Available online at www.sciencedirect.com

ScienceDirect

Neurocomputing 70 (2006) 187–199

NEUROCOMPUTING

www.elsevier.com/locate/neucom

Ensemble classifiers based on correlation analysis for DNA microarray classification

Kyung-Joong Kim*, Sung-Bae Cho

Department of Computer Science, Yonsei University, 134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea

Received 24 March 2005; received in revised form 14 March 2006; accepted 29 March 2006
Communicated by J. Zhang
Available online 14 June 2006

Abstract

Since accurate classification of DNA microarray is a very important issue for the treatment of cancer, it is more desirable to make a decision by combining the results of various expert classifiers rather than by depending on the result of only one classifier. In spite of the many advantages of mutually error-correlated ensemble classifiers, they are limited in performance. It is difficult to create an optimal ensemble for DNA analysis that deals with few samples with large features. Usually, different feature sets are provided to learn the components of the ensemble expecting the improvement of classifiers. If the feature sets provide similar information, the combination of the classifiers trained from them cannot improve the performance because they will make the same error and there is no possibility of compensation. In this paper, we adopt correlation analysis of feature selection methods as a guideline of the separation of features to learn the components of ensemble. We propose two different correlation methods for the generation of feature sets to learn ensemble classifiers. Each ensemble classifier combines several other classifiers learned from different features and based on correlation analysis to classify cancer precisely. In this way, it is possible to systematically evaluate the performance of the proposed method with three benchmark datasets. Experimental results show that two ensemble classifiers whose components are learned from different feature sets that are negatively or complementarily correlated with each other produce the best recognition rates on the three benchmark datasets.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Ensemble; DNA microarray; Correlation analysis

1. 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 [15]. This enables us to generate massive gene expression data. However, it has also led to many statistical and analytical challenges (due to the large number of genes but relatively few samples). We can analyze gene information rapidly and precisely by managing all this information at one time using several statistical methods and machine learning [10].

Cancer classification in practice (which relies on clinical and histopathological information) can be often incomplete or misleading. For this reason, DNA microarray technology has been applied to the field of accurate prediction and diagnosis of cancer. Molecular-level diagnostics with gene expression profiles can offer precise, objective and systematic cancer classification. Especially, accurate classification is a very important issue for the treatment of cancer. Since gene expression data usually consist of a huge number of genes, several researchers have been studying the problems of cancer classification using data mining methods, machine learning algorithms and statistical methods [1,9]. However, most researchers have evaluated only the performance of the feature selection method and the classifier in classifying gene expression data.

Many researchers have worked on the ensemble of multiple classifiers to improve the performance of classification in data mining or machine learning. The ensemble classifier increases not only the accuracy of the classification, but also leads to greater confidence in the results.

*Corresponding author.
E-mail addresses: kjkim@cs.yonsei.ac.kr (K.-J. Kim), sbcho@cs.yonsei.ac.kr (S.-B. Cho).

0925-2312/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2006.03.002

Theoretically, the performance of the ensemble classifier improves when the combined classifiers are mutually independent. Representative ensemble methods such as average combination, voting, weighted voting, the Bayesian approach and neural networks have been applied to many fields. However, these methods do not assure independent combined classifiers. On the other hand, 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 [1,8].

Usually, the classification of gene expression data requires two steps: feature selection and classification. There are many different kinds of feature selection and classification methods. The most important problem is their proper selection because the classification accuracy is highly sensitive to the choice. Because it is difficult to test all combinations (feature selection + classifier), making a robust classifier given various gene expression datasets is crucial. Recently, an ensemble of classifiers has gain popularity because of their improved generalization capability compared to the single classifier. In this paper, we attempt to propose a framework to construct an ensemble for high-performance classification of gene expression data. We argue that correlation analysis of various feature selection methods is useful for forming an ensemble.

This research aims to form an effective ensemble composed of classifiers based on correlation analysis of feature selection methods. This is so that we can classify gene expression datasets that are very different from the usual datasets. We attempted to use several methods for extracting informative features and combining classifiers learned from the negatively or complementarily correlated features, respectively. We adopted seven feature selection methods. These feature selection methods included the Pearson's and Spearman's correlation coefficients, the Euclidean distance, the cosine coefficient, information gain, mutual information and signal-to-noise ratio. Optimal feature–classifier pairs were chosen with correlation analysis. We adopted four types of classifiers to show the effectiveness of the proposed ensemble creation methods based on correlation analysis of feature selections. The classification methods included multi-layer perceptron (MLP), *k*-nearest neighbor (KNN), the support vector machine (SVM) and the structure adaptive self-organizing map system (SASOM). The methods for combining the classifiers were majority voting, weighted voting, the neural network approach and the Bayesian approach. We test the proposed method in three benchmark cancer datasets, and systematically analyze the usefulness of the ensemble classifiers on the basis of the correlation analysis given various settings.

2. DNA microarrays

Uncovering broad patterns of genetic activity, providing new understanding of gene functions and generating unexpected insights into biological mechanisms are the goals of microarray-based studies [15]. With the development and application of DNA microarrays, the expression of almost all human genes can now be systematically examined in human malignancies [14]. DNA sequences are initially transcribed into mRNA sequences. These mRNA sequences are translated into the amino acid sequences of the proteins that perform various functions. Measuring mRNA levels can provide a detailed molecular view of the genes. Measuring gene expression levels under different conditions is important for expanding our knowledge of gene functions. Gene expression data can help provide better understanding of cancer. It can also allow for the classification of individual tumors by their gene expression patterns, which may also describe and predict therapeutic resistance and sensitivity patterns [30].

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 the DNA spot is less than 250 μm, and macroarrays when the diameter is bigger than 300 μm. The arrays with small solid substrates are also referred to as DNA chips. Gene information can be investigated in a short time, because so many genes can be put on the DNA microarray to be analyzed.

There are two representative DNA microarray technologies: cDNA microarray technology and oligonucleotide microarray technology. cDNA microarrays are composed of thousands of individual DNA sequences printed in a high-density array on a glass microscope slide using a robotic arrayer. High-density oligonucleotide microarrays [15,21,31] are made using spatially patterned, light-directed combinatorial chemical synthesis and contain hundreds of thousands of different oligonucleotides on a small glass surface.

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, as shown in Fig. 1. For mRNA samples, the two samples are reverse-transcribed into Cdna and labeled using different mixed fluorescent dyes (red-fluorescent dye Cy5 and green-fluorescent dye Cy3). After the hybridization of these samples with the arrayed DNA probes, the slides are captured as images using a 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 [7,11]

$$\text{gene\_expression} = \log_2 \frac{\text{Int(Cy5)}}{\text{Int(Cy3)}}, \qquad (1)$$

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

The complexity of microarray data calls for data analysis tools that will effectively aid in biological data mining.

