Accurate Cancer Classification Using Expressions of Very Few Genes

Lipo Wang, Feng Chu, and Wei Xie

Abstract—We aim at finding the smallest set of genes that can ensure highly accurate classification of cancers from microarray data by using supervised machine learning algorithms. The significance of finding the minimum gene subsets is three-fold: 1) It greatly reduces the computational burden and "noise" arising from irrelevant genes. In the examples studied in this paper, finding the minimum gene subsets even allows for extraction of simple diagnostic rules which lead to accurate diagnosis without the need for any classifiers. 2) It simplifies gene expression tests to include only a very small number of genes rather than thousands of genes, which can bring down the cost for cancer testing significantly. 3) It calls for further investigation into the possible biological relationship between these small numbers of genes and cancer development and treatment. Our simple yet very effective method involves two steps. In the first step, we choose some important genes using a feature importance ranking scheme. In the second step, we test the classification capability of all simple combinations of those important genes by using a good classifier. For three "small" and "simple" data sets with two, three, and four cancer (sub)types, our approach obtained very high accuracy with only two or three genes. For a "large" and "complex" data set with 14 cancer types, we divided the whole problem into a group of binary classification problems and applied the 2-step approach to each of these binary classification problems. Through this "divide-and-conquer" approach, we obtained accuracy comparable to previously reported results *but with only 28 genes rather than 16,063 genes*. In general, our method can significantly reduce the number of genes required for highly reliable diagnosis.

Index Terms—Cancer classification, gene expression, fuzzy, neural networks, support vector machines.

1 Introduction

Compared with traditional tumor diagnostic methods based mainly on the morphological appearance of the tumor, the method using gene expression profiles is more objective, accurate, and reliable [1]. With the help of gene expression obtained from microarray technology, heterogeneous cancers can be classified into appropriate subtypes. Recently, different kinds of machine learning and statistical methods, such as artificial neural network [2], evolutionary algorithm [3], and nearest shrunken centroids [4], have been used to analyze gene expression data.

Supervised machine learning can be used for cancer prediction as follows: First, a classifier is trained with a part of the samples in the cancer data set. Second, one uses the trained classifier to predict the samples in the rest of the data set to evaluate the effectiveness of the classifier. The challenge of this problem lies in the following two points:

- In a typical gene expression data set, there are only very few (usually from several to several tens) samples of each type of cancers. That is, the training data are scarce.
- A typical gene expression data set usually contains expression data of a large number of genes, say, several thousand. In other words, the data are high dimensional.

In 2003, Tibshirani et al. successfully classified the lymphoma data set [5] with only 48 genes by using a statistical method called nearest shrunken centroids with an accuracy of 100 percent [6]. For the SRBCT data, Khan et al. classified all 20 testing samples with 96 genes [2]. They used a two-layered linear neural network. In 2002, Tibshirani et al. applied nearest shrunken centriods to the SRBCT data set [4]. They obtained 100 percent accuracy with 43 genes. For the method of nearest shrunken centroids, it categorizes each sample to the class whose centriod is nearest to the sample. The difference between standard nearest centroids and nearest shrunken centroids is that the latter uses only some important genes rather than all the genes to calculate the centroids. In 2003, Deutsch reduced the number of genes required to correctly classify the four cancer subtypes in the SRBCT data set to 12 genes [3]. In the same year, Lee and Lee also obtained 100 percent accuracy in this data set with an SVM classifier and the separability-based gene importance ranking [7], [8]. They used at least 20 genes to obtain this result. At the same time, they generated three principal components (PCs) from the 20 top genes. Their SVM also obtained 100 percent accuracy in the space defined by these three principal components. For the liver cancer data set, Chen et al. used 3,180 genes (represented by 3,964 cDNA) to classify HCC and the nontumor samples [9].

In [10], Ambroise and McLachlan indicated that testing results could be overoptimistic, caused by the "selection bias," if the testing samples were not excluded from the gene selection process. In fact, taking advantage of testing samples in any step of the classifier-building process, e.g.,

- *L. Wang and F. Chu are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Block S1, 50 Nanyang Avenue, Singapore 639798.*
E-mail: elpwang@ntu.edu.sg, chufeng@pmail.ntu.edu.sg.
- *W. Xie is with the Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613. E-mail: wxie@i2r.a-star.edu.sg.*

Manuscript received 4 Sept. 2005; revised 6 Jan. 2006; accepted 24 Feb. 2006; published online 9 Jan. 2007.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0099-0905.
Digital Object Indentifier No. 10.1109/TCBB.2007.1006.

1545-5963/07/$25.00 © 2007 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

feature selection, parameter tuning, model selection, etc., will induce bias. Therefore, to honestly evaluate a classifier in a given data set, the testing samples must be totally excluded from the classifier building process. According to this criterion, almost all of the above reported results are overoptimistic because all of the above classifiers more or less used the information of the testing samples in their training process.

In this paper, we propose a simple yet very effective method that leads to accurate cancer classification using expressions of only a very few genes. Furthermore, we evaluated our methods in an honest way, which excluded the influence of the bias [10].

This paper is organized as follows: We first introduce our procedure to find the minimum gene combinations. Then, the numerical results of four data sets demonstrate the effectiveness of our approach. In the final part, we discuss the results and related findings.

2 Method

Our proposed method is comprised of two steps. In Step 1, we rank all genes in the training data set using a scoring scheme. Then, we retain the genes with high scores. In Step 2, we test the classification capability of all simple combinations among the genes selected in Step 1 using a good classifier. We note that this paper proposes neither new feature ranking measures nor new classifiers. We shall describe two mechanisms for each of Step 1 and Step 2.

2.1 Step 1: Gene Importance Ranking

In Step 1, we compute the importance ranking of each gene using a feature ranking measure, two of which are described below. We then retain only the most important genes for Step 2.

