Appl Intell (2007) 26:243–250
DOI 10.1007/s10489-006-0020-4

Cancer classification using ensemble of neural networks with multiple significant gene subsets

Sung-Bae Cho · Hong-Hee Won

Published online: 12 November 2006

Abstract Molecular level diagnostics based on microarray technologies can offer the methodology of precise, objective, and systematic cancer classification. Genome-wide expression patterns generally consist of thousands of genes. It is desirable to extract some significant genes for accurate diagnosis of cancer because not all genes are associated with a cancer. In this paper, we have used representative gene vectors that are highly discriminatory for cancer classes and extracted multiple significant gene subsets based on those representative vectors respectively. Also, an ensemble of neural networks learned from the multiple significant gene subsets is proposed to classify a sample into one of several cancer classes. The performance of the proposed method is systematically evaluated using three different cancer types: Leukemia, colon, and B-cell lymphoma.

Keywords DNA microarray · Gene expression data · Cancer classification · Ensemble classifier · Multiple significant gene subsets

1 Introduction

The expression levels of thousands of genes can be simultaneously measured under particular experimental environments and conditions due to the significant advancement of DNA microarray technology [1]. This technology makes it possible to understand life on the molecular level, and enables to generate large-scale gene expression data. It has led to many statistical and analytical challenges from the problems in biology because it produces large amount of gene data. We can analyze the gene information very rapidly and precisely by managing them at once using several statistical and machine learning methods [2].

Cancer classification in clinical practice relied on clinical and histopathological information may turn out to be 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 the problems of cancer classification using data mining methods, machine learning algorithms and statistical methods to efficiently analyze these data [3, 4].

A significant number of 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. In general, in order to obtain a good ensemble, we have to try increasing the accuracy of the individual networks as well as increasing the diversity among the individual networks. But this does not mean that the individual networks should be mutually independent. There is no method that could guarantee to generate an ensemble comprising fully independent individual networks [5]. Schapire's boosting [6] and Breiman's bagging [7] are popular for generating individual networks, and averaging and voting are prevailing methods to combine individual predictions of classifiers.

During the last decades, computer-aided cancer classification attracts much attention. Zhou et al. proposed an

S.-B. Cho (✉) · H.-H. Won
Dept. of Computer Science, Yonsei University, 134
Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
e-mail: sbcho@cs.yonsei.ac.kr

automatic pathological diagnosis procedure called neural ensemble-based detection and realized an early stage lung cancer diagnosis system [8]. Satos-Carcia et al. also proposed an ensemble model of artificial neural networks to predict cardio-respiratory morbidity after pulmonary resection for non-small cell lung cancer [9]. Abbass presented an evolutionary artificial neural network approach based on the pareto-differential evolution algorithm augmented with local search for the prediction of breast cancer [10]. Futschik *et al.* applied evolving fuzzy neural networks to classify cancer tissue that is illustrated on the case studies of leukemia and colon cancer [11].

There was extensive work on building ensembles with the help of mutually exclusive features to improve the performance [12]. They exploited the negative correlation of the features using the fact that Pearson's correlation coefficient was negatively correlated with Euclidean distance, and showed the improvement of the performance by combining classifiers learned with two sets of the features [12]. Here the mutually exclusive features were from the fact that two features became more similar as Euclidean distance was smaller and Pearson's correlation coefficient was larger. However, the results were not verified enough because only one specific dataset was used. It is necessary to analyze systematically the performance of classifiers using a variety of benchmark datasets.

In this paper, we propose an ensemble of neural networks learned from multiple significant gene subsets to classify a sample into several cancer classes. The proposed method is tested in three cancer types, and the usefulness of multiple significant gene subsets is systematically analyzed for the ensemble classifier. The rest of the paper is organized as follows. In Section 2, DNA microarray technology is briefly introduced. Section 3 describes the proposed neural network ensemble with multiple significant gene subsets. Some experimental results are presented in Section 4. Finally, in Section 5, conclusions are drawn and some of the future works are discussed.

2 DNA microarray

DNA arrays consist of a large number of DNA molecules spotted in a systemic order on a solid substrate. DNA arrays can be categorized as microarrays when the diameter of DNA spot is less than 250 microns. The arrays with the small solid substrate are also referred to as DNA chips. It is so powerful that we can analyze the information of at least hundreds of genes at once.

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, 13].