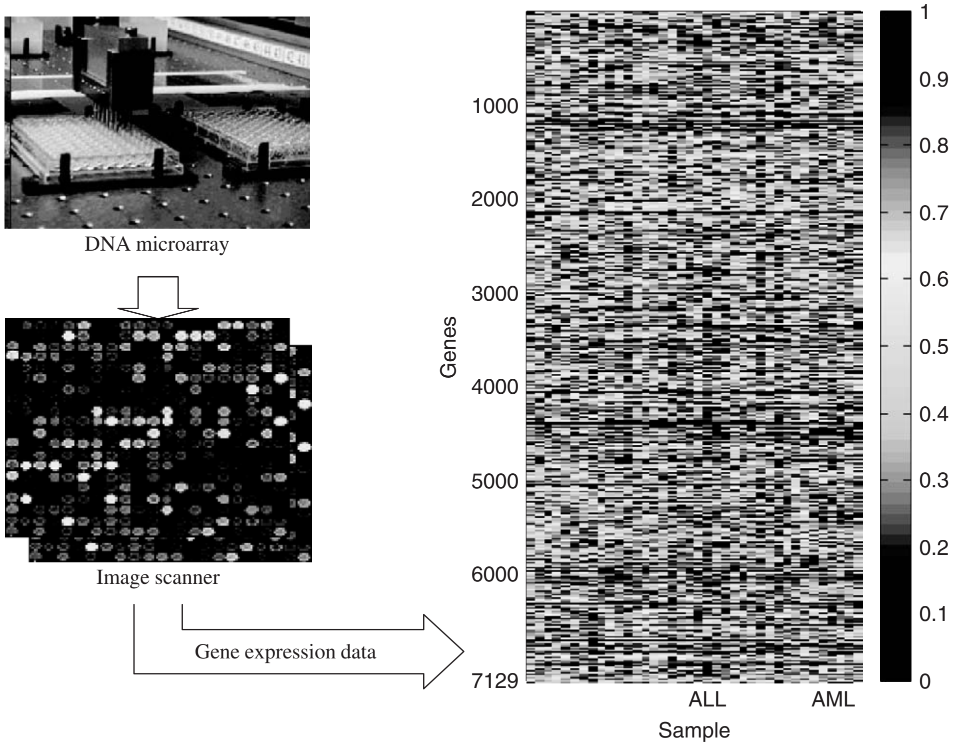

Fig. 1. General process of acquiring the gene expression data from DNA microarray. (This is an example of Leukemia cancer and there are two types of cancers including acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). A sample comes from patient.)

Since Golub's pioneering work [13], many researchers have been working on the classification of gene expression profiles. They have used various feature selection methods to select informative genes and several classifiers to classify the samples accurately. Golub's group used neighborhood analysis and weighted voting to classify acute leukemia samples [13]. They used 50 genes (the most closely correlated with the AML-ALL distinction) in the known (training) samples to compute the prediction strength. The classifier made strong predictions for 29 of 34 test samples.

In 2001, [18] proposed a classification scheme based on artificial neural networks (ANN). They first filtered genes expressed under minimal levels, and applied principal component analysis to reduce the dimensionality. They used 3-fold cross-validation, so that 2/3 of the data were used to train ANN and the remaining data were used to test the proposed system. This method was applied to small round blue cell tumor (SRBCT) data composed of four classes, and the sensitivities of the ANN model for diagnostic classification was 93% for Ewing family of tumors (EWS), 96% for rhabdomyosarcoma, and 100% for neuroblastoma and Burkitt lymphomas. They also revealed that 96 genes are the optimal number of genes for classifying SRBCT data.

In other experiments, Li used GA and KNN for gene selection and classification, respectively [19]. The GA/KNN method was applied to the systematical analysis of sensitivity, reproducibility and stability of the experiments. In the case of Nguyen's work, partial least squares were used for gene selection, and logistic discriminant and quadratic discriminant methods were used for the classification of expression profiles [29]. Ovarian, leukemia, lymphoma, colon and NCI60 datasets were used to verify the proposed method. The related works are summarized in Table 1.

The classification of gene expression data is an active research area and there are many approaches. An ensemble that uses multiple classifiers has shown better performance than other methods. Many researchers have been working on the ensemble of multiple classifiers to improve classification performance. [34] attempted to generate an ensemble of diverse base classifiers, building each base classifier using different subsets of features to obtain high accuracy of classification. They used several combinations of classifiers and compared the performance systematically. [33] used bagging and boosting to investigate the performance of ensemble machine learning in classifying gene expression data on cancer classification problems. They calculated the accuracy, reliability, sensitivity and specificity to measure the performance and the bagging recorded superior results to the single classifier and Adaboost methods. Related works on the ensemble approach are summarized in Table 2. Our contribution is a systematic approach for forming effective ensembles based on correlation analysis. Most previous works form ensembles without using the systematic method.

Table 1
Related works for classification of DNA microarray data

Researcher	Feature selection	Classification	Data	Remark
Golub [16]	Neighborhood analysis	Weighted votes	Leukemia	Pioneering study
Ben-Dor [4]	TNoM score	Nearest neighbor SVM	Colon Ovarian Leukemia	Comparing clustering with classification
Furey [13]	Signal-to-noise ratio	SVM	Ovarian Leukemia Colon	Applying SVM 70.9–83.9% of accuracy
Li [10]	GA	KNN	Lymphoma Colon	Several statistical analysis
Khan [29]	PCA	ANN	SRBCT	Applying ANN to multiclass data Revealing optimal number of genes for SRBCT
Nguyen [15]	PLS PCA	Discriminant analysis	Ovarian Leukemia Lymphoma Colon NCI60	Applying to various datasets
Cho [30]	Pearson correlation coefficients Cosine coefficients Information gain	MLP KNN SVM Decision tree	Leukemia	Using various feature selection methods and classifiers
Bicciato [31]	PCA	Soft independent modeling of class analogy	Leukemia SRBCT	Applying modified PCA to the multiclass problem

3. Ensemble classifiers based on correlation analysis

Ensemble classifiers consist of two stages: informative features selection based on correlation analysis and ensemble classifiers learned from the informative features. Informative features selection based on correlation analysis includes three steps:

- defining the ideal feature vector,
- selecting informative genes based on similarity with the ideal feature vector, and
- selecting informative gene subsets for the ensemble classifiers using correlation analysis.

The basic idea of the ensemble classifier scheme is to develop several trained classifiers with feature sets selected by correlation analysis and combine them. Through correlation analysis, we can select informative features for the ensemble classifiers. This naturally raises the question of obtaining a consensus on the results of the individual classifiers. This section is organized as shown in Fig. 2.

The idea of this approach is to increase the diversity of the accurate classifiers for the ensemble. The correlation analysis of features can make mutually exclusive feature sets that are critical to generate highly diverse and accurate classifier sets. In the case of positively correlated features given correlation among the ranks from the feature selection methods, the selected features of the two feature selection methods are very similar because they provides similar ranks. If one feature gets relatively high ranks in one method, it gets high ranks in other methods. This is positively correlated. But in negatively correlated features, some high ranked features in one method get relatively low ranks because they use different metrics to evaluate the relevance of features. Though negatively correlated feature selection methods cannot provide perfect mutually independent feature sets and sometimes there can be overlapped genes, there can be relatively high probability of independent gene sets compared to positively correlated feature selection methods. The usefulness of negatively correlated classifiers for the ensemble is frequently addressed by many publications [5,23,35]. To achieve the goal, we used negatively correlated features and expected increased diversity in the ensemble. If the behavioral characteristics of the input data (features) are positively correlated, it might generate similar classifiers and there is only a small performance variation.