2.1.1 T-Test

The t-score (TS) [4], [11] of gene i is defined as follows:

$$TS_i = max\left\{\left|\frac{\overline{x}_{ik} - \overline{x}_i}{m_k s_i}\right|, k = 1, 2, \dots K\right\}, \quad (1)$$

where

$$\overline{x}_{ik} = \sum_{j \in C_k} \overline{x}_{ij}/n_k, \quad (2)$$

$$\overline{x}_i = \sum_{j=1}^{n} x_{ij}/n, \quad (3)$$

$$s_i^2 = \frac{1}{n-K}\sum_k \sum_{j \in C_k} (x_{ij} - \overline{x}_{ik})^2, \quad (4)$$

$$m_k = \sqrt{\frac{1}{n_k} - \frac{1}{n}}. \quad (5)$$

There are K classes. $max\{y_k, k = 1, 2, \dots K\}$ is the maximum of all y_k. C_k refers to class k that includes n_k samples. x_{ij} is the expression value of gene i in sample j. $\overline{x}_{ik}$ is the mean expression value in class k for gene i. n is the total number of samples. $\overline{x}_i$ is the general mean expression value for gene i. s_i is the pooled within-class standard deviation for gene i. In fact, the TS used here is a $t\text{-statistic}$ between the centroid of a specific class and the overall centroid of all the classes [4], [11]. Another possible model for TS could be a $t\text{-statistic}$ between the the centroid of a specific class and the centroid of all the other classes.

2.1.2 Class Separability

Another frequently used method for gene importance ranking is the class separability (CS) [8]. The CS of gene i is defined as:

$$CS_i = SB_i/SW_i, \quad (6)$$

where

$$SB_i = \sum_{k=1}^{K} (\overline{x}_{ik} - \overline{x}_i)^2, \quad (7)$$

$$SW_i = \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \overline{x}_{ik})^2. \quad (8)$$

For gene i, SB_i is the sum of squares of the interclass distances (the distances between samples of different classes). SW_i is the sum of squares of the intraclass distances (the distances of samples within the same class). A larger CS indicates a greater ratio of the interclass distance to the intraclass distance and, therefore, can be used to measure the capability of genes to separate different classes.

In fact, the CS used here is very similar to the F-statistic that is also widely used for ranking genes in literature (see, e.g., [12], [13]). The difference between the CS and the F-statistic F is:

$$CS = F \cdot (K-1)/\left(\sum_{k=1}^{K} n_k - K\right). \quad (9)$$

Because the term $(K-1)/(\sum_{k=1}^{K} n_k - K)$ in (9) is a constant for a specific data set, the CS can be regarded as a simplification of F-statistic. The two methods will lead to the same ranking results for the same data set.

2.2 Step 2: Finding the Minimum Gene Subset

After selecting some top genes in the importance ranking list, we attempt to classify the data set with only one gene. We input each selected gene into our classifier. If no good accuracy is obtained, we go on classifying the data set with all the possible 2-gene combinations within the selected genes. If still no good accuracy is obtained, we repeat this procedure with all of the 3-gene combinations and so on until we obtain a good accuracy. In this paper, we used the following two classifiers to test gene combinations.

2.2.1 Fuzzy Neural Network (FNN)

We apply an FNN (Fig. 1) [14], [15] which we proposed earlier. This FNN combines the features of initial fuzzy model self-generation, partition validation, parameter optimization, and rule-base simplification.

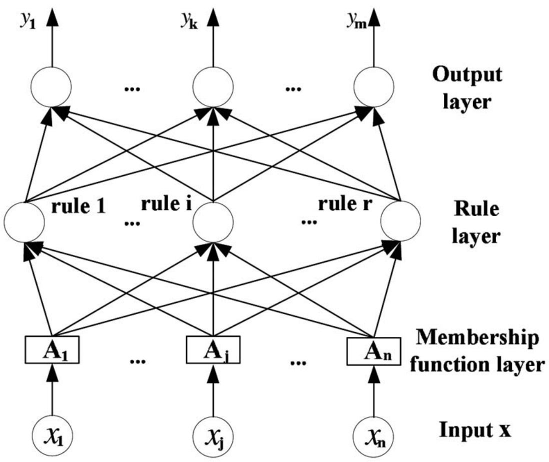


Fig. 1. The structure of our FNN proposed in [14], [15]. The FNN consists of four layers, i.e., the input layer, the input membership function layer, the rule layer, and the output layer. The input membership function layer converts numerical values to categorical values. Each rule node is connected to all input membership function nodes and output nodes for this rule. Each rule node performs a product of its inputs. The input membership functions act as fuzzy weights between the input layer and the rule layer. Links between the rule layer, the output layer, and the input membership functions are adjusted during the learning process. In the output layer, each node receives inputs from all the rule nodes connected to this output node and produces the actual output of the network.

2.2.2 Support Vector Machines (SVMs)

In addition to the FNN, we also used another classifier, i.e., the SVM. SVMs, pioneered by Vapnik [16], are able to find the optimal hyperplane that maximizes the boundaries between patterns. This feature makes SVM a powerful tool for pattern recognition tasks. In fact, SVMs have already been used in gene expression data analysis [7], [17]. In this work, we applied a group of C-SVMs with radial basis kernel functions [16].

For both the FNN and the SVMs, we carried out 5-fold cross-validation (CV) in the training data set to tune their parameters. We have included CV accuracy for all of the data sets. The FNN has one parameter, i.e., the learning rate (η) that needs to be tuned. At the beginning of the tuning process, we assign an initial value η_0 to η (e.g., 0.1). After that, we build up an FNN and then test the FNN using 5-fold CV in the training data set. We subsequently adjust η a little and build, then cross-validate the FNN again. We select the η value that leads to the smallest CV error. For the SVMs, the two parameters to be tuned, i.e., C and γ [16], are selected similarly.

3 Results

3.1 Lymphoma Data

In the lymphoma data set [5] (http://llmpp.nih.gov/lymphoma), there are 42 samples derived from diffuse large B-cell lymphoma (DLBCL), nine samples from follicular lymphoma (FL), and 11 samples from chronic lymphocytic leukaemia (CLL). The entire data set includes the expression data of 4,026 genes. In this data set, a small part of the data is missing. A k-nearest neighbor algorithm was applied to fill those missing values [18].

In the first step, we randomly divided the 62 samples into two parts: 31 samples for training, 31 samples for testing. We ranked the entire set of 4,026 genes according to their t-scores (TSs) in the training data set. Then, we picked out the 100 genes with the highest TSs (Table 8 at http://www.ntu.edu.sg/home5/pg02317674). In this paper, every gene is labeled after its importance rank. For example, gene 4 means the gene ranked fourth in Table 8 (http://www.ntu.edu.sg/home5/pg02317674). Through its ID in the microarray (for example, GENE1622X), the real name of each gene can be found on the Web page of the lymphoma data set.