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

where Int(Cy5) and Int(Cy3) are the intensities of red and green colors. Since more than thousands of genes are put on the DNA microarray, it is helpful to investigate the genome-wide information in short time.

3 Ensemble of neural networks with multiple significant gene subsets

An ensemble of neural networks with multiple significant gene subsets is here proposed to identify the corresponding cancer classes. The framework of the proposed neural network ensemble with significant gene subsets is shown in Fig. 1. The basic idea is to develop several pairs of trained classifier with multiple significant gene subsets, and to classify a given input pattern by utilizing a combination method. Then it naturally raises the question of obtaining a consensus on the results of individual classifiers. The ensemble

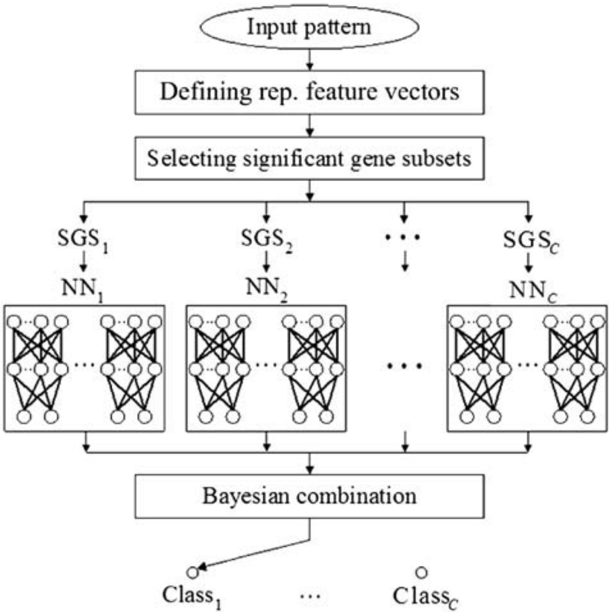


Fig. 1 The overview of the proposed ensemble classifier

Table 1 Mathematical formula for similarity measure of g_i and g_{rep_j} (for all formula, $\sum = \sum_{l=1}^{M}$)

$$PC(g_i,\ g_{\text{rep}\_j}) = \frac{M\sum g_i g_{\text{rep}\_j} - \sum g_i \sum g_{\text{rep}\_j}}{\sqrt{\left(M\sum g_i^2 - \left(\sum g_i\right)^2\right)\left(M\sum g_{\text{rep}\_j}^2 - \left(\sum g_{\text{rep}\_j}\right)^2\right)}}$$

$$SC(g_i,\ g_{\text{rep}\_j}) = 1 - \frac{6\sum (D_g - D_{g_{\text{rep}\_j}})^2}{M(M^2-1)}$$

(D_g and $D_{g_{\text{rep}\_j}}$ are the rank matrices of g_i and g_{rep_j})

$$ED(g_i,\ g_{\text{rep}\_j}) = \frac{1}{\sqrt{\sum (g_i - g_{\text{rep}\_j})^2}}$$

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

classifiers consist of two stages: Significant gene subsets selection based on representative vectors and ensemble classifier learned from significant gene subsets. Significant gene subsets selection based on representative vectors includes two steps: Defining the representative feature vectors that are highly discriminatory for the cancer classes, and selecting significant gene subsets based on the similarity with the representative feature vectors.

3.1 Multiple significant gene subsets

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

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

where e is the expression level and $i = 1, \ldots, N$. We want to know the locations of informative k features out of N.

Suppose there are different C classes of cancer and g_{rep_j} is a representative gene vector representing class of cancer $c_j (j = 1, \ldots, C)$. If it is possible to know representative vector g_{rep_j} for class of cancer c_j, we can simply measure the correlation and similarity of g_i to the classes, which tells the feature-goodness. We defined two representative gene patterns, g_{rep_j}, to obtain a standard of good features and utilized the features by scoring the respective similarity with each representative gene pattern [14]. We showed that two representative gene vectors are negatively correlated to represent two different aspects of classification boundaries. The first pattern is high in class A and low in class B, and the second one is low in class A and high in class B. The former is a binary vector which consists of 1 for all the samples in class A and 0 for all the samples in class B, while the latter is another binary vector which is composed of 0 for all the samples in class A and 1 for all the samples in class B. This method can be applied to multi-class problems as well as binary classification task.