Table 2
Related works on the ensemble classifiers in medical area

Researcher	Method	Dataset	Remark
Ben-Dor [4]	Boosting	Colon Ovarian Leukemia	—
Tsymbal [25]	Simple Bayesian classifier Cross-validation majority Weighted voting Dynamic selection Dynamic voting Dynamic voting with selection	Small, medium and large acute abdominal pains	Ensembles of simple Bayesian classifiers
Tan [26]	C 4.5 (decision tree) Bagging Adaboost	Leukemia Breast cancer Central nervous system Lung cancer	Systematic comparison of bagging, Adaboost and single C 4.5
Lo [27]	Backpropagation-artificial neural network Voting Average	Breast microcalcification	Combining three sources: image processing, BI-RADS and history
Masulli [28]	Random Voronoi resampling	Synthetic Leukemia	Applying to both linear and non-linear cases

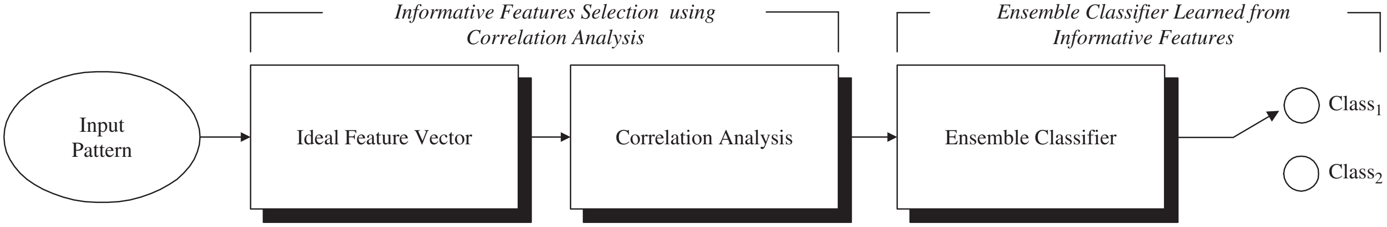
Informative Features Selection using Correlation Analysis
Ensemble Classifier Learned from Informative Features
Input Pattern
Ideal Feature Vector
Correlation Analysis
Ensemble Classifier
Class$_1$
Class$_2$

Fig. 2. The overview of ensemble classifier framework.

3.1. Ideal feature vectors

Among the thousands of genes whose expression levels are measured, not all are needed for classification. Microarray data consist of a large number of genes in small samples. For efficient classification, we need to find out the informative features from input observation. This process is referred to as gene selection [19].

Suppose that we have a $M \times N$ training set where M is the number of samples (input vector) and N is the number of features (dimensionality of the input vector). The ith feature of samples, g_i, can be expressed as

$$g_i = (e_1, e_2, e_3, \ldots, e_M), \tag{2}$$

where e is the data and $i = 1$–N. We want to know the locations of the informative k features out of M. Suppose g_{ideal} is an ideal vector representing class c_j. 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 the classes, which tells the feature goodness. Modeling g_{ideal}, we can use prior knowledge and intuitional experience about classes. An ideal gene pattern that belongs to the tumor class is defined by $g_{\text{ideal}} = (1,1,\ldots,1,0,0,\ldots,0)$, so that all the elements from the tumor samples are 1 and the others are 0

$$g_{\text{ideal}} = (e'_1, e'_2, e'_3, \ldots, e'_M). \tag{3}$$

Measuring the similarity of g_i and g_{ideal} using similarity measures such as the Pearson's correlation coefficient (PC), the Spearman coefficient (SC), the Euclidean distance (ED) and the coefficient (CC) in Table 1, the most similar 25 genes are used for classification. The most informative 25 genes are selected using information-theoretic feature selection methods such as information gain (IG), mutual information (MI) and signal-to-noise ratio in Table 3. For a gene g_i that is continuous-valued, we create a new Boolean attribute g'_i that is true if $g_i < \sigma_i$ and false otherwise. How to select the best value for the threshold σ_I is described in a machine learning text book [28]. In IG formula, $P(g_i|c_j)$ means the relative frequency of samples

Table 3
Mathematical formula for similarity measure of g_i and g_{ideal}

$$\mathrm{PC}(g_i, g_{\text{ideal}}) = \frac{\sum g_i g_{\text{ideal}} - (\sum g_i \sum g_{\text{ideal}})/N}{\sqrt{\left(\sum_{g_i}^{2} - (\sum g_i)^2/N\right)\left(\sum g_{\text{ideal}}^2 - (\sum g_{\text{ideal}})^2/N\right)}}$$

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

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

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

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

$$\mathrm{IG}(g_i, c_j) = P(g_i|c_j)\log\frac{P(g_i|c_j)}{P(c_j)P(g_i)} + P(\bar{g}_i|c_i)\log\frac{P(\bar{g}_i|c_j)}{P(c_j)P(\bar{g}_i)}$$

$$\mathrm{MI}(g_i, c_j) = \log\frac{P(g_i, c_j)}{P(g_i)P(c_j)}$$

$$\mathrm{SN}(g_i, c_j) = \frac{\mu_{c_j}(g_i) - \mu_{\bar{c}_j}(g_i)}{\sigma_{c_j}(g_i) + \sigma_{\bar{c}_j}(g_i)}$$

($g_i < \sigma_i$) given the label is c_j. In MI formula, $P(g_i,c_j)$ is the relative frequency of samples ($g_i < \sigma_i$) labeled as c_j.

3.2. Correlation analysis

Theoretically, the more features, the more effective the classifier is to solve problems [5]. But features that have overlapped in the feature space may cause the redundancy of irrelevant information and result in the counter-effect of overfitting. When there are N feature selection methods, the set of non-linear transformation functions that change the observation space into the 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 [5]:

$$I(\phi_k) = \frac{a\sum A_i}{N/2\sum_{j=1, j\neq i}^{N} d_{ij}} + b, \tag{4}$$

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 the constants. The number of genes is m

$$d_{ij} = \sum_{p=1}^{m} \mathrm{PC}(\varphi_i(g_p), \varphi_j(g_p)). \tag{5}$$

The higher the dependency of a pair of features, the smaller the amount of classification information $I(\phi_k)$. As the extent of the area occupied by the features becomes larger, the amount of classification information $I(\phi_k)$ also increases. If we keep the number of features larger, the numerator of the equation is larger because the extent of the area occupied by the features becomes wider. Although the numerator increases, $I(\phi_k)$ mainly decreases without keeping d_{ij} small. Therefore, it is more desirable to use a 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. Negatively correlated feature sets are a good example of mutually independent features.

The feature sets chosen by the negatively correlated methods could be very disjointed, and the classifiers with these feature selection methods are trained in a less dependent feature space. In positively correlated cases, there are many common genes between the two feature sets, and this means that the ensemble classifiers learn from the highly correlated feature sets. Since the two sets of classifiers are trained in mutually dependent feature spaces, it is hard to expect any performance improvement when the classifiers are combined by the ensemble method. Thus, we have proposed two different approaches for achieving greater performance improvement.