We applied our FNN to classify the lymphoma microarray data set. At first, we added the selected 100 genes one by one to the network according to their TS ranks, starting with the gene ranked 1 in Table 8 (http://www.ntu.edu.sg/home5/pg02317674). That is, we first used only a single gene that is ranked 1 as the input to the network. We trained the network with the training data set and, subsequently, tested the network with the test data set. We repeated this process with the first two genes in Table 8 (http://www.ntu.edu.sg/home5/pg02317674), then three genes, and so on. We found that the FNN performed very well: It reached 100 percent 5-fold CV accuracy for the training data with only the first eight genes in Table 8 (http://www.ntu.edu.sg/home5/pg02317674). And, its corresponding testing accuracy was 96.8 percent.

The excellent performance of our FNN motivated us to search for the smallest gene subsets that can ensure highly accurate classification for the entire data set. We first attempted to classify the data set using only one gene. We fed each of the 4,026 genes in the lymphoma data set into our FNN. The best 5-fold CV accuracy was 90.32 percent and the best testing accuracy was 80.65 percent. Second, we tested all possible combinations of two genes within the 100 genes in Table 8 (http://www.ntu.edu.sg/home5/pg02317674). Fig. 2 shows the CV procedure used here. This procedure totally excluded the testing samples from the classifier building process and, hence, excluded the influence of bias. Other CVs conducted in this paper also used a similar scheme. To our pleasant surprise, among all 4,950 such possible 2-gene combinations, the CV accuracy for the training data reached 100 percent for the FNN in 174 combinations. The corresponding testing accuracies for these combinations varied from 80.6 percent (six errors in 31 testing samples) to 100 percent. The average accuracy was 93.85 percent. Table 1 shows the detailed results.

Since only two genes were required for classification, it became possible for us to visualize the gene profiles with respect to the distinct lymphoma subtypes. Here, we picked out some combinations that also reached 100 percent testing accuracy for visualization. We take note that these combinations only represent some of the best performing combinations we found, which are used to visually present our findings and related discussions. However, the performance of the classifier should only be estimated by the averaged accuracy of all the combinations in Table 1, i.e., 93.85 percent.

Fig. 3 shows four plots of 2-gene combinations: (78, 52), (78, 8), (4, 87), (4, 40). In each of the four plots, the clusters of

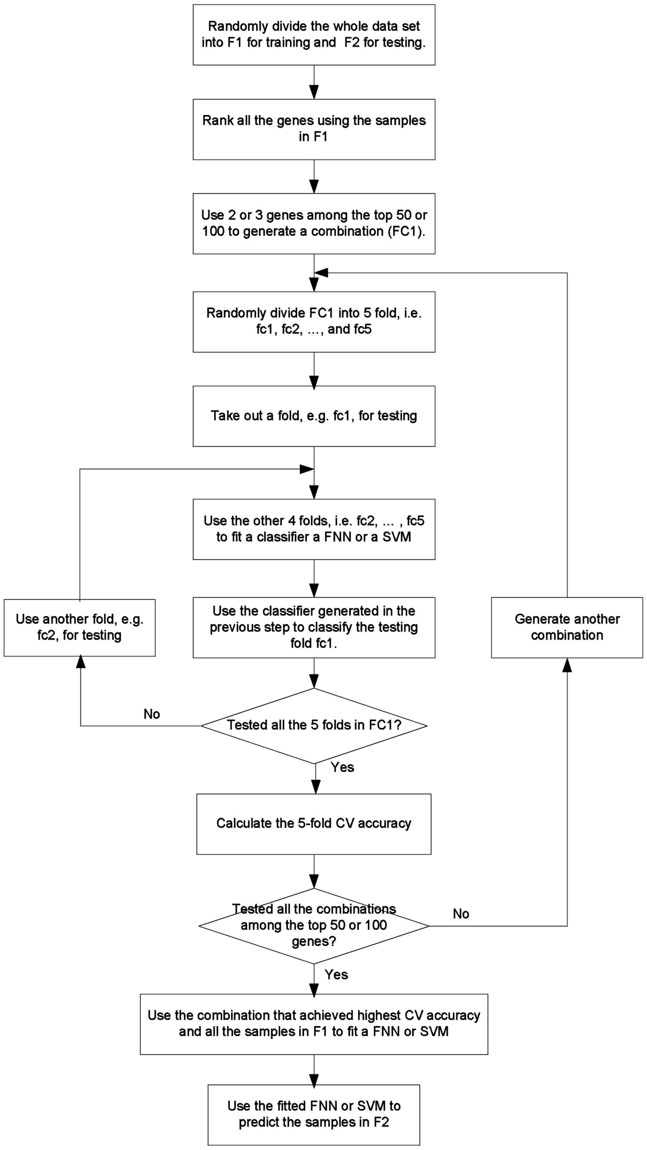

Randomly divide the whole data set into F1 for training and F2 for testing.
Rank all the genes using the samples in F1
Use 2 or 3 genes among the top 50 or 100 to generate a combination (FC1).
Randomly divide FC1 into 5 fold, i.e. fc1, fc2, ..., and fc5
Take out a fold, e.g. fc1, for testing
Use the other 4 folds, i.e. fc2, ... , fc5 to fit a classifier a FNN or a SVM
Use another fold, e.g. fc2, for testing
Use the classifier generated in the previous step to classify the testing fold fc1.
Generate another combination
No
Tested all the 5 folds in FC1?
Yes
Calculate the 5-fold CV accuracy
Tested all the combinations among the top 50 or 100 genes?
No
Yes
Use the combination that achieved highest CV accuracy and all the samples in F1 to fit a FNN or SVM
Use the fitted FNN or SVM to predict the samples in F2

Fig. 2. The procedure of the cross validation (CV) used in this paper. Here, the classifier building and the parameter tuning are carried out during the CV using only the training samples. The testing samples are totally independent from the CV; hence, the testing results obtained are unbiased.

DLBCL, CLL, FL are very clear and the boundaries can be easily drawn. Fig. 4 shows the corresponding expression profiles of the four combinations. We obtain the following simple prediction rules from Fig. 3d, which allow doctors to make an accurate diagnosis of the three subtypes of lymphoma. For a lymphoma patient:

1. the patient has DLBCL if and only if the expression level of gene 4 (i.e., GENE1622X) is greater than -0.75;
2. the patient has CLL if and only if the expression level of gene 40 (i.e., GENE540X) is less than -1;
3. the patient has FL otherwise, i.e., if and only if the expression level of gene 4 (i.e., GENE1622X) is less than -0.75 and the expression level of gene 40 (i.e., GENE540X) is greater than -1.