Representative gene vector g_{rep_j} is a good arbitrator to distinguish between the class j and the other classes. SGS_j (Significant gene subset j) is defined as the gene set selected on the basis of the jth representative gene vector g_{rep_j}. SGS_j is defined as follows:

$$\text{SGS}_j = \left\{ g_i \,|\, \underset{1 \le i \le N}{\arg\max}\{Sim(g_i,\ g_{\text{rep}\_j})\} \right\} \tag{3}$$

where $Sim(X, Y)$ is the similarity of vectors X and Y. Correlation analysis and distance measure methods are used in order to measure the similarity of gene vector g_i and representative gene vectors. In this paper, we utilize four methods for $Sim(X,\ Y)$. Similarity measures are summarized in Table 1: Pearson correlation coefficient (*PC*), Spearman correlation coefficient (*SC*), Euclidean distance (*ED*), and cosine coefficient (*CC*) are listed in order. In case of *ED*, we define the similarity of vectors as an inverse of their distance.

To select significant gene subsets based on the similarity with the representative gene vector, we have to normalize each gene into the range of 0 to 1 as follows:

$$e_i = \frac{e_i - \min(e_i)}{\max(e_i) - \min(e_i)} \tag{4}$$

where $\min(e_i)$ is the minimum value of e_i and $\max(e_i)$ is the maximum value of e_i.

In this paper, informative genes are selected based on the representative ones by calculating the similarity of gene vector g_i and representative gene vectors. We have selected k, the number of genes in SGS, as 25. There is no report on the optimal number of genes to be selected, but our preliminary study indicates that 25–30 is appropriate [12].

3.2 Ensemble classifier

For classification, we have used 3-layered multi-layer perceptrons with 30 hidden nodes, 2 output nodes, 0.01–0.50 of learning rate and 0.9 of momentum. The neural networks are trained by error back propagation algorithm with the selected significant gene subsets. NN_j defines the neural network

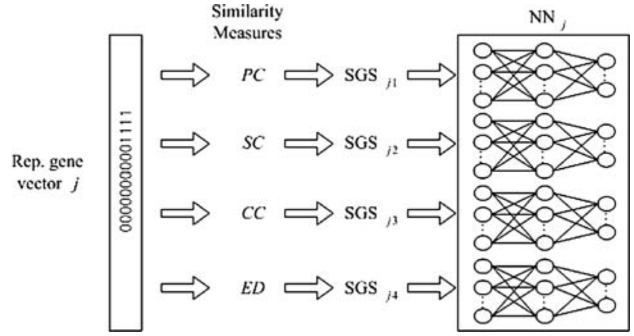


Fig. 2 The jth neural networks (NN_j) with four SGS_j's selected by four similarity measures

trained by the SGS_j as shown in Figs. 1 and 2. Since the multiple significant gene subsets represent different classes of cancer, the classifiers learned with those sets can be complementary. Combining the complementary classifiers can help increasing the performance of the classification [4].

Bayesian approach has been chosen among several alternatives such as voting, weighted voting, Bayesian approach and neural network for the final decision of the ensemble classifier. The Bayesian approach can overcome 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, the 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. When k classifiers are combined, c_i, $i = 1, \ldots, m$, is the class of a sample, $c(\mathrm{classifier}_j)$ is the class of the jth classifier, and w_i is a priori possibility of the class c_i computed from training data, the Bayesian combination is conducted as follows.

For training, we calculate the probabilities of each class given individual classifiers' prediction, which is expressed as $P(c_i \mid c(\mathrm{classifier}_j))$. For testing, each individual classifier provides their result classes for the test samples. After that, classifier decides the result class whose probability is the highest by calculating the following equation:

$$c_{\text{ensemble}} = \underset{1 \le i \le m}{\arg\max} \left\{ w_i \prod_{j=1}^{k} P(c_i | c(\mathrm{classifier}_j)) \right\} \tag{5}$$

We calculate w_i by defining it as the ratio of the number of i class samples to one of total samples. To apply the Bayesian approach to an ensemble method it would be better that each neural network is independent. Since the multiple significant gene subsets are selected based on the similarity of different representative vectors, they are fundamentally different from each other even though it does not guarantee the independency. Moreover, the outputs of the neural network are not just likelihoods or binary logical values near zero or one. Instead, they are estimates of Bayesian a posteriori probabilities of a classifier [15]. Note here that the outputs are not probabilities, but rather estimates of probabilities (hopefully good ones, but not necessarily so). Since the neural networks considered here were trained using a simple gradient technique, it is possible for the network to become stuck in local minima, or for the network's limited size to preclude it from computing an accurate estimate.