In complementarily correlated approaches, two different feature selection methods are negatively correlated but they can have simultaneously common features and negatively correlated features. In negatively correlated approaches, the correlation between the ideal vector and the individual gene is considered. There are two feature sets: one is composed of features correlated with the ideal vector of the normal sample and another contains features correlated with the ideal vector of the cancer sample. Genes from the two sets show completely negative correlations.

- *Correlation analysis for complementarily correlated features*: Complementary correlation means partial correlation (there are negatively correlated genes and positively correlated genes). Using feature selection, we get a set of informative features from the data. In order to choose complementarily correlated features, we plotted the distribution of g_i from two feature selection methods. If the two features are correlated, the distribution will be in the (−) direction, otherwise in the (+) direction. The overlapped genes of the correlated features can discriminate classes, and the other genes (not overlapped among the combined features) can supplement to search the solution spaces complementarily. For example, with the Colon dataset found in the experimental results, gene 1659 and gene 550 are high-ranked in terms of both the Pearson's correlation coefficient and the cosine coefficient, and gene 440 is high-ranked in terms of both the Euclidean distance and the cosine coefficient. This subset of two features might play an important role in classification. Experimental results show that the ensemble classifier of complementarily correlated features works better than the ensemble of uncorrelated features or base classifiers.
- *Correlation analysis for negatively correlated features*: To utilize the informative features to train classifiers, we attempted 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). We then selected the sets of informative genes with high similarity to each ideal gene vector. Since the Pearson's correlation coefficient of two ideal gene vectors is −1, the 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. 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. Similarity measures used for negative correlation are the Pearson correlation coefficient, the Spearman correlation coefficient, the Euclidean distance and the cosine coefficient, as shown in Table 3.

3.3. Ensemble classifier

We used the MLP, the self-organizing map (SOM), the SASOM, the SVM, the decision tree and the KNN systems as classifiers [5].

We applied complementarily correlated features to a classification framework. Given k features and n classifiers, there were $k \times n$ feature–classifier combinations. There were ${}_{k\times n}C_m$ possible ensemble classifiers when m feature–classifier combinations were selected as the ensemble classifiers. We plotted the distribution of g_i from the feature selection methods and used complementarily correlated features for the ensemble classifiers. Classifiers were trained using the complementarily correlated features selected, and finally a combining module was used to find the outputs. After the classifiers train independently with some features to produce their own outputs, the final answer can be judged by a combining module, where the majority voting, weighted voting or Bayesian combination can be adopted.

We also combined the neural networks learned from negatively correlated gene subsets, since combining the heterogeneous classifiers can help increase the performance of the classification. We chose the Bayesian approach as the ensemble classifier. The Bayesian approach can solve the problem of a tie-break between the ensemble classifiers by using a priori knowledge of each combined classifier.

- *Majority voting*: This is a simple ensemble method that selects the class most favored by the base classifiers. Majority voting has some advantages in that it does not require any previous knowledge nor does it require any additional complex computation to decide. Where c_i is the class i ($i = 1,\ldots m$), and $s_i(\text{classifier}_j)$ is 1 if the output of the jth classifier classifier$_j$ equals the class i otherwise 0, majority voting is defined as follows:

$$c_{\text{ensemble}} = \operatorname*{arg\,max}_{1\leqslant i\leqslant m}\left\{\sum_{j=1}^{k} s_i(\text{classifier}_j)\right\}. \tag{6}$$

- *Weighted voting*: A poor classifier can affect the result of the ensemble in majority voting because it gives the same weight to all classifiers. Weighted voting reduces the effect of the poor classifier by giving a different weight to a classifier based on the performance of each classifier. The weights of the classifiers are determined from the accuracy on the training dataset. Where w_j is the weight of the jth classifier, weighted voting is defined as follows:

$$c_{\text{ensemble}} = \operatorname*{arg\,max}_{1\leqslant i\leqslant m}\left\{\sum_{j=1}^{k} w_j s_i(\text{classifier}_j)\right\},$$
$$w_i = \frac{1-E_i}{\sum_k (1-E_k)}. \tag{7}$$

- *Bayesian combination*: While the majority voting method combines classifiers with their results, the Bayesian combination makes the error possibility of each classifier affect the final result. The method combines the classifiers with different weights by using the previous knowledge of each classifier. Where the k classifiers are combined, c_i, $i = 1,\ldots,m$, is the class of a sample, $c(\text{classifier}_j)$ is the class of the jth classifier, and w_i is the a priori possibility of class c_i, the Bayesian combination is defined as follows:

$$c_{\text{ensemble}} = \operatorname*{arg\,max}_{1\leqslant i\leqslant m}\left\{\eta \prod_{j=1}^{k} P(c_i|c(\text{classifier}_j))\right\}. \tag{8}$$

4. Experiments

4.1. Environment

We used three well-known gene expression profiles which were obtained from the Internet: a Leukemia dataset, a Lymphoma cancer dataset and a Colon cancer dataset. Microarray samples, while clearly noisy, contain in addition significant "atypical" sources of variation: across labs, across people carrying out the experiments, type of arrays used, etc. It is critical to try to deal with such types of variations so that the resulting classification scheme can be generalized in practice. Nowadays, we can download many sets of gene expression data that deal with different diseases (brain tumors, breast cancer, colon cancer, leukemia Lymphoma cancer, etc). However, only some of them are frequently used for benchmark purposes. Among them, the Leukemia, Lymphoma and Colon sets are the most popular. Lymphoma data from Stanford University is profiled using a specialized DNA microarray called a Lympochip and colon cancer data from Princeton University uses a general oligonucleotide DNA microarray. Each data is profiled using different types of machines and from different labs.

In our experiments, three representative cancer datasets were used. The Leukemia dataset consisted of 72 samples: 25 samples of AML and 47 samples of ALL [1]. A total of 38 out of 72 samples were used as training data and the remaining samples were used as test data. Each sample contained 7129 gene expression levels. The Colon dataset consisted of 62 samples of colon epithelial cells taken from colon-cancer patients [1]. Each sample contained 2000 gene expression levels. A total of 31 out of 62 samples were used as training data and the remaining samples were used as test data. The Lymphoma dataset consisted of 24 samples of GC B-like and 23 samples of activated B-like [26]. A total of 22 out of 47 samples were used as training data and the remaining samples were used as test data (http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode = view&paper_id = 43).

B cell diffuse large cell Lymphoma (B-DLCL) is a heterogeneous group of tumors, based on significant variations in morphology, clinical presentation and response to treatment. Gene expression profiling has revealed two distinct tumor subtypes of B-DLCL: germinal center B cell-like DLCL and activated B cell-like DLCL. This Lymphoma cancer dataset consisted of 24 samples of GC B-like and 23 samples of activated B-like (http://www.genome-stanford.edu/lymphoma).

The Colon dataset consisted of 62 samples of Colon epithelial cells taken from Colon cancer patients. Each sample was taken from tumors and normal healthy parts of the Colons of the same patients and measured using high-density oligonucleotide arrays. Each sample contained 2000 gene expression levels. Although the original data consisted of 6000 gene expression levels, 4000 out of 6000 were removed based on confidence in the measured expression levels. A total of 40 of 62 samples were from Colon cancer samples and the remaining ones were from normal samples (http://microarray.princeton.edu/oncology/affydata/index.html). Because the number of samples is very small, it is not easy to partition the datasets into three parts (training/validation/test sets). Many previous works in this field also used the training/test partition.

