to understand life on the molecular level. The development of DNA microarray technology enables to generate large-scale gene expression data. It has led to many statistical and analytical challenges from the problems in biology because it has been produced large amount of genes. We can analyze the gene information very rapidly and precisely by managing them at one time [2] using several statistical methods and machine learning.

Cancer classification in clinical practice relied on clinical and histopathological information can be often incomplete or misleading. DNA microarray technology has been applied to the field of accurate prediction and diagnosis of cancer and expected that it would help them. Molecular level diagnostics with gene expression profiles can offer the methodology of precise, objective, and systematic cancer classification. Especially accurate classification of cancer is very important issue for treatment of cancer. Since the gene expression data usually consist of huge number of genes, several researchers have been studying many problems of cancer classification using data mining methods, machine learning algorithms and statistical methods to efficiently analyze these data [3, 4]. However, most researchers have partly evaluated only the performance of the feature selection methods and classifiers. There was extensive work that not only evaluated the individual classifiers, but combined the individual classifiers based on a correlation of the features to improve the performance [5].

Many researchers have worked on the ensemble of the multiple classifiers to improve the performance of classification. The ensemble classifier increases not only the performance of the classification, but also the confidence of the results. Theoretically, the performance of the ensemble classifier grows larger when the combined classifiers are mutually independent. Representative ensemble methods such as average combination, voting, weighted voting, Bayesian approach and neural network have been applied to many fields of data mining and machine learning. However, these methods do not assure that the combined classifiers would be independent. On the other hand, the methods such as boosting (bootstrap resampling), bagging (bootstrap aggregating) and arcing (adaptively resampling and combining) produce diverse sample data, train heterogeneous classifiers with the data, and combine the classifiers [4, 6, 7, 8].

In this paper, we propose the method varying the data to create an effective ensemble [9]. For diagnosis of cancer, ensemble approach has been adapted to reduce false alarm rate with high detection rate [10]. While the breast cancer data consist of a number of samples and few features, our experimental data have a lot of features and few samples. Though creating an effective ensemble by varying the data is identical goal, the difference is that we vary features of data but they varied samples of data in the paper [10]. We use the negative correlation of the features for that purpose. We define two ideal feature vectors for a standard of good feature, and utilize the features selected by scoring the similarity with each ideal feature vector. Two ideal feature vectors are the one high in class A and low in class B, and the other one low in class A and high in class B. Since the vectors have negative correlation, the sets of genes similar to each ideal vector are also negatively correlated. The negatively correlated features represent two different aspects of classification boundary for gene expression data. We can search in a much wider solution space by combining classifiers learned from these features. In this paper, we propose the paired neural network classifier trained with negatively correlated features. We test the proposed method in three benchmark cancer datasets, and systematically analyze the usefulness of the negative correlation.

0-7803-7898-9/03/$17.00 ©2003 IEEE

II. BACKGROUNDS

A. cDNA Microarray

DNA arrays consist of a large number of DNA molecules spotted in a systemic order on a solid substrate. Depending on the size of each DNA spot on the array, DNA arrays can be categorized as microarrays when the diameter of DNA spot is less than 250 microns, and macroarrays when the diameter is bigger than 300 microns. The arrays with the small solid substrate are also referred to as DNA chips. It is so powerful that we can investigate the gene information in short time, because at least hundreds of genes can be put on the DNA microarray to be analyzed.

DNA microarrays are composed of thousands of individual DNA sequences printed in a high density array on a glass microscope slide using a robotic arrayer. The relative abundance of these spotted DNA sequences in two DNA or RNA samples may be assessed by monitoring the differential hybridization of the two samples to the sequences on the array. For mRNA samples, the two samples are reverse-transcribed into cDNA, labeled using different fluorescent dyes mixed (red-fluorescent dye Cy5 and green-fluorescent dye Cy3). After the hybridization of these samples with the arrayed DNA probes, the slides are imaged using scanner that makes fluorescence measurements for each dye. The log ratio between the two intensities of each dye is used as the gene expression data [1].

$$gene_expression = \log_2 \frac{\mathrm{Int}(Cy5)}{\mathrm{Int}(Cy3)} \quad (1)$$

where Int(Cy5) and Int(Cy3) are the intensities of red and green colors. Since at least hundreds of genes are put on the DNA microarray, we can investigate the genome-wide information in short time.