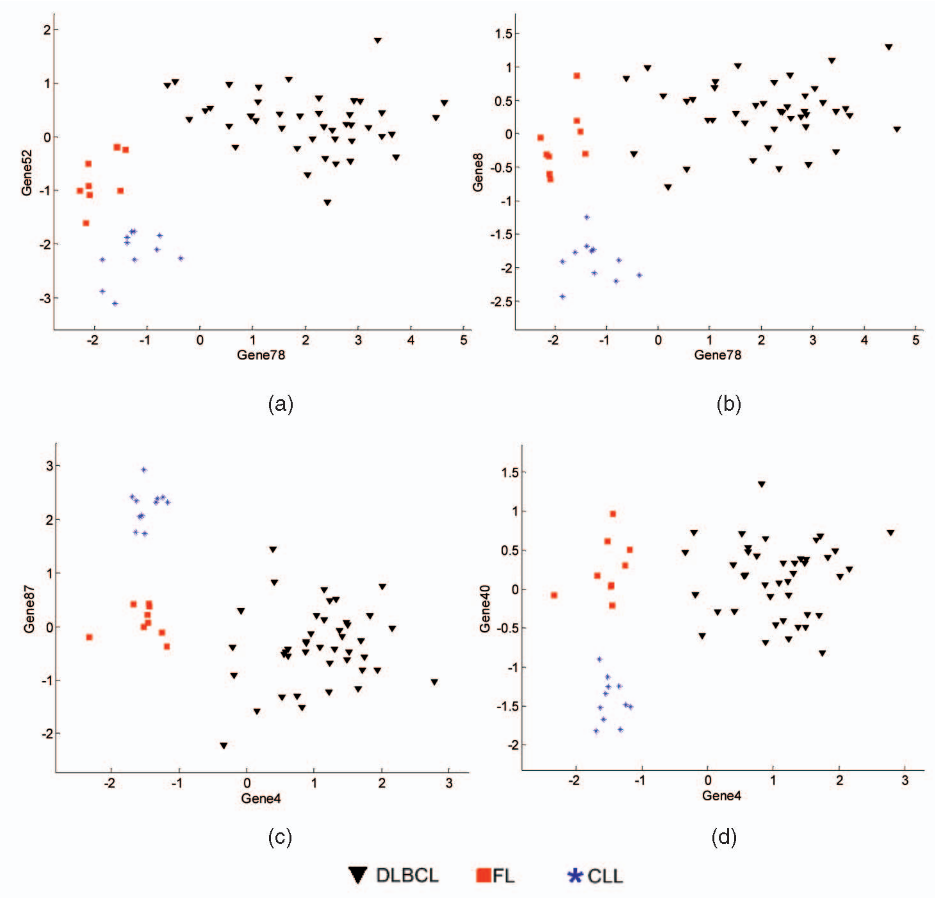
TABLE 1
Two-Gene Combinations in the Lymphoma Data Set that Lead to 100 Percent CV Accuracy in the Training Set and Their Respective Testing Accuracy for the FNN

Gene 1	Gene 2	5-fold CV error in Training	Number of Mistakes in Testing
1	62	0	3
1	78	0	1
2	4	0	1
2	24	0	1
2	37	0	1
2	78	0	0
2	83	0	1
2	89	0	2
2	92	0	1
3	24	0	4
3	31	0	4
3	37	0	3
3	78	0	1
3	89	0	3
4	5	0	1
…	…	…	…
…	…	…	…
78	84	0	1
78	86	0	2
78	88	0	2
78	94	0	2
78	99	0	2
79	92	0	3
84	92	0	4
86	92	0	3
87	92	0	1
92	97	0	6
Averaged Number of Errors			**1.908**
Averaged Accuracy			**93.85%**

The complete table can be obtained at http://www.ntu.edu.sg/home5/pg02317674.

In recent years, the method of honestly evaluating classifiers in the context of gene expression-based cancer classification has been discussed in the literature [10], [19], [20], [21], [22]. A *double cross validation* or *nested cross validation* scheme was applied in [20], [21], [22]. To examine our results more completely, we also applied the double CV to the lymphoma data set. Fig. 6 is the flow chart of the double CV algorithm that we used. Compared to the double CV in [20], the procedure we used is different in the following two aspects:

1. Besides the CV in the outer loop, we also conducted a CV in the inner loop to evaluate the classifier. However, in [20], Freyhult et al. only used one fold of the samples to test the classifiers in the inner loop. Considering the small number of samples in an inner fold, e.g., only five or six samples for the lymphoma data set, our scheme is more likely to find promising gene combinations.
2. Our method used all of the inner CV data for gene ranking and greatly reduced the computing time. Otherwise, if we wanted to totally exclude the



(a) (b)

(c) (d)

Fig. 3. Gene expression levels of 2-gene combinations that perfectly separate the three types of lymphoma subtypes, i.e., DLBCL, FL, and CLL: (a) (78, 52), (b) (78, 8), (c) (4, 87), and (d) (4, 40). Here, the genes are labeled according to their t-score (TS) ranks, for example, the two genes in (a) are ranked 78 and 52 according to their TSs. A total of 27 such 2-gene combinations are found by the FNN, which achieved 100 percent CV accuracy in the training process and also achieved 100 percent testing accuracy. These combinations are the best performing ones in Table 1. However, the performance of the classifier should only be estimated by the averaged accuracy of all the combinations in Table 1.