For feature selection, each gene was scored based on the similarity measure, and the 25 top-ranked genes were chosen as the feature of the input pattern. The feature selection is based on only training data. There currently are no reports on the optimal number of genes, but our previous study indicates that 25 is a reasonable number [5]. For classification, we used a 3-layered MLP with eight hidden nodes, two output nodes, 0.05 learning rate and 0.9 momentum. The KNN was used with $k = 5$. Similarity measures used in the KNN were the Pearson's correlation coefficient and the Euclidean distance. The SASOM system was used by a 4×4 initial map with rectangular topology, 0.05 initial learning rate, 1000 initial maximum iteration, 10 initial radius, 0.02 final learning rate, 10,000 final maximum iteration and three final radius. The number of nodes in the SASOM system will increase by dynamically splitting the nodes. We used SVMs with a linear function and an RBF function as the kernel function. The input of each classifier is a vector composed of values of the selected features and the output is a class label (cancer/normal) of the sample. Details of the construction of classifiers are described in [6]. The parameters are derived from many experiments.

In the complementarily correlated features section, the size of the ensemble was 7 and the members were chosen from 42 combinations (six classifiers and seven feature selection methods) by using correlation analysis. In the negatively correlated genes section, the size of the ensemble ranged from 2 to 4. There are relatively small cases of the negatively correlated feature selection methods because its condition is more rigorous than complementarily correlated one. For two ideal vectors, there were four different distance measures for selecting the feature sets. In total, there were eight different feature sets and eight different MLPs trained from them. From these, we chose the members of the ensemble using the proposed correlation analysis.

4.2. Results of analysis of the ensemble with complementarily correlated features

Fig. 3 shows the correlation of the Euclidean distance, the Pearson's correlation coefficient and the cosine coefficient on the basis of plotted distribution. As shown in Fig. 3, they are correlated in the Colon dataset. Table 4 shows the genes ranked by them and the value of the genes when using each method. There are some overlapped genes among them. The bold-faced figures mean the overlapped genes of those features. This indicates that the overlapped genes of the correlated features can discriminate classes and the other genes (not overlapped among the combined features) can supplement to search the solution spaces complementarily. For example, gene 1659 and gene 550 of the Colon dataset are high-ranked in both the Pearson's correlation coefficient and the cosine coefficient, and gene 440 is high-ranked in both the Euclidean distance and the cosine coefficient. This subset of two features might play an important role in classification.

There were six classifiers and seven feature selection methods used in this paper, which produced 42 feature–classifier pairs. Because the size of the ensemble is seven, there were ${}_{42}C_7$ candidates possible. Recognition rates by the ${}_{42}C_7$ ensemble classifiers are shown in Table 5. The

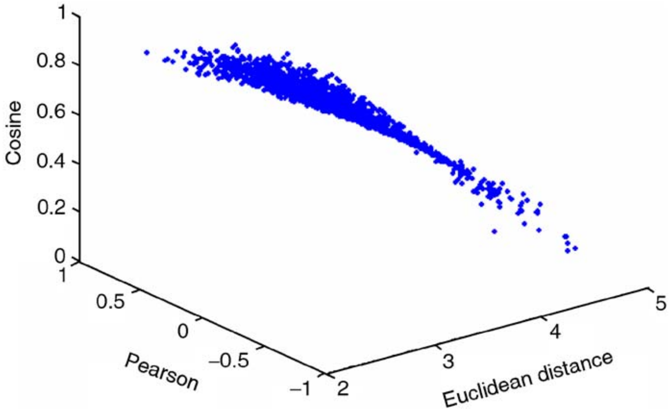

Fig. 3. Correlation of Euclidean distance, Pearson's correlation coefficient and cosine coefficient.

Table 4
Genes ranked by Euclidean distance, Pearson's correlation coefficient and cosine coefficient

Rank	Euclidean	Pearson	Cosine
1	619 (2.262385)	619 (0.681038)	619 (0.895971)
2	767 (2.335303)	1771 (0.664378)	1772 (0.875472)
3	704 (2.374358)	1659 (0.634084)	767 (0.874914)
4	187 (2.388404)	550 (0.631655)	1771 (0.873892)
5	207 (2.410640)	187 (0.626262)	1659 (0.870115)
6	887 (2.473033)	1772 (0.621581)	187 (0.867285)
7	635 (2.474971)	1730 (0.615566)	704 (0.866679)
8	1915 (2.498611)	1648 (0.614949)	1208 (0.866029)
9	1046 (2.506833)	365 (0.614591)	550 (0.864547)
10	1208 (2.512257)	1208 (0.603313)	1546 (0.856904)
11	482 (2.520699)	1042 (0.602160)	251 (0.855841)
12	1771 (2.525080)	1060 (0.601712)	1915 (0.855784)
13	1993 (2.529032)	513 (0.596444)	440 (0.855453)
14	62 (2.546894)	767 (0.594119)	1263 (0.854854)
15	1772 (2.547455)	1263 (0.591725)	1060 (0.854829)
16	1194 (2.549244)	138 (0.587851)	965 (0.854137)
17	1594(2.551892)	1826 (0.584774)	1648 (0.854119)
18	199 (2.557360)	1546 (0.582293)	1942 (0.853586)
19	1867 (2.587469)	141 (0.579073)	513 (0.852270)
20	959 (2.589989)	1227 (0.574537)	1042 (0.851993)
21	440 (2.593881)	704 (0.569022)	1993 (0.851753)
22	480 (2.594514)	1549 (0.562828)	365 (0.851205)
23	1546 (2.604907)	1489 (0.561003)	1400 (0.849531)
24	399 (2.613609)	1724 (0.559919)	207 (0.849084)
25	1060 (2.614100)	1209 (0.559778)	271 (0.848481)

Table 5
The best recognition rate by ensemble classifier (%)

		Leukemia	Colon	Lymphoma
MV	Best	97.1	93.5	100.0
	Average	85.1	72.9	77.3
WV	Best	97.1	93.5	100.0
	Average	87.3	73.1	79.1
BC	Best	97.1	93.5	100.0
	Average	92.8	74.8	85.2

purpose of this exhaustive experiment is to show the effectiveness of the proposed ensemble creation methods based on correlation analysis. In the negatively correlated features, two neural networks were used and they were trained using negatively correlated features. MV means the ensemble classifier using the majority voting method, WV means the ensemble classifier using the weighted voting method, and BC means the ensemble classifier using the Bayesian combination. While there is little difference in the best recognition rate of the ensemble classifier according to the ensemble method or the number of combined classifiers, there is a difference in the average recognition rate of the ensemble classifier.