4 Experiments

4.1 Datasets

We have used three representative benchmark cancer datasets including Leukemia, Colon, and Lymphoma data. All datasets have two classes ($C = 2$) where one is cancer and the other is normal. 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 identified as training data and the remaining were identified as test data. Each sample contains 7129 gene expression levels (Available at: http://www.genome.wi.mit.edu/MPR). Colon dataset consists of 62 samples of colon epithelial cells taken from colon-cancer patients. Each sample contains 2000 gene expression levels. 40 of 62 samples are colon cancer samples and the remaining is normal samples. 31 out of 62 samples were identified as training data and the remaining was identified as test data (Available at: http://www.sph.uth.tmc.edu:8052/hgc/default.asp). Lymphoma dataset consists of 24 samples of GC B-like and 23 samples of activated B-like. 22 out of 47 samples were identified as training data and the remaining was identified as test data (Available at: http://genome-www.stanford.edu/lymphoma). For all datasets, dataset partition mentioned above was used for comparison experiments with other classifiers, and 10-fold cross validation was conducted for the other experiments.

4.2 Results analysis

Because Leukemia data have two different classes including AML and ALL, NN_1 defines the neural network trained with the SGS_1 representing AML class and NN_2 defines the neural network trained with the SGS_2 representing ALL class. Figure 3 shows the expression level of the significant gene subsets (SGS_1 and SGS_2), top-ranked genes in terms of the similarity with the representative vector genes, chosen by Pearson's correlation coefficient in Leukemia cancer dataset. They are the expression levels of genes chosen by Pearson's correlation coefficient using Representative Gene A and Representative Gene B respectively. 1–27 samples are of the class of ALL and 28–38 samples are of the class of AML.

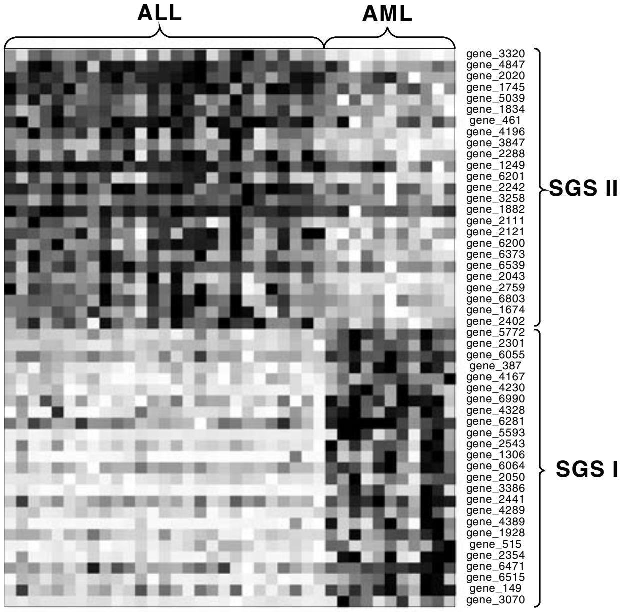


Fig. 3 The significant gene subsets (SGS_1 and SGS_2) chosen by PC in Leukemia dataset

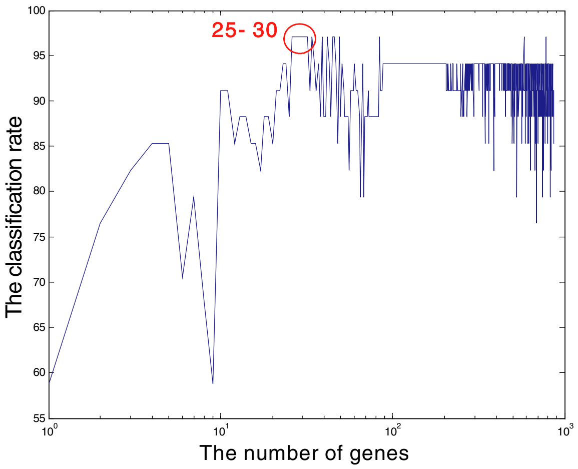


Fig. 4 Classification rates with respect to the number of genes in Leukemia dataset