B. Oligonucleotide Microarray

Affymetrix (Inc, Santa Clara, CA) has developed GeneChip® oligonucleotide array. High-density oligonucleotide DNA probe array technology employs photolithography and solid-phase DNA synthesis.

High-density oligonucleotide chip arrays are made using spatially patterned, light-directed combinatorial chemical synthesis, and contain up to hundreds of thousands of different oligonucleotides on a small glass surface. Synthetic linkers, modified with a photochemically removable protecting groups, are attached to a glass surface, and light is directed through a photolithographic mask to a specific areas on the surface to produce localized deprotection. Specific hydroxyl-protected deoxynucleotides are incubated with the surface, and chemical coupling occurs at those sites that have been illuminated in the preceding step. As the chemical cycle is repeated, each spot on the array contains a short synthetic oligonucleotide, typically 20-25 bases long. The oligonucleotides are designed based on the knowledge of the DNA target sequences, to ensure high-affinity and specificity of each oligonucleotide to a particular gene. This allows cross-hybridization with the other similar sequenced gene and local background to be estimated and subtracted [1].

III. PAIRED NEURAL NETWORK WITH NEGATIVELY CORRELATED FEATURES

The framework of the proposed paired neural network with negatively correlated features is shown in Fig. 1. The basic idea of paired neural network scheme is to develop several pairs of trained neural networks with two ideal feature vectors, and to classify a given input pattern by utilizing combination methods. Then it naturally raises the question of obtaining a consensus on the results of each individual network.

The features are selected from DNA microarray data using two different ideal feature vectors, Ideal feature A and Ideal feature B. The neural network classifiers are trained with the selected features. MLP I is defined as the result of MLP trained by the feature set selected based on Ideal Gene A (1,1,...,1,0,0,...,0) and MLP II is defined as the result of MLP trained by the feature set selected based on Ideal Gene B (0,0,...,0,1,1,...,1). MLP I is MLP_{A1}~MLP_{Ak} and MLP II is MLP_{B1}~MLP_{Bk} in Fig. 1. Since the negatively correlated features represent two different aspects of data, the classifiers learned with two negatively correlated feature sets respectively are mutually complementary. Combining the heterogeneous classifiers helps increasing the performance of the classification.

A. Negative Correlation

Theoretically, the more features we may concern, the more

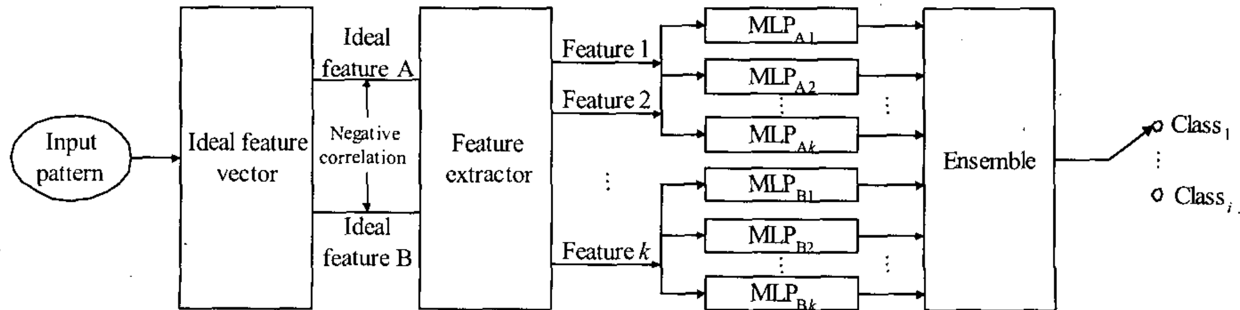

Fig. 1. Overview of the paired neural network with negatively correlated features

effective the classifier is to solve the problems [5]. But features that have overlapped in feature space may cause the redundancy of irrelevant information and result in the counter effect such as overfitting. When there are N feature selection methods, the set of non-linear transformation functions that change observation space into feature space is $\phi = \{\varphi_1, \varphi_2, \varphi_3, \ldots, \varphi_N\}$, and $\phi_k \in 2^{\phi}$, $I(\phi_k)$, the amount of classification information provided by the set of features ϕ_k, is defined as follows:

$$I(\phi_k) = \frac{a\sum A_i}{\frac{N}{2}\sum_{j=1, j\neq i}^{N} d_{ij}} + b \tag{2}$$

where d_{ij} is the dependency of the ith and the jth elements, A_i is the extent of the area which is occupied by the ith element from the feature space, and a and b are constant.