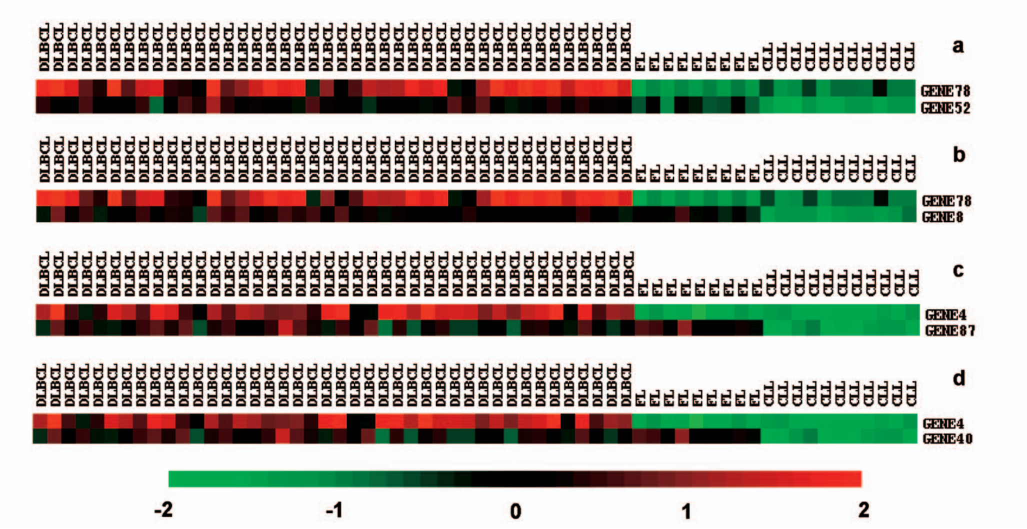


Fig. 4. The expression profiles of the 2-gene combinations in the lymphoma data set shown in Fig. 3: (a) (78, 52), (b) (78, 8), (c) (4, 87), and (d) (4, 40). Here, all of the genes are labeled according to their TS ranks. All four combinations can separate DLBCL, FL, and CLL. For example, in (d), gene 4 draws a boundary between DLBCL and other types; gene 40 draws a boundary between CLL and other types. Therefore, Gene 4 and Gene 40 jointly separate the three types.

testing samples from gene ranking in the inner CV, we had to rank genes and search for gene combinations for each fold. The computing time would be nearly 10 times that of our scheme for a 10-fold inner CV. Although our inner CV accuracy included the "selection bias" as indicated in [10], we found such

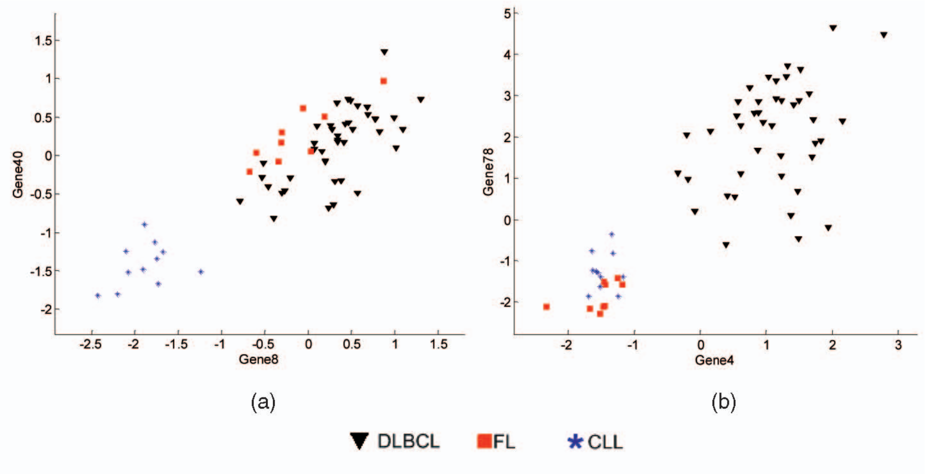


Fig. 5. Gene expression levels of 2-gene combinations that do not separate the three lymphoma subtypes: (a) (8, 40) and (b) (4, 78). Although these genes have high TS ranks, their combinations cannot separate the three subtypes because of poor cooperation.

inner results were accurate enough for picking out promising gene combinations, as shown in the following results. More importantly, such a scheme would never change the independence of the testing samples in the outer CV.

For the lymphoma data set, we used an SVM as the classifier and searched for gene combinations among the top 100 genes (TSs). The double CV achieved 93.55 percent accuracy, i.e., four errors in 62 samples. To facilitate easy verification of our double CV process and results, we have put all of the results on http://www.ntu.edu.sg/home5/pg02317674. File foldlym describes how the data set was randomly divided into 10 folds. It also includes the actual and the predicted labels for each sample. File lymDCV contains the winmax matrix for each fold. In addition, it also indicates the gene combination that was randomly picked out.

However, simply combining genes with high ranks does not ensure high classification accuracy. The cooperation of genes is also of great importance. To illustrate this point, we plot gene combinations (8, 40) and (4, 78) versus the cancer types in Fig. 5. Fig. 3c and Fig. 5b both show that gene 4 is very capable of classifying DLBCL because DLBCL can be picked out according to the X-axis (gene 4) value alone. Similarly, Figs. 3d and 5a show that gene 40 has good capacity to classify CLL. Through the cooperation of gene 4 and gene 40, DLBCL, FL, and CLL are totally classified (Fig. 3d). However, if gene 40 works with a gene that is not good at separating FL from CLL, for example, gene 8 (Fig. 5a), the cooperation between these two genes will be poor and the accuracy will be low (only 63.675 percent). The poor cooperation between gene 4 and gene 78 (both with high ranks) shown in Fig. 5b also substantiates the importance of cooperation between genes.

Furthermore, the expression profile in Fig. 4 can also support this idea of appropriate gene combinations. In Fig. 4, the expression value of gene 4 shows obvious difference between DLBCL and the other two types, i.e., FL and CLL. There is a clear boundary in the figure. This difference tells us that gene 4 is good at classifying DLBCL. Similarly, it can be found that gene 40 has a good capability of classifying CLL.