The best recognition rate of the ensemble classifiers is 97.1% in the Leukemia dataset, 93.5% in the Colon dataset, and 100.0% in the Lymphoma dataset. Compared with the best recognition rates of the base classifiers 97.1%, 83.9% and 92.0% for each dataset, the performance of the ensemble classifiers is superior. The best result of the Leukemia dataset (97.1%) is obtained by all the basis classifiers except the SASOM system. In other datasets, the performance of the ensemble classifiers outperformed the best classifiers. For all the datasets, the ensemble classifier with all classifiers produced the worst result. If we observe the classifiers of the best ensemble classifier, we find features more important to the result. In other words, in the ensemble classifiers, there must be classifiers that use the Euclidean distance and the Pearson's correlation coefficient. The other classifier is the one with the cosine coefficient, mutual information or information gain. This fact is also prominent in the Lymphoma dataset. Most of the best ensemble classifiers are classifiers with information gain, signal-to-noise ratio and the Euclidean distance or classifiers with information gain, signal-to-noise ratio and the Pearson's correlation coefficient.

4.3. *Results of analysis of ensemble with negative correlated features*

Fig. 4. shows the expression level of the informative gene subsets selected on the basis of negative correlation, top-ranked genes in terms of similarity with ideal vector genes, chosen by the Pearson's correlation coefficient method in the Leukemia dataset. They are the expression levels of the genes chosen by the Pearson's correlation coefficient. About 1–27 samples are the class of ALL and 28–38 samples are the class of AML. As shown in Fig. 4, the expression levels of 50 genes are clearly distinguishable between the two classes. The 25 genes are underexpressed in ALL and overexpressed in AML. In Fig. 3, the other 25 genes are overexpressed in ALL and underexpressed in AML. The expression levels of 25 genes selected based on ideal genes (0,0,…,0,1,1,…,1) are more separable between the ALL and AML classes than those of the other 25 genes selected based on ideal genes (1,1,…,1,0,0,…,0).

By using correlation analysis, two feature sets that are negatively correlated are generated and two MLPs (referred to as MLP I and MLP II) are trained using each of them. In distance-based gene selection methods (Euclidean, correlation coefficient, Pearson correlation and Spearman correlation), the distance between gene expression data and ideal feature vector is calculated. If there are N samples, the size of ideal feature vector is N. Its element can have 0 or 1. If the sample's class is cancer the value of element is 1 and vice versa. Let us assume that the gene expression values of gene 1 are as follows. The number of sample size is 10. Gene $1 = (0.5, 0.4, 0.3, 0.2, 0.6, 0.7, 0.8, 0.4, 0.9, 0.3)$. If the sample from 1 to 5 is cancer and the others are normal, the ideal feature vector is $(1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$. The distance between the two vectors is used as a ranking measure. On the other hand, it is possible to use 1 for normal and 0 for cancer in ideal vector. In this case, the ideal vector is $(0, 0, 0, 0, 0, 1, 1, 1, 1, 1)$.

Based on the ideal feature vectors, the feature selection measure can choose different feature subsets. If the ideal vector is $(1,1,1,1,1,0,0,0,0,0)$, the genes which are expressed highly for cancer sample and are regulated for normal sample get high rank. However, if the ideal vector is $(0,0,0,0,0,1,1,1,1,1)$, the genes which are expressed low for cancer sample and high for normal sample get high rank. This produces two different gene subsets that have opposite characteristics and we call it as negatively correlated. MLP I is trained from the gene subsets using ideal feature vector $(1,1,1,1,1,0,0,0,0,0)$ and MLP II is trained from the gene subsets using ideal feature vector $(0,0,0,0,0,1,1,1,1,1)$. MLP I and MLP2 are not related to the pair of feature selection methods. For each feature selection method (only distance-based methods), there are two gene subsets (each subset has 25 genes).

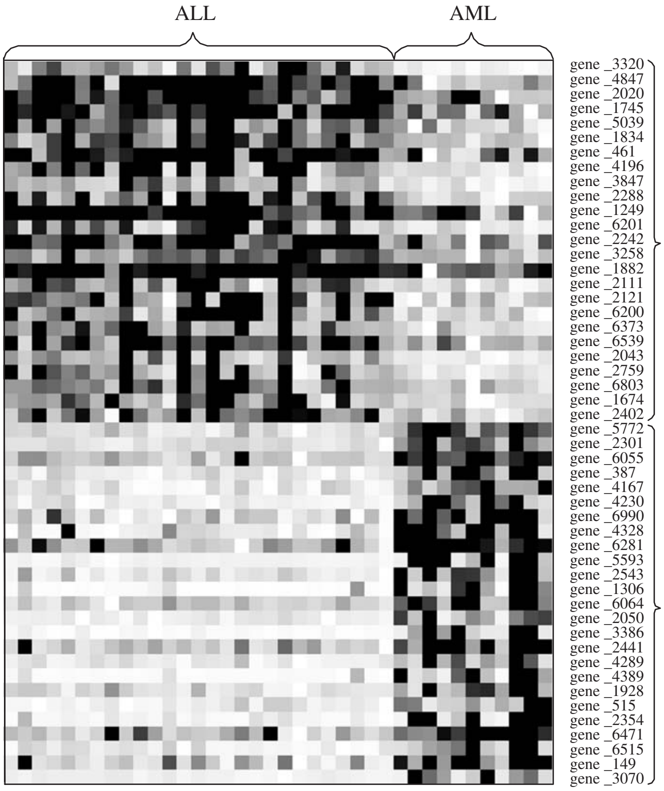

Fig. 4. The informative gene subsets chosen by Pearson's correlation coefficient in Leukemia dataset.

Table 6 shows the best recognition rate of the basis classifiers in each dataset. The basis classifier referred the single classifier (one feature selection method + one classifier). In this case, MLP is used as a classifier. Because the proposed method can only be considered for the distance-based methods (PC, SC, ED and CC), other three feature selection methods (IG, MI and SN) are naturally ignored. Basis classifier means the combination of the four distance-based feature selection methods and MLP. We have identified the single classifier as "single classifier in MLP I" and "single classifier in MLP II." The purpose of the experimentation is the comparison of "single classifier in MLP I," "single MLP II," "ensemble of MLPs from MLP I," "ensemble of MLPs from MLP II," "ensemble of MLPs from MLP I and MLP II." The table shows the performance of single classifier in MLP I and single classifier in MLP II.

In the Leukemia dataset, single classifier in MLP I with the Pearson's correlation coefficient produces the best recognition rates, 97.1%, among the feature–classifier combinations. In the Colon dataset, single classifier in MLP I with the cosine coefficient produces the best recognition rate, 83.9%. In the Lymphoma dataset, single classifier in MLP II with Spearman's correlation coefficient produces the best recognition rate, 88.0%. While single classifier in MLP I outperforms single classifier in MLP II in the Leukemia dataset and the Colon dataset, single classifier in MLP II outperforms single classifier in MLP I in the Lymphoma dataset.

Fig. 5 shows the average and the best recognition rates of the ensemble classifiers for the Lymphoma dataset. In case of the negatively correlated feature set (MLP I + MLP II), eight diverse feature sets were produced with two ideal feature vectors (ideal Gene A and ideal Gene B) and four similarity measures. The classifiers learned with eight diverse feature sets were combined using the Bayesian approach. The average recognition rate means the average of all possible ${}_8C_k$ ($k = 2$, 3 and 4) combinations of ensemble classifiers.