The higher dependency of a pair of features is, the smaller amount of classification information $I(\phi_k)$ is. As the extent of the area occupied by features is larger, the amount of classification information $I(\phi_k)$ is bigger. If we keep the number of features larger, the numerator of the equation is larger because the extent of the area occupied by features becomes wider. Although the numerator of the equation becomes larger, $I(\phi_k)$ will be mainly decreased without keeping d_{ij} small. Therefore, it is more desirable to use small number of mutually independent features than to unconditionally increase the number of features to enlarge $I(\phi_k)$, the amount of classification information provided to the classifier by the set of features. Correlation between feature sets can be induced from the distribution of feature numbers, or using mathematical analysis using statistics.

Therefore, it is more important to explore and utilize the informative features to train classifiers, rather than increase the number of features we use. The informative features are some genes highly related with particular classes for classification, which are called informative genes [11]. We have utilized the informative genes for the classification.

When there are $M \times N$ gene expression data having M samples and N genes and M samples are divided into two kinds of class, A and B, gene data g_i is defined as a vector as follows:

$$g_i = (e_1, e_2, e_3, \cdots, e_M),\ i=1 \sim N \tag{3}$$

We want to know the locations of informative k features out of N. If it is possible to know representative vector g_{ideal} for class c_j, we can simply measure the correlation and similarity of g_i to classes, which tells the feature-goodness. Modeling g_{ideal}, we should use prior knowledge and intuitional experience about classes.

Suppose g_{ideal} is an ideal vector representing class c_j.

$$g_{ideal} = (e_1', e_2', e_3', \cdots, e_M') \tag{4}$$

In this paper, we attempt to define two ideal feature vectors as the one high in class A and low in class B (1,1,...,1,0,0,...,0), and the other one low in class A and high in class B (0,0,...,0,1,1,...,1) as shown in Fig. 2 and select the sets of informative genes with high similarity to each ideal gene vector. Since Pearson's correlation coefficient of two ideal gene vectors is -1, two vectors are perfectly negatively correlated. The sets of gene vectors are also highly negatively correlated. The informative features selected by negative correlation represent two different aspects of training data. We can search in a much wider solution space by combining these features.

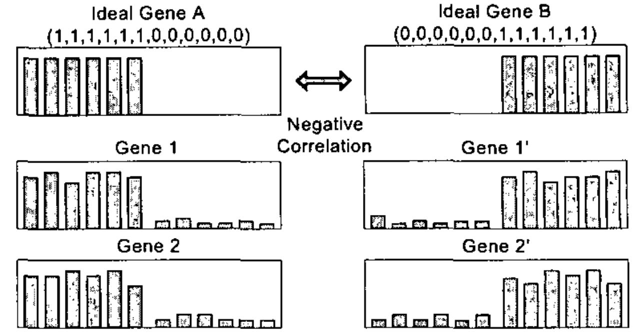

Fig. 2. Informative genes selected using negative correlation

B. Feature Selection Methods

Among thousands of genes whose expression levels are measured, not all are needed for classification. Microarray data consist of large number of genes in small samples. We need to select the informative genes for classification. This process is referred to as gene selection [11]. It is also called feature selection in machine learning. Informative genes highly correlated with class are selected generally using statistical correlation analysis, clustering method, etc. There have been various studies that extract the informative features for the classification using principal component analysis or genetic algorithm. The principal component analysis helps to extract the informative genes by transforming the feature space.