The expression levels of 50 genes (SGS_1 and SGS_2) are clearly distinguishable between two classes as shown in Fig. 3. The 25 genes of them are under-expressed in ALL and over-expressed in AML. These include Leukotriene C4 synthase (LTC4S) gene, Zyxin, FAH Fumarylacetoacetate, and LYN V-yes-1 Yamaguchi sarcoma viral related oncogene homolog. In particular, Zyxin has been reported to be a good distinguishable feature in other works [15, 16]. In Fig. 3, the other 25 genes are over-expressed in ALL and under-expressed in AML. These include C-myb gene extracted from Human (c-myb) gene, complete primary cds, and five complete alternatively spliced cds, ACTN2 Actinin alpha 2, Hepatocyte growth factor-like protein gene and PTGIR Prostaglandin I2 (prostacyclin) receptor (IP). The expression levels of 25 genes (SGS_1) are more separable between ALL and AML than those of the other 25 genes (SGS_2). This result has been also come in another analysis (PCA 3-D plot and classification). But it is not always true that SGS_1 is more informative than SGS_2. While SGS_1 outperforms SGS_2 in Leukemia dataset and Colon dataset, SGS_2 is slightly superior to SGS_1 in Lymphoma dataset.

Figure 4 shows test classification rates with respect to the number of genes selected by Pearson's correlation coefficient on the basis of representative gene vector 1 in Leukemia dataset. The x-axis represents the logarithmic scale of the number of genes. The best performance is achieved in the feature subsets of 25–30 genes. In this range, the performance is stable while it is oscillating in other range. The classification rate of the feature subsets of 25–30 genes is 97.1%. The result gives a fair justification why we have chosen the most similar 25 genes for SGS.

Figure 5 shows principal component analysis (PCA) 3-D plot of ALL (full circles) and AML (void circles). (a) is the result of SGS_1 using PCA, (b) is the one of SGS_2 using PCA, and (c) is the one of SGS_1 and SGS_2 using PCA in Leukemia dataset. In Fig. 5(a), ALL samples are gathered together very closely in three-dimensional space. This is caused by the feature genes selected on the basis of SGS_1. In Fig. 5(b), AML samples are gathered together compared with ALL samples. On the other hand, ALL samples and AML samples are gathered together respectively in Fig. 5(c) because it uses simultaneously two informative gene subsets ($SGS_1 + SGS_2$) together.

Table 2 shows the performance of neural network classifier comparing with other ones in terms of the recognition rate. Column is the list of similarity measures: Pearson's correlation coefficient, Spearman's correlation coefficient, Euclidean distance and cosine coefficient. For all datasets, one of two NN classifiers provides the highest average performance though KNN shows the similar performance for colon dataset. Here NN_1 and NN_2 mean the neural network classifiers with SGS_1 and SGS_2, respectively.

Table 3 demonstrates the performance of single neural network classifiers. Here, ($NN_1 + NN_2$) means the one with the multiple significant gene subsets, ($SGS_1 + SGS_2$). The result of the classifier ($NN_1 + NN_2$) with the gene subsets, ($SGS_1 + SGS_2$), provides even higher performance in Colon and Lymphoma datasets, and it also shows the highest performance in Leukemia dataset.