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 (Figs. 5–7),

Table 6
Recognition rate with features and classifiers (%) (the best recognition rate for each combination)

	Leukemia		Colon		Lymphoma	
	Single classifier in MLP I	Single classifier in MLP II	Single classifier in MLP I	Single classifier in MLP II	Single classifier in MLP I	Single classifier in MLP II
PC	97.1	79.4	74.2	77.4	64.0	72.0
SC	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

but outperforms in the best recognition rate (Table 7). While the best recognition of the ensembles of MLP I and MLP II decreases as the number of combined classifiers increases, the best recognition of the ensemble of the negatively correlated coefficient feature set increases. Increasing the number of base classifiers dose not always guarantee the performance improvement [36]. In Fig. 5, the statistical significance test between MLP I + MLP II and MLP II where the ensemble size is 4 shows that the difference is meaningful ($t = 2.055 \approx 2.09$, $p = 0.05$, the number of experiments is 10). In Lymphoma, the difference among the accuracies is clear. In Colon, the difference between MLP I and others is clear. However, the difference between MLP II and MLP I + MLP II is near 0. Because they are relatively clear, we do not use statistical significant test. The case (member is 4) in Lymphoma is very special and worth to analyze because the proposed MLP I + MLP II shows the best performance.

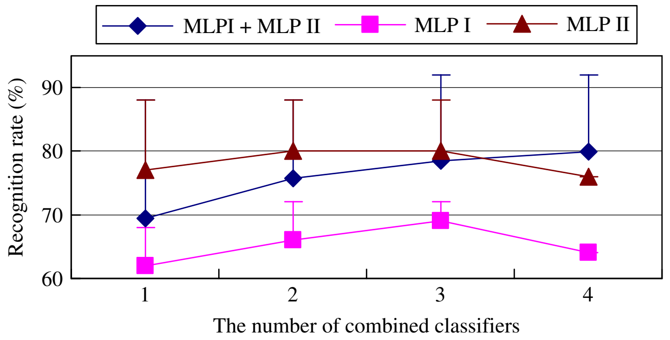

Fig. 5. Recognition rate of the ensemble in Lymphoma dataset (MLP I means “ensemble of MLPs from MLP I,” MLP II means “ensemble of MLPs from MLP II,” and “MLP I + MLP II” means “ensemble of MLPs from MLP I and MLP II”).

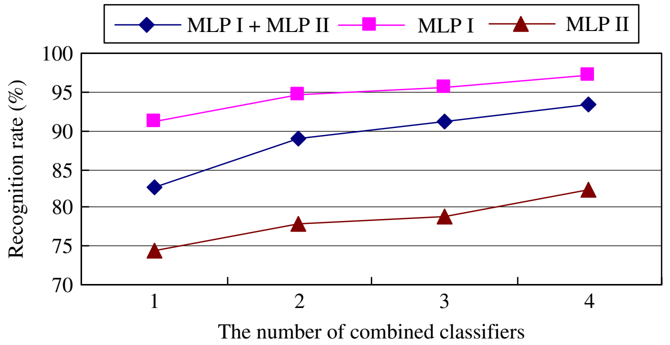

Fig. 6. Recognition rate of the ensemble in Leukemia.

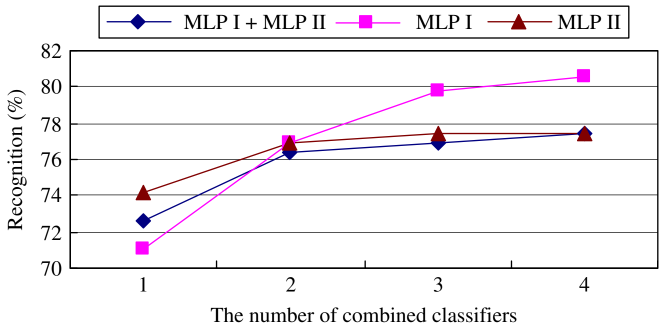

Fig. 7. Recognition rate of the ensemble in Colon.

Table 7
Best recognition rate of ensemble classifier learned negatively correlated feature sets (%)

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

4.4. Related works

Table 8 shows relevant works on cancer classification in three benchmark cancer datasets, the Leukemia, Colon and Lymphoma datasets. Many researchers have been studying problems of cancer classification [1,9,12,19,22,29] and clustering [2,13,16,32] using gene expression profile data and attempting to propose optimal classification techniques to work out these problems. Some produce better results than others, as shown in Table 8. For the Leukemia dataset, our method produces the best recognition rate, 97.1%, while other methods produce 84.6–95.8%. For the Colon dataset, our method produces a result of 83.9–93.5% compared to that of the others (72.6–94.1%.) For the Lymphoma dataset, our method produces also the best recognition rate of 92.0–100.0%, compared to that of the others (90.0–98.1%.) For all datasets, our methods and the methods in Nguyen and Rocke outperform the others [29]. The maximum accuracy for leukemia datasets published by other researchers is 100% in [5,24]. However,

Table 8
Relevant works on cancer classification (%)

Author	Leukemia	Colon	Lymphoma
Our method	97.1	83.9–93.5	92.0–100.0
Furey et al. [13]	94.1	90.3	—
Li et al. [14]	94.1	—	—
Li et al. [10]	84.6~	94.1~	—
Ben-Dor et al. [4]	91.6–95.8	72.6–80.6	—
Dudoit et al. [3]	95.0~	—	90.0~
Nguyen et al. [15]	94.2–96.4	87.1–93.5	96.9–98.1

that specific dataset is known for the mis-labeling of one sample. 97.1% is the maximum accuracy that was achieved if the label correction is taken into account.

5. Concluding remarks

In order to predict the cancer class of patients, we have illustrated a classification framework that combines sets of classifiers using correlation information. The results clearly show that the suggested ensemble classifiers work. We could also improve classification performance by combining classifiers from two independent features and by combining independent sets of classifiers learned from negatively correlated features, even when we use simple combination methods of voting and the Bayesian approach. For all the datasets, the Bayesian combination is the best among the three ensemble methods.

The experimental results also imply some correlations between features, which might guide researchers to choose or devise the best ensemble classifiers for their problems in bioinformatics. Based on the results, we have developed the optimal feature–classifier combination to produce the best performance.

Moreover, the ensemble classifier with negative correlation outperformed the ensemble classifiers without negative correlation. We confirmed that high correlation and the negative correlation on the basis of the correlation analysis enable the ensemble classifier to work better by providing appropriate information for the classification to classifiers. Our experimental results showed the possibility of performance improvement by using correlation analysis. Though it showed relatively low performance in averaged accuracy, it is possible to form the best ensemble. There could be many strategies to find the best one. Finding the strategy must be dealt in another paper as a future work.

Feature selection is a very important issue in classifying gene expression data and there are two basic approaches: filter and wrapper approaches [17]. Li's work is a kind of wrapper approach [20]. In this paper, we have chosen the filter approach and the adoption of another feature selection method needs to be investigated in the future.

Acknowledgements

This research was supported by Brain Science and Engineering Research Program sponsored by Korean Ministry of Commerce, Industry and Energy.

References

[1] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, N. Yakhini, Tissue classification with gene expression profiles, J. Comput. Biol. 7 (2000) 559–584.

[2] A. Ben-Dor, R. Shamir, Z. Yakhini, Clustering gene expression patterns, J. Comput. Biol. 6 (1999) 281–297.