In this paper, informative genes are selected based on the similarity of gene vector g_i and ideal gene vectors A and B. The most similar 25 genes are used for classification. Correlation analysis and distance measure methods are used in order to measure the similarity of gene vector g_i and ideal gene vectors A and B.

Using the statistical correlation analysis, we can see the linear relationship and the direction of relation between two variables. Correlation coefficient r varies from -1 to $+1$, so that the data distributed near the line biased to (+) direction will have positive coefficients, and the data near the line biased to (–) direction will have negative coefficients. There are representative methods such as Pearson's correlation coefficient (*PR*) and Spearman's correlation coefficient (*SP*).

The similarity between two input vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be thought of as distance, which measures on how far the two vectors are located. The distance between g_{ideal_tumor} and g_i tells us how much likely the g_i is to the tumor class. Calculating the distance between them, if it is bigger than certain threshold, the gene g_i would belong to tumor class; otherwise g_i belong to normal class. We have adopted Euclidean distance (*ED*) and

cosine coefficient (*CC*).

Feature selection methods are summarized in Table I: Pearson correlation coefficient (*PR*), Spearman correlation coefficient (SP), Euclidean distance (*ED*), and cosine coefficient (*CC*) are listed in order.

TABLE I
MATHEMATICAL FORMULA FOR EACH FEATURE SELECTION METHOD

$$PR(g_i, g_{ideal}) = \frac{\sum g_i g_{ideal} - \frac{\sum g_i \sum g_{ideal}}{N}}{\sqrt{\left(\sum g_i^2 - \frac{(\sum g_i)^2}{N}\right)\left(\sum g_{ideal}^2 - \frac{(\sum g_{ideal})^2}{N}\right)}} \quad (5)$$

$$SP(g_i, g_{ideal}) = 1 - \frac{6\sum (D_g - D_{ideal})^2}{N(N^2-1)} \quad (6)$$

(D_g and D_{ideal} are the rank matrices of g_i and g_{ideal})

$$ED(g_i, g_{ideal}) = \sqrt{\sum (g_i - g_{ideal})^2} \quad (7)$$

$$CC(g_i, g_{ideal}) = \frac{\sum g_i g_{ideal}}{\sqrt{\sum g_i^2 \sum g_{ideal}^2}} \quad (8)$$

C. Paired Neural Network

A feed-forward multilayer perceptron (MLP) is error backpropagation neural network that is applied in many fields due to its powerful and stable learning algorithm. A neural network can be considered as a mapping device between an input set and an output set. Mathematically, a neural network represents a function F that maps I into O; $F : I \rightarrow O$, or $y=f(x)$ where $y \in O$ and $x \in I$. Since the classification problem is a mapping from the feature space to some set of output classes, we can formalize the neural network as a classifier.

Suppose a two-layer neural network classifier with T neurons in the hidden layer, and c neurons in the output layer. Here T is the number of features, c is the number of classes, and H is an appropriately selected number. The network is fully connected between adjacent layers. The operation of this network can be thought of as a nonlinear decision-making process: Given an unknown input $X = (x_1, x_2, \ldots, x_T)$ and the class set $\Omega = \{w_1, w_2, \ldots, w_c\}$, each output neuron produces y_i of belonging to this class by

$$P(w_i \mid X) \approx y_i = f\left\{\sum_{k=1}^{H} w_{ik}^{om} f\left(\sum_{j=1}^{T} w_{kj}^{mi} x_j\right)\right\} \quad (9)$$

where w_{kj}^{mi} is a weight between the jth input neuron and the kth hidden neuron, w_{ik}^{om} is a weight from the kth hidden neuron to the ith class output, and f is a sigmoid function such as $f(x) = 1/(1+e^{-x})$. The neuron having the maximum value is selected as the corresponding class.

The weight-update rule in backpropagation algorithm is defined as follows:

$$\Delta w_{ji}(n) = \eta \delta_j x_{ji} + \alpha \Delta w_{ji}(n-1) \quad (10)$$

where $\Delta w_{ji}(n)$ is the weight update performed during the nth iteration through the main loop of the algorithm, η is a positive constant called the learning rate, δ_j is the error term associated with j, x_{ji} is the input from neuron i to neuron j, and $0 \leq \alpha < 1$ is a constant called the *momentum*.