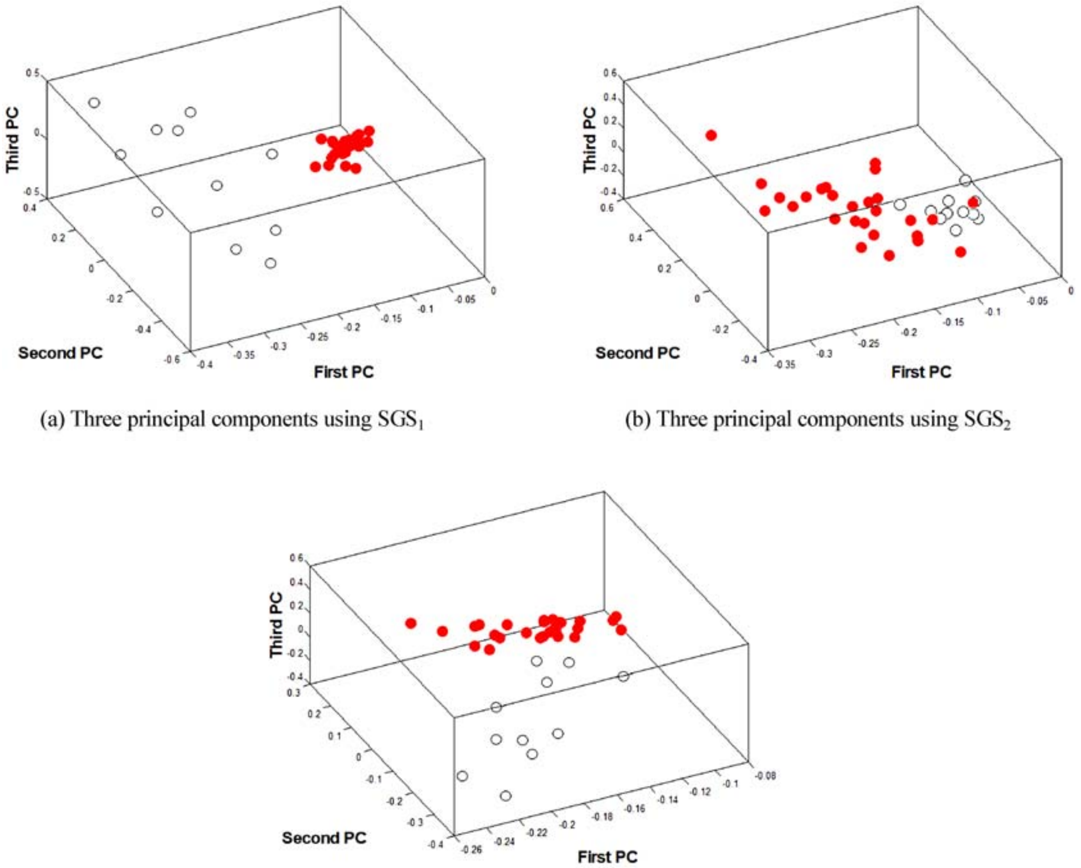
Table 4 summarizes the average performances by cross validation of all ensemble and single neural network classifiers with all different feature vectors. In both of ensemble and single neural net classifiers, the experiment using multiple significant gene subsets ($SGS_1 + SGS_2$) provides the

Table 2 Performance comparison with neural networks and other classifiers (%)

	Leukemia				Colon				Lymphoma			
	NN_1	NN_2	SVM	KNN	NN_1	NN_2	SVM	KNN	NN_1	NN_2	SVM	KNN
PC	**97.1**	79.4	79.4	94.1	74.2	77.4	64.5	71.0	64.0	72.0	60.0	76.0
SC	82.4	79.4	58.8	82.4	58.1	64.5	64.5	61.3	60.0	**88.0**	44.0	60.0
ED	91.2	61.8	70.6	82.4	67.8	77.4	64.5	**83.9**	56.0	72.0	56.0	68.0
CC	94.1	76.5	85.3	94.1	**83.9**	77.4	64.5	80.7	68.0	76.0	56.0	72.0
Avg.	**91.2**	74.3	73.5	88.3	71.0	**74.2**	64.5	**74.2**	62.0	**77.0**	54.0	69.0

Table 3 Performance comparison of single classifiers with different gene subsets (10-fold cross validation, %)

	Leukemia			Colon			Lymphoma		
	NN_1	NN_2	$NN_1 + NN_2$	NN_1	NN_2	$NN_1 + NN_2$	NN_1	NN_2	$NN_1 + NN_2$
PC	88.8 (±11.2)	94.3 (±7.4)	94.5 (±7.2)	73.5 (±21.4)	86.0 (±15.0)	87.6 (±15.0)	64.5 (±21.4)	85.0 (±14.1)	91.5 (±11.1)
SC	91.6 (±15.3)	92.9 (±7.5)	94.5 (±7.2)	82.6 (±20.8)	75.2 (±21.4)	87.6 (±15.5)	87.5 (±14.4)	67.0 (±24.1)	91.5 (±11.1)
ED	80.7 (±22.3)	91.8 (±7.1)	**95.7** (±6.9)	68.8 (±26.1)	84.3 (±16.7)	82.9 (±20.0)	72.5 (±18.6)	86.0 (±21.2)	**93.5** (±10.6)
CC	83.0 (±14.8)	93.2 (±9.3)	94.5 (±7.2)	72.1 (±24.3)	80.7 (±18.7)	**89.3** (±13.8)	70.0 (±17.5)	85.5 (±16.7)	91.5 (±11.1)
Avg.	86.0	93.0	**94.8**	74.3	81.6	**86.9**	73.6	80.9	**92.0**



(a) Three principal components using SGS_1

(b) Three principal components using SGS_2

(c) Three principal components using SGS_1 and SGS_2

Fig. 5 Principal component analysis (PCA) 3-D plot

Table 4 Performance comparison of all ensemble and single classifiers with different gene subsets (10-fold cross validation, %)