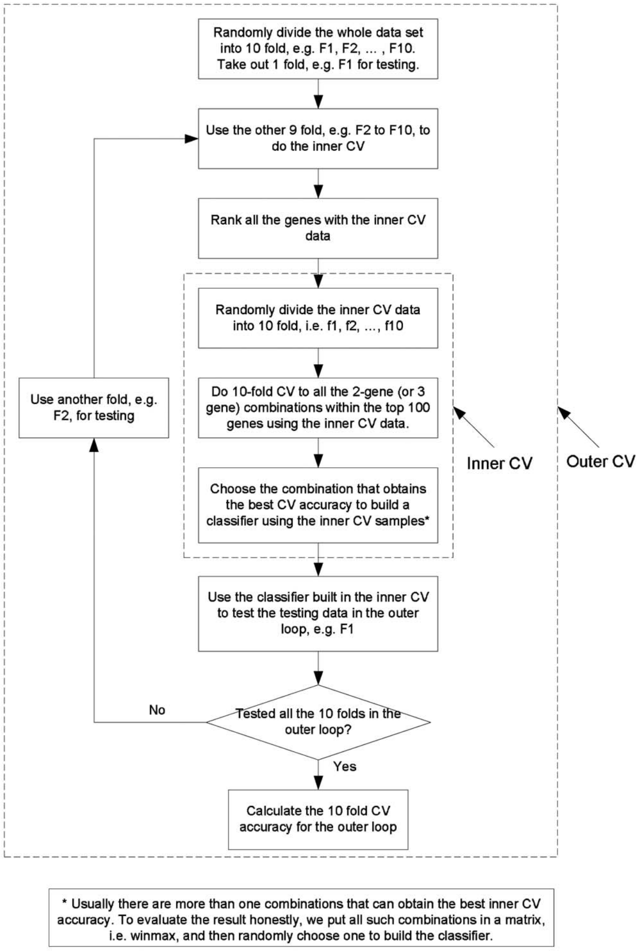
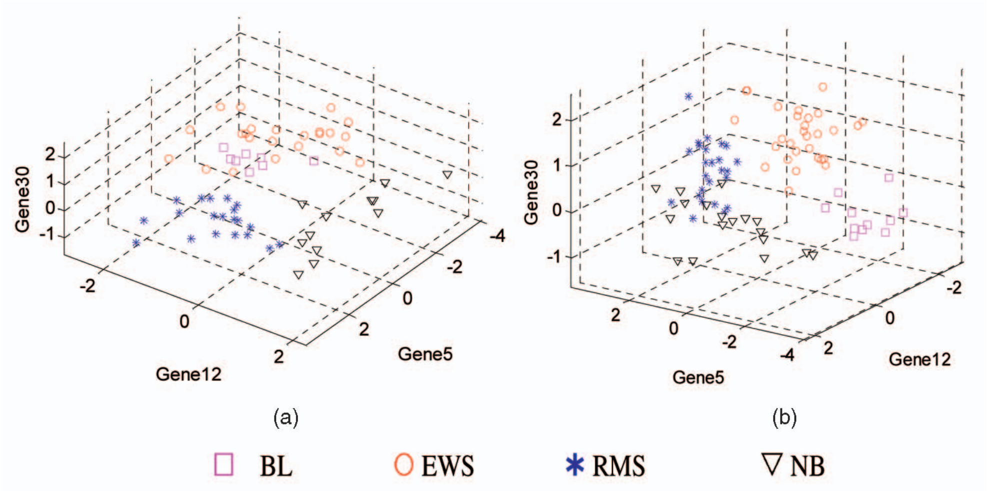

Randomly divide the whole data set into 10 fold, e.g. F1, F2, ... , F10. Take out 1 fold, e.g. F1 for testing.
Use the other 9 fold, e.g. F2 to F10, to do the inner CV
Rank all the genes with the inner CV data
Randomly divide the inner CV data into 10 fold, i.e. f1, f2, ..., f10
Do 10-fold CV to all the 2-gene (or 3 gene) combinations within the top 100 genes using the inner CV data.
Choose the combination that obtains the best CV accuracy to build a classifier using the inner CV samples*
Use the classifier built in the inner CV to test the testing data in the outer loop, e.g. F1
Tested all the 10 folds in the outer loop?
No
Yes
Use another fold, e.g. F2, for testing
Calculate the 10 fold CV accuracy for the outer loop
Inner CV
Outer CV
* Usually there are more than one combinations that can obtain the best inner CV accuracy. To evaluate the result honestly, we put all such combinations in a matrix, i.e. winmax, and then randomly choose one to build the classifier.

Fig. 6. The procedures of the double cross validation (CV). In the flow, there are two loops, i.e., the inner loop and the outer loop. In the inner loop, genes are first ranked according to their TSs. After that, we use all of the 2-gene combinations to build classifiers and do internal 10-fold CV. Then, we use the classifier that achieves the best CV accuracy in the internal loop to classify the testing samples in the outer loop. We repeat the process until all of the samples are cross validated in the outer loop. Because the testing samples in the outer loop are totally isolated from the gene ranking and classifier building process conducted in the internal loop, the CV results obtained in the outloop are unbiased.



(a)

(b)

Fig. 7. Three-dimensional views of gene expression levels of 3-gene combination (5, 12, 30) that separates the four SRBCT subtypes, i.e., BL, EWS, RMS, NB: (a) A view in which RMS and NB can be seen to be separated from other subtypes. (b) A view in which BL can be seen to be separated from EWS. From the two views, the four subtypes are clearly separable. Here, all of the genes are labeled according to their TS ranks. Our FNN found only one such 3-gene combination that achieved 100 percent 5-fold CV accuracy.

Gene 4 and gene 40 are able to jointly classify the three lymphoma types with 100 percent accuracy.

3.2 SRBCT Data

The SRBCT data set [2] (http://research.nhgri.nih.gov/microarray/Supplement) contains the expression data of 2,308 genes. There are, in total, 63 training samples and 25 testing samples already provided in [2]; five of the testing samples are not SRBCTs. The 63 training samples contain 23 Ewing family of tumors (EWS), 20 rhabdomyosarcoma (RMS), 12 neuroblastoma (NB), and eight Burkitt lymphomas (BL). And, the 20 SRBCT testing samples contain six EWS, five RMS, six NB, and three BL.

In the first step, we ranked the entire set of 2,308 genes according to their TSs [4], [11] in the training data set. Then, we picked out the 50 genes with the highest TSs as shown in Table 9 (http://www.ntu.edu.sg/home5/pg02317674). We applied our FNN as the classifier to the SRBCT microarray data set.

We used the expression data of the gene ranked 1 in Table 9 (http://www.ntu.edu.sg/home5/pg02317674) to train and then test the FNN. We repeated this process with the top two genes in Table 9 (http://www.ntu.edu.sg/home5/pg02317674), then the top three genes, and so on. The testing error decreased to 0 when the top 16 genes were input into the FNN.

To further reduce the number of genes required for accurate classification of the four cancer subtypes, we tested all possible combinations of one gene and two genes within the 50 selected genes. None of them can lead to 100 percent CV accuracy for the training data.

Then, we tested all possible combinations of three genes within the 50 selected genes. Among all 19,600 such 3-gene combinations, we found the 5-fold CV accuracy for the training data reached 100 percent with the combination of Gene 5, Gene 12, and Gene 30. The testing accuracy of this combination is 95 percent (one error for the 20 testing samples). Because this combination only contains three genes, it is still possible for us to visualize it. Fig. 7 shows two different views of the 3D plot of this 3-gene combination. In the view of Fig. 7a, RMS and NB are well separated from other types. In the view of Fig. 7b, BL and EWS are well separated with a clear boundary. From these two views, it is clear that the four SRBCT subtypes are separated from one another.