The network presented above trains on a set of example patterns and discovers relationships that distinguish the patterns. A network of a finite size, however, does not often load a particular mapping completely or it generalizes poorly. Increasing the size and number of hidden layers most often does not lead to any improvements.

We have chosen the Bayesian approach among several alternatives such as average combination, voting, weighted voting, Bayesian approach and neural network for final decision of the ensemble classifier. The Bayesian approach can solve the problem of tie-break in ensemble by using a priori knowledge of each combined classifier. While majority voting combines the classifiers with their results, Bayesian combination makes the error possibility of each classifier affect the final result. The method combines classifiers with different weights by using the previous knowledge of each classifier. Where k classifiers are combined, c_i, $i=1,\ldots,m$, is the class of a sample, $c(classifier_j)$ is the class of the jth classifier, and w_i is a priori possibility of the class c_i, Bayesian combination is defined as follows:

$$c_{ensemble} = \arg\max_{1 \leq i \leq m} \left\{ w_i \prod_{j=1}^{k} P(c_i \mid c(classifier_j)) \right\} \quad (11)$$

IV. EXPERIMENTS

A. Experimental Environment

Three representative datasets, leukemia cancer dataset, colon cancer dataset and lymphoma cancer dataset, are used in this paper among several microarray datasets from published cancer gene expression studies. Leukemia dataset consists of 72 samples: 25 samples of acute myeloid leukemia (AML) and 47 samples of acute lymphoblastic leukemia (ALL). 38 out of 72 samples were used as training data and the remaining were used as test data. Each sample contains 7129 gene expression levels. Colon dataset consists of 62 samples of colon epithelial cells taken from colon-cancer patients. Each sample contains 2000 gene expression levels. 31 out of 62 samples were used as training data and the remaining were used as test data. Lymphoma dataset consists of 24 samples of GC B-like and 23 samples of activated B-like. 22 out of 47 samples were used as training data and the remaining were used as test data.

For feature selection, each gene is scored based on the feature selection methods described in Table I, and 25 top-ranked genes are chosen as the feature of the input pattern. There is no report on the optimal number of genes, but our previous study indicates that 25 is reasonable [5].

For classification, we have used 3-layered MLP with 5~15

hidden nodes, 2 output nodes, learning rate of 0.01~0.50 and momentum of 0.9 and finally chosen 8 hidden nodes and learning rate of 0.10 with high performance. The maximum learning iteration is fixed to 100 in order to prevent overfitting.

The negative correlation feature set consists of MLP I and MLP II. For comparative study, we have compared the result of the ensemble of the negatively correlated feature set to the result of the ensemble of MLP I and the result of the ensemble of MLP II. We have combined the classifiers learned with the negatively correlated feature set using Bayesian approach, analyzed the results and evaluated them. In case of the negatively correlated feature set, we have produced 8 feature-classifier combinations using 4 feature selection methods and 2 classifiers (MLP I and MLP II). We have tried to combine 2, 3 and 4 classifiers among 8 classifiers, and analyzed the results of ensemble. We have conducted all ${}_8C_k$ (k=2, 3 and 4) combinations of ensemble, and have investigated the best recognition rate and average recognition rate.

B. *Result Analysis*

Table II shows the recognition rate of the basis classifiers in each dataset. Column is the list of feature selection methods: Pearson's correlation coefficient (PR), Spearman's correlation coefficient (SP), Euclidean distance (ED) and cosine coefficient (CC). In Leukemia dataset, MLP I with Pearson's correlation coefficient and MLP I with information gain produce the best recognition rate, 97.1%, among the feature-classifier combinations. In Colon dataset, MLP I with cosine coefficient produces the best recognition rate, 83.9%. In Lymphoma dataset, MLP II with Spearman's correlation coefficient produces the best recognition rate, 88.0%.

The result of MLP I is different from that of MLP II on each dataset. While MLP I outperforms MLP II in Leukemia dataset and Colon dataset, MLP II outperforms MLP I in Lymphoma dataset. These results are caused by the characteristics of the datasets. We expect that it is more informative to use the set of genes selected using Ideal Gene A, so that MLP I can outperform MLP II. Table II shows that MLP I with Pearson's correlation coefficient and MLP I is superior to MLP II with Pearson's correlation coefficient.