	Ensemble classifiers			Single classifiers		
	NN_1	NN_2	$NN_1 + NN_2$	NN_1	NN_2	$NN_1 + NN_2$
Leukemia	93.2 (±9.5)	94.5 (±6.9)	**95.9** (±6.4)	86.0 (±16.4)	93.0 (±7.6)	94.8 (±6.8)
Colon	75.0 (±19.1)	84.3 (±15.5)	**87.9** (±17.0)	74.3 (±22.9)	81.6 (±17.9)	86.9 (±15.8)
Lymphoma	87.5 (±18.8)	87.0 (±18.1)	**93.0** (±10.9)	73.6 (±19.4)	80.9 (±20.3)	92.0 (±10.5)

Table 5 Performance comparison of 25 genes (e_i) and 50 genes (e_i and $1 - e_i$) with colon dataset (10-fold cross validation, %)

	Ensemble classifiers		Single classifiers	
	NN_1	NN_2	NN_1	NN_2
25 genes	75.0 (±19.1)	84.3 (±15.5)	**74.3** (±22.9)	**81.6** (±17.9)
50 genes	**76.9** (±21.9)	**85.7** (±18.1)	72.1 (±10.0)	81.5 (±4.5)

best recognition rate, and ensemble classifier with ($SGS_1 + SGS_2$) outperforms single classifier with the same gene subsets.

Overall, neural network with the multiple significant gene subsets shows much better performance than that with only single significant gene subset, and ensemble classifiers provide a little better performance than single ones.

Finally, we have conducted an experiment by expanding the number of genes as 50, 25 in SGS_1 and SGS_2 (e_i) and 25 of their opposite expression values ($1 - e_i$) using colon dataset, and compared the result with that with 25 genes in SGS_1 and SGS_2, respectively, in order to check if there could be any correction if the presence of a gene can inhibit the manifestation of each cancer class. As summarized in Table 5, the result with 50 genes is nearly the same to that with 25 genes. Though it looks a little higher in ensemble classifiers, the difference is not significant.

4.3 Related works

Several machine learning techniques have been previously used for classifying gene expression data, including k nearest neighbor [17], decision tree, multi-layer perceptron, support vector machine [18], discriminant analysis [3, 19, 20], boosting, and self-organizing map [13]. Gene expression data have small number of samples with a number of features, so it seems that simple classifiers such as discriminant analysis, k nearest neighbor, and support vector machine among many classification methods outperform others. Choosing a feature selection method in classification is also very important as well as choosing a classifier because gene expression data are likely to be noisy.

Table 6 shows relevant works on cancer classification in three benchmark cancer datasets: Leukemia, Colon and Lymphoma datasets. Many researchers have been studying several problems of cancer classification [3, 4, 17–20] and clustering [13, 21–23] using these gene expression profiles and attempting to propose the optimal classification technique to work out these problems. Some produce better results than others as shown in Table 5. For Leukemia dataset, the proposed method produces the second best recognition rate, 95.9%, while other methods produce 91.6–95.8%. For Colon, the proposed method produces the result of 87.9% comparable to others 72.6–94.1%. For Lymphoma, the proposed method produces the best recognition rate of 93.0%, compared with others 84.6–98.1%. For all the datasets, the proposed method and the methods (discriminant analysis, principal component analysis and partial least square) in Nguyen et al. outperform the other methods [20]. Even though Nguyen's method produces higher performance than the proposed method, it exploits dimension reduction methods such as PCA and PLS. Since they transform the original features to new ones, the features extracted by them cannot be analyzed, which could be a serious weakness in bioinformatics that requires explainable results.

Table 6 Relevant works on cancer classification (%)

Author	Leukemia	Colon	Lymphoma
The proposed method	95.9	87.9	93.0
Furey et al. [18]	94.1	90.3	–
Li et al. 2000 [19]	94.1	–	–
Li et al. 2001 [17]	–	94.1∼	84.6∼
Ben-Dor et al. [4]	91.6–95.8	72.6–80.6	–
Dudoit et al. [3]	95.0∼	–	90.0∼
Nguyen et al. [20]	94.2–96.4	87.1–93.5	96.9–98.1

5 Concluding remarks

In order to classify the cancer class of patients, we have proposed a classification framework that combines a set of neural networks trained with multiple significant gene subsets, which are selected based on the representative gene vectors. The results clearly show that the suggested ensemble classifier works and we can improve the classification performance by combining neural networks learned from multiple significant gene subsets, even when we use simple combination methods of Bayesian approach.