Except for (5, 12, 30), no other gene combinations obtained 100 percent CV accuracy for the training data. Although some combinations achieved 100 percent testing accuracy (e.g., the combination (7, 13, 18) in Fig. 8), they made a few errors for the training data. Such combinations should be excluded from the evaluation of the classifier. Otherwise, the result would be biased because it took advantage of the testing data.

In Table 2, we compared the numbers of genes required in different methods for the SRBCT data set. This comparison clearly points out that our method can significantly reduce the number genes required for accurate classification.

We did not randomly divide the SRBCT data for evaluation because the provider of the SRBCT data had already divided the data into training and testing sets [2] and this data partition has been used by other authors [3], [4], [7]. Hence, we also followed this partition for comparison.

3.3 Liver Cancer Data

The liver cancer data set [9] (http://genome-www.stanford.edu/hcc/) has two classes, i.e., the nontumor liver and HCC. The data set contains 156 samples and the expression data of 1,648 important genes. Among them, 82 are HCCs and the other 74 are nontumor livers. We randomly divided the data into 78 training samples and 78 testing samples. In this data set, there are some missing values. We also used the k-nearest neighbor method to fill those missing values [18].

In this data set, we followed the same steps as we did in the lymphoma and the SRBCT data sets. First, we chose 100 important genes in the training data set. Then, we tested all possible 1-gene and 2-gene combinations within the 100 important genes. Among all of the 5,050 combinations, we found that combination (1, 7) and (2, 82) reached

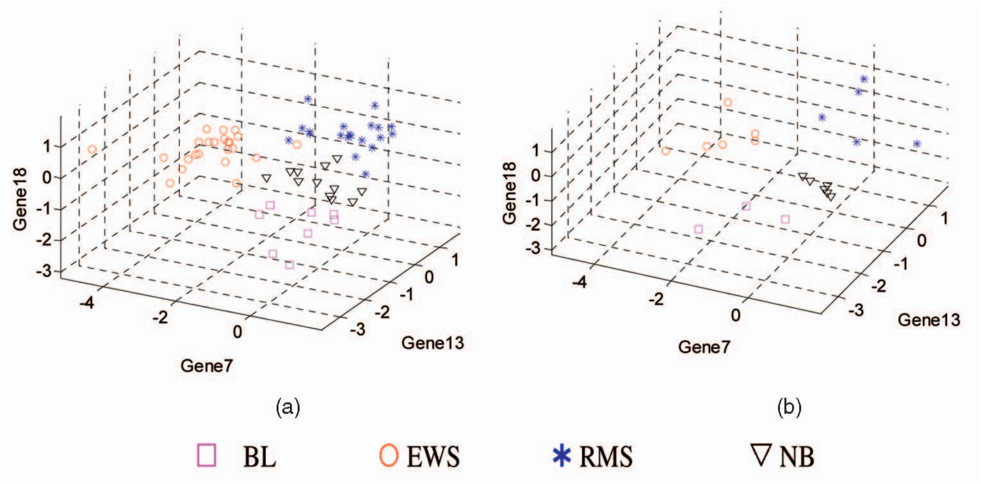


Fig. 8. Three-dimensional views of gene expression levels of the 3-gene combination (7, 13, 18) that separates the four SRBCT subtypes, i.e., BL, EWS, RMS, NB: (a) A view that includes only the training samples. (b) A view that includes only the testing samples. Although (b) is a perfect separation of the four subtypes, the combination (7, 13, 18) should be excluded from the evaluation of the classifier because it made one error for the training data and had a lower 5-fold CV accuracy than the combination (5, 12, 30).

100 percent CV accuracy for the training data. Their corresponding test accuracies are 100 percent and 96.2 percent, respectively. The averaged accuracy is 98.1 percent, which can be used to estimate the performance of the classifier.

The clone IDs of genes 1 and 7 are IMAGE: 301122 and IMAGE: 128461, respectively. We plot the gene expression levels of this combination in Fig. 9. From this plot, we found that the nontumor liver samples and HCC samples were separated quite well. Thus, we obtain the following simple prediction rules, which allow doctors to make an accurate diagnosis of HCC and nontumor liver:

1. the tissue is from a nontumor liver if the expression level of gene 1 (i.e., IMAGE: 301122) is greater than 0.345 and the expression level of gene 7 (i.e., IMAGE: 128461) is greater than -0.65;
2. otherwise, the tissue is from HCC.

We also conducted 10-fold double CV to the whole liver cancer data set. The double CV accuracy is 96.15 percent, i.e., six errors in 156 samples. The detailed results are given in files foldliver and liverDCV that can be obtained at http://www.ntu.edu.sg/home5/pg02317674.

For the liver cancer data set, Chen et al. [9] used 3,180 genes (represented by 3,964 cDNA) to classify HCC from the nontumor samples. In comparison with Chen et al.'s work [9], our method also greatly reduced the number of genes required to obtain an accurate result.

TABLE 2
Comparisons of Results for the SRBCT Data Set Obtained by Different Approaches

Method	Accuracy	Number of genes required
MLP neural network [2]	100% (biased)	96
Nearest shrunken centroids [4]	100% (biased)	43
Evolutionary algorithm [3]	100% (biased)	12
SVM [7]	100% (biased)	20 or 3 PCs
Our FNN	95% (unbiased)	3

Here, the results obtained in [2], [3], [4], and [7] are all biased, whereas our results for the FNN are unbiased.

3.4 GCM Data

The above three data sets, i.e., the lymphoma, the SRBCT, and the liver cancer data, are relatively small data sets. They include three, four, and two (sub)types of cancers, respectively. In such small data sets, exhaustive searches for 2-gene or 3-gene subsets are possible because they do not require much computing time. Let us suppose that one gene subset needs 0.5 seconds to process. Then, it will cost 22.46 hours to search for all of the possible 3-gene subsets from the 100 genes obtained by gene ranking and selection. However, if we search for all possible 5-gene subsets in the same data, it will cost 435.7 days. Therefore, exhaustive search is only effective for searching subsets with a small number of genes.