TABLE II
RECOGNITION RATE WITH FEATURES AND CLASSIFIERS (%)

	Leukemia		Colon		Lymphoma	
	MLP I	MLP II	MLP I	MLP II	MLP I	MLP II
PR	97.1	79.4	74.2	77.4	64.0	72.0
SP	82.4	79.4	58.1	64.5	60.0	88.0
ED	91.2	61.8	67.8	77.4	56.0	72.0
CC	94.1	76.5	83.9	77.4	68.0	76.0
Mean	91.2	74.3	71.0	74.2	62.0	77.0

Fig. 3, 4 and 5 show the average recognition rates of the ensemble classifiers for benchmark datasets. The x-axis means the number of combined classifiers, and the y-axis means the recognition rate. In case of the negatively correlated feature set (MLP I + MLP II), 8 diverse feature sets have been produced with 2 ideal feature vectors (Ideal Gene A and Ideal Gene B) and 4 feature selection methods (Pearson's correlation coefficient, Spearman's correlation coefficient, Euclidean distance and cosine coefficient). The classifiers learned with 8 diverse feature sets have been combined using Bayesian approach. The average recognition rate means the average of all possible ${}_8C_k$ (k=2, 3 and 4) combinations of ensemble classifiers. As the number of combined classifiers is increasing, average recognition rate of the ensemble classifier is also increasing. The performance of the ensemble classifier is superior to the basis in all benchmark datasets. The best recognition rate of ensemble classifier is 97.1% in Leukemia dataset, 87.1% in Colon dataset, and 92.0% in Lymphoma dataset as shown in Table III. Compared with the best recognition rates of base classifiers, 97.1%, 83.9%, and 92.0% on the datasets respectively in Table II, the performance of ensemble is better.

Compared with the results of MLP I and MLP II, the negatively correlated features set (MLP I + MLP II) does not outperform in the average recognition rate, but outperforms in the best recognition rate. While the best recognition of the ensemble of MLP I and MLP II is decreasing as the number of combined classifiers is increasing, the best recognition of the ensemble of the negatively correlated coefficient feature set is increasing.

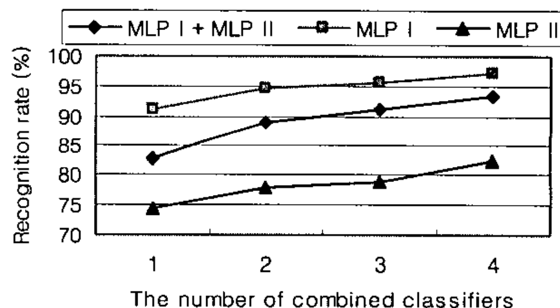

Fig. 3. Recognition rate of the ensemble in Leukemia

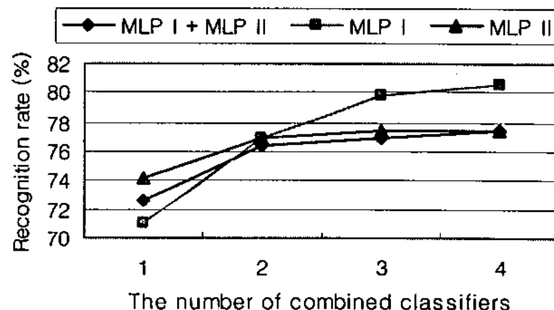

Fig. 4. Recognition rate of the ensemble in Colon

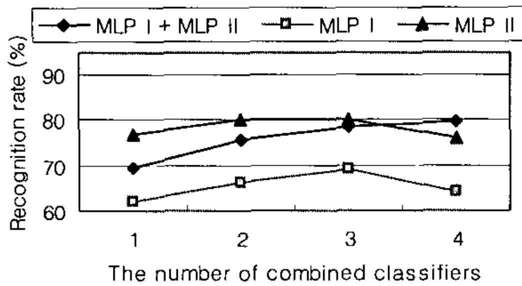

Fig. 5. Recognition rate of the ensemble in Lymphoma

TABLE III
THE BEST RECOGNITION RATE OF ENSEMBLE CLASSIFIERS (%)