The expression level of multiple significant gene subsets selected based on the representative gene vectors is clearly distinguishable between two classes. These patterns give enough information for the classification. The result of NN_1 is different from that of NN_2 on each dataset. While NN_1 outperforms NN_2 in Leukemia dataset, NN_2 outperforms NN_1 in Colon and Lymphoma datasets.

The experimental results show that the performance of the proposed ensemble classifier is superior to those of the individual neural networks in all benchmark datasets. Moreover, the neural network ensemble with multiple significant gene subsets outperforms the ensemble classifiers only with each significant gene subset. We have confirmed that multiple significant gene subsets enable the ensemble classifier to work better by providing salient features for the classification to neural network classifiers.

Obviously, the proposed method can be also applied to multi-class cancer classification problems. In the future work, we will have to conduct the same experiments with multi-class datasets to strengthen the results obtained and the experiments with the datasets of larger samples. Besides, we need to study on gene selection for extracting more representative genes. There have been only a few works that attempt to deal with the issue to model the number of genes. Since most of them are not systematic, a study on using expert information will be one of the important topics in this area. Finally, the correlation between genes and cancer classes should be investigated more in depth.

Acknowledgments This research was supported by the Ministry of Information and Communication, Korea under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment, IITA-2005-(C1090-0501-0019). The authors would like to thank the four anonymous reviewers for their constructive comments on the preliminary draft of this paper.

References

1. Harrington CA, Rosenow C, Retief J (2000) Monitoring gene expression using DNA microarrays. Curr Opin Microbiol 3:285–291
2. Eisen MB, Brown PO (1999) DNA arrays for analysis of gene expression. Methods Enzymbol 303:179–205
3. Dudoit S, Fridlyand J, Speed TP (2000) Comparison of discrimination methods for the classification of tumors using gene expression data, technical report 576, Department of Statistics, University of California, Berkeley
4. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini N (2000) Tissue classification with gene expression profiles. J Comput Biol 7:559–584
5. Zhou Z-H, Wu J, Tang W (2002) Ensembling neural networks: many could be better than all. Artif Intell 137(1–2):239–263
6. Scharpire RE (1990) The strength of weak learnability. Mach Learn 5(2):197–227
7. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
8. Zhou Z-H et al. (2002) Lung cancer cell identification based on artificial neural network ensembles. Artif Intell Med 24(1):25–36
9. Santos-Garcia G et al. (2004) Prediction of postoperative morbidity after lung resection using an artificial neural network ensemble Artificial Intelligence in Medicine
10. Abbass HA (2002) An evolutionary artificial neural networks approach for breast cancer diagnosis. Artif Intell Med 25:265–281
11. Futschik ME, Reeve A, Kasabov N (2003) Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue. Artif Intell Med 28:165–189
12. Cho S-B, Ryu J (2002) Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features. Proc IEEE 90(11):1744–1753
13. Shamir R, Sharan R (2001) Algorithmic approaches to clustering gene expression data. In: Jiang T, Smith T, Xu Y, Zhang MQ (eds), Current topics in computational biology. MIT Press
14. Golub TR, Slonim DK, Tamayo P, Huard C, GaasenBeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Blomfield CD, Lander ES (1999) Molecular classification of cancer: Class discovery and class prediction by gene-expression monitoring. Science 286:531–537
15. Richard MD, Lippmann RP (1991) Neural network classifiers estimate Bayesian a posteriori probabilities. Neural Comput 3:461–483
16. Li J, Wong L (2002) Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. Bioinformatics 18(5):725–734
17. Li L, Weinberg CR, Darden TA, Pedersen LG (2001) 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):1131–1142
18. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10):906–914
19. Li W, Yang Y (2000) How many genes are needed for a discriminant microarray data analysis, Critical Assessment of Techniques for Microarray Data Mining Workshop
20. Nguyen DV, Rocke DM (2002) Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18(1):39–50
21. Tamayo P (1999) Interpreting patterns of gene expression with self-organizing map: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 96:2907–2912
22. Hartuv E, Schmitt A, Lange J, Meier-Ewert S, Lehrach H, Shamir R (2000) An algorithm for clustering cDNA fingerprints. Genomics 66(3):249–256
23. Ben-Dor A, Shamir R, Yakhini Z (1999) Clustering gene expression patterns. J Comput Biol 6:281–297