[3] S. Bicciato, A. Luchini, C.Di. Bello, PCA disjoint models for multiclass cancer analysis using gene expression data, Bioinformatics 19 (2003) 571–578.

[4] S.-B. Cho, Exploring features and classifiers to classify gene expression profiles of acute leukemia, Int. J. Pattern Recogn. Artif. Intell. 16 (7) (2002) 1–13.

[5] S.-B. Cho, J. Ryu, Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features, Proc. IEEE 90 (11) (2002) 1744–1753.

[6] S.-B. Cho, H.-H. Won, Data mining for gene expression profiles from DNA microarray, J. Software Eng. Knowledge Eng. 31 (6) (2003) 593–608.

[7] J. Derisi, V. Iyer, P. Brosn, Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278 (1997) 680–686.

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

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

[10] M.B. Eisen, P.O. Brown, DNA arrays for analysis of gene expression, Methods Enzymol. 303 (1999) 179–205.

[11] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Bostein, Cluster analysis and display of genome-wide expression patterns, Proc. Natl. Acad. Sci. USA 95 (1998) 14863–14868.

[12] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (10) (2000) 906–914.

[13] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. GaasenBeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Blomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene-expression monitoring, Science 286 (1999) 531–537.

[14] G.M. Hampton, H.F. Frierson Jr., Classifying human cancer by analysis of gene expression, Trends Mol. Med. 9 (1) (2003) 5–19.

[15] C.A. Harrington, C. Rosenow, J. Retief, Monitoring gene expression using DNA microarrays, Curr. Opin. Microbiol. 3 (2000) 285–291.

[16] E. Hartuv, A. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrach, R. Shamir, An algorithm for clustering cDNA fingerprints, Genomics 66 (3) (2000) 249–256.

[17] I. Inza, P. Larranaga, R. Blanco, A.J. Cerrolaza, Filter versus wrapper gene selection approaches in DNA microarray domains, Artif. Intell. Med. 31 (2004) 91–103.

[18] J. Khan, et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat. Med. 7 (2001) 673–679.

[19] L. Li, C.R. Weinberg, T.A. Darden, L.G. Pedersen, Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method, Bioinformatics 17 (12) (2001) 1131–1142.

[20] J. Li, et al., Discovery of significant rules for classifying cancer diagnosis data, Bioinformatics 19 (2003) 93–102.

[21] R.J. Lipshutz, S.P.A. Fodor, T.R. Gingeras, D.J. Lockhart, High density synthetic oligonucleotide arrays, Nat. Genet. 21 (1999) 20–24.

[22] W. Lisss, Y. Yang, How many genes are needed for a discriminant microarray data analysis, Critical Assessment of Techniques for Microarray Data Mining Workshop, 2000.

[23] Y. Liu, X. Yao, Simultaneous training of negatively correlated neural networks in an ensemble, IEEE Trans. Systems, Man Cybern. 29 (6) (1999) 716–725.

[24] Liu, et al., A combinational feature selection and ensemble neural network method for classification of gene expression data, BMC Bioinform. 5 (1) (2004) 136–147.

[25] J.Y. Lo, et al., Computer-aided classification of breast microcalcification clusters: merging of features from image processing and radiologists, Proc. Int. Soc. Opt. Eng. (2003) 882–889.

[26] I.S. Lossos, A.A. Alizadeh, M.B. Eisen, W.C. Chan, P.O. Brown, D. Bostein, L.M. Staudt, R. Levy, Ongoing immunoglobulin somatic mutation in germinal center B cell-like but not in activated B cell-like diffuse large cell lymphomas, Proc. Natl. Acad. Sci. USA 97 (18) (2000) 10209–10213.

[27] F. Masulli, S. Rovetta, An ensemble approach to variable selection for classification of DNA microarray data, in: Proceedings of the International Joint Conference on Neural Networks, 2003, pp. 3089–3094.

[28] T.M. Mitchell, Machine Learning, McGraw-Hill, New York, 1997 (pp. 72–73).

[29] D.V. Nguyen, D.M. Rocke, Tumor classification by partial least squares using microarray gene expression data, Bioinformatics 18 (1) (2002) 39–50.

[30] D.A. Rew, DNA microarray technology in cancer research, Eur. J. Surg. Oncol. 27 (2001) 504–508.

[31] R. Shamir, R. Sharan, Algorithmic approaches to clustering gene expression data, in: T. Jiang, T. Smith, Y. Xu, M.Q. Zhang (Eds.), Current Topics in Computational Biology, MIT Press, Cambridge, MA, 2001.

[32] P. Tamayo, Interpreting patterns of gene expression with self-organizing map: Methods and application to hematopoietic differentiation, Proc. Natl. Acad. Sci. USA 96 (1999) 2907–2912.

[33] A.C. Tan, D. Gilbert, Ensemble machine learning on gene expression data for cancer classification, Appl. Bioinform. 2 (2003) 75–83.

[34] A. Tsymbal, S. Puuronen, Ensemble feature selection with the simple Bayesian classification in medical diagnosis, in: Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems, 2002, pp. 225–230.

[35] X. Yao, Y. Liu, G. Lin, Ensemble learning via negative correlation, Neural Networks 12 (10) (1999) 1399–1404.

[36] Z.H. Zhou, J. Wu, W. Tang, Ensembling neural networks: many could be better than all, Artif. Intell. 137 (1-2) (2002) 239–263.

Kyung-Joong Kim (Student Member, IEEE) received the B.S. and M.S. degree in computer science from Yonsei University, Seoul, Korea, in 2000 and 2002, respectively. Since 2002, he has been a Ph.D. student in the Department of Computer Science, Yonsei University. His research interests include evolutionary neural network, robot control and agent architecture.

Sung-Bae Cho (Member, IEEE) received the B.S. degree in computer science from Yonsei University, Seoul, Korea, in 1988 and the M.S. and Ph.D. degrees in computer science from Korea Advanced Institute of Science and Technology (KAIST), Taejeon, Korea, in 1990 and 1993, respectively. From 1991 to 1993, he worked as a Member of the Research Staff at the Center for Artificial Intelligence Research at KAIST. From 1993 to 1995, he was an Invited Researcher of Human Information Processing Research Laboratories at ATR (Advanced Telecommunications Research) Institute, Kyoto, Japan. In 1998, he was a Visiting Scholar at University of New South Wales, Canberra, Australia. Since 1995, he has been a Professor in the Department of Computer Science, Yonsei University. His research interests include neural networks, pattern recognition, intelligent man–-machine interfaces, evolutionary computation and artificial life.

Dr. Cho is a Member of the Korea Information Science Society, INNS, the IEEE Computer Society and the IEEE Systems, Man and Cybernetics Society. He was awarded outstanding paper prizes from the IEEE Korea Section in 1989 and 1992, and another one from the Korea Information Science Society in 1990. In 1993, he also received the Richard E. Merwin prize from the IEEE Computer Society. In 1994, he was listed in Who's Who in Pattern Recognition from the International Association for Pattern Recognition and received the best paper awards at International Conference on Soft Computing in 1996 and 1998. In 1998, he received the best paper award at World Automation Congress. He was listed in Marquis Who's Who in Science and Engineering in 2000 and in Marquis Who's Who in the World in 2001.