	Generalization	Sensitivity	Specificity
Leukemia			
MLP I	97.1	92.9	100.0
MLP II	82.4	64.3	95.0
MLP I + MLP II	97.1	92.9	100.0
Colon			
MLP I	80.6	95.0	54.5
MLP II	77.4	95.0	45.5
MLP I + MLP II	87.1	95.0	72.7
Lymphoma			
MLP I	64	36.4	85.7
MLP II	76	72.7	78.6
MLP I + MLP II	92	90.9	92.9

This paper shows that the paired neural network classifier works and we can improve the classification performance by combining independent sets of classifiers learned from negatively correlated features, even when we use simple combination method like Bayesian approach.

V. CONCLUDING REMARKS

In this paper, we have proposed two ideal gene vectors to extract the informative genes for classification, selected the sets of genes based on the defined ideal gene vectors. We have combined the independent classifiers learned with the selected informative gene sets. In order to evaluate the usefulness of the proposed negative correlation, we have extracted the features using negative correlation and combined the neural network classifiers learned from negatively correlated features using Bayesian approach on three benchmark datasets.

The expression level of genes selected by negative correlation is clearly distinguishable between two classes. These patterns give enough information for the classification. The result of MLP I is different from that of MLP II on each dataset. While MLP I outperforms MLP II in Leukemia dataset and Colon dataset, MLP II outperforms MLP I in Lymphoma dataset. These results are caused by the characteristics of the dataset.

The experimental results show that the performance of the ensemble classifier is superior to the basis in all benchmark datasets. Moreover, the paired neural network classifier with negative correlation outperforms the ensemble classifiers without negative correlation. We have confirmed that negative correlation enables the ensemble classifier to work better by providing enough information for the classification to neural network classifiers.

ACKNOWLEDGEMENTS

This work was supported by Biometrics Engineering Research Center and a grant of Korea Health 21 R&D Project, Ministry of Health & Welfare, Republic of Korea.

REFERENCES

[1] Harrington, C. A., Rosenow, C., and Retief, J., "Monitoring gene expression using DNA microarrays," *Curr. Opin. Microbiol.*, vol. 3, pp. 285-291, 2000.

[2] Eisen, M. B. and Brown, P. O., "DNA arrays for analysis of gene expression," *Methods Enzymbol*, vol. 303, pp. 179-205, 1999.

[3] Dudoit, S., Fridlyand, J. and Speed, T. P., "Comparison of discrimination methods for the classification of tumors using gene expression data," *Technical Report 576*, Department of Statistics, University of California, Berkeley, 2000.

[4] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, N., "Tissue classification with gene expression profiles," *Journal of Computational Biology*, vol. 7, pp. 559-584, 2000.

[5] Cho, S.-B. and Ryu, J., "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features," *Proc. of the IEEE*, vol. 90 , no. 11, pp. 1744-1753, 2002.

[6] Dettling, M. and Bühlmann, P., "How to use boosting for tumor classification with gene expression data," *Technical Report*, Department of Statistics, ETH Zürich, 2002.

[7] Parmanto, B., Munro, P. W. and Doyle, H. R., "Reducing variance of committee prediction with resampling techniques," *Connection Science*, vol. 8, pp. 405-426, 1996.

[8] Raviv, Y. and Intrator, N., "Bootstrapping with noise: An effective regularization techniques," *Connection Science*, vol. 8, pp. 355-372, 1996.

[9] Sharkey, A. J. C. and Sharkey, N. E., "Diversity, selection, and ensembles of artificial neural nets," *Proc. of Third Int. Conf. on Neural Networks and Their Applications, IUSPIM, France*, pp. 205-212, 1997.

[10] Sharkey, A. J. C., "Adapting an ensemble approach for the diagnosis of breast cancer," *Proc. Int. Conf. Artificial Neural Networks, Skövde, Sweden*, pp. 281-286, 1998.

[11] Li, L., Weinberg, C. R., Darden, T. A. and Pedersen, L. G., "Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, no. 12, pp. 1131-1142, 2001.