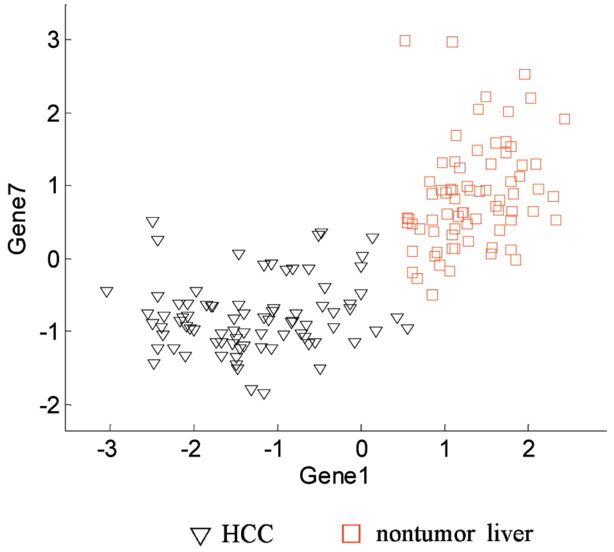


Fig. 9. Expression levels of Gene 1 and Gene 7 in the liver cancer data set. Here, all of the genes are labeled according to their TS ranks. The HCCs and the nontumor livers are clearly separable from the plot of (1, 7).

TABLE 3
General Information for the GCM Data Set

Type Number	Cancer Name	Abbr.	Number of Training Samples	Number of Testing Samples
1	Breast	B	8	3
2	Prostate	P	8	2
3	Lung	L	8	3
4	Colorectal	CR	8	3
5	Lymphoma	Ly	16	6
6	Bladder	BL	8	3
7	Melanoma	M	8	2
8	Uterus	U	8	2
9	Leukemia	Le	24	6
10	Renal	R	8	3
11	Pancreas	PA	8	3
12	Ovary	Ov	8	3
13	Mesothelioma	Ms	8	3
14	Brain	C	16	4

The cancer types included the number of training and testing samples for each type.

However, this limitation does not mean that exhaustive search cannot deal with large and complex data sets. In fact, for data sets with more classes, we can divide the whole classification problem into a group of binary classification problems with "one-against-one" or "one-against-all" schemes [23]. For each binary classification problem, we then use the proposed 2-step exhaustive search. Using this "divide-and-conquer" method, we can greatly reduce the computation time required.

To test the effectiveness of our method for more complex problems, we applied it to the GCM data set that includes 14 types of cancers [24] (http://www-genome.wi.mit.edu/mpr/GCM.html). This data set contains expression data of 16,306 genes and the total of 198 samples has already been divided into two parts, i.e., 144 for training and the other 54 for testing. Table 3 gives the general information of the data set.

We used the "one-against-all" scheme [23] to process the GCM data set. That is, we built 14 binary SVM classifiers and each binary SVM classifier output a probability that a sample belonged to a cancer type. Finally, a sample was categorized to the type of the largest probability. For each binary SVM classifier, we ranked all of the genes according to their TSs and then tested all of the 2-gene combinations within the top 50 genes. After that, we picked out the combination that achieved the best 5-fold CV accuracy in the training data for prediction. For each of the binary SVM classifiers, usually more than one combination achieved the best accuracy. To honestly evaluate the SVM classifier, we randomly chose a combination from these candidates and then used the 14 binary classifier to jointly cross-validate and test the GCM data. We repeated this process 100 times. The best 5-fold CV accuracy we obtained for the training data was 83.3 percent (24 errors in the 144 samples). Seven groups of binary SVMs achieved such CV accuracy; their testing accuracies varied from 64.8 percent to 77.8 percent. The averaged testing accuracy was 69.8 percent, which could be used to evaluate the performance of the group of SVMs.

Since all 14 of the binary SVMs used only two genes, we plotted the 14 2-gene combination that jointly achieved 77.8 percent testing accuracy in Figs. 10 and 11. Table 4 summarizes the genes used by each binary SVM and the best 5-fold CV accuracy it achieved. The confusion matrices that describe the training and testing results are shown in Table 5 and Table 6, respectively. However, we should note that these results only represent the best performing combinations. The general performance should be estimated by the averaged testing accuracy, i.e., 69.8 percent.

For the GCM data set, the best CV accuracy reported is 81.25 percent (27 errors in 144 samples) and the best testing accuracy reported is 77.8 percent (12 errors in 54 samples) with all 16,063 genes [24]. In [24], Ramaswamy et al. claimed that the testing samples were independent. However, it was not confirmed that the classifier building process was totally independent of the testing samples. For example, it is unknown how they decided the number of genes being used by the classifier. In [25], Li et al. obtained nearly 70 percent testing accuracy using all 16,063 genes. From their presentation, it was not clear whether their testing scheme was honest. In [26], Tan et al. honestly tested the GCM data and obtained 67.39 percent testing accuracy with 134 genes. Compared with these results, our 14 2-gene binary SVMs jointly reached comparable accuracy with only 28 genes.

We did not randomly divide the GCM data set for evaluation because the provider of the GCM data had already divided the data into training and testing sets [24]. Hence, we also followed this partition for comparison.

3.5 Results with Class Separability Importance Ranking

To find out how the gene importance ranking scheme influences the classification result, we also used another ranking scheme, i.e., the class separability (CS) [8]. In the SRBCT data set, we tested the combinations within the top 100 genes selected by the new ranking scheme. We also found the only 3-gene combination that leads to 100 percent accuracy with TS. However, the three genes rank (23, 40, 22) in CS rather than (5, 12, 30) in TS ranking [4], [11]. In the liver cancer data set, we also found the only 2-gene (ranked (1, 7) in TS but (1, 16) in CS) combination with 100 percent accuracy. In the lymphoma data set, we found 216 2-gene combinations within the top 100 genes of the new rank that reached 100 percent CV accuracy. These 216 combinations included most of the combinations found with the TS [4], [11] importance ranking.

4 Discussion

For our purpose of finding the smallest gene subsets for accurate cancer classifications, both TS and CS are highly effective ranking schemes, whereas both SVM and FNN are sufficiently good classifiers. Many other gene importance ranking schemes and classifiers may also be used in our approach.

As we have known from the results in the lymphoma data set, the gene combination that gives good separation may not be unique. Furthermore, it is found that highly correlated genes may undertake very similar tasks in a combination. Therefore, if a gene in a combination is replaced by another highly correlated gene, the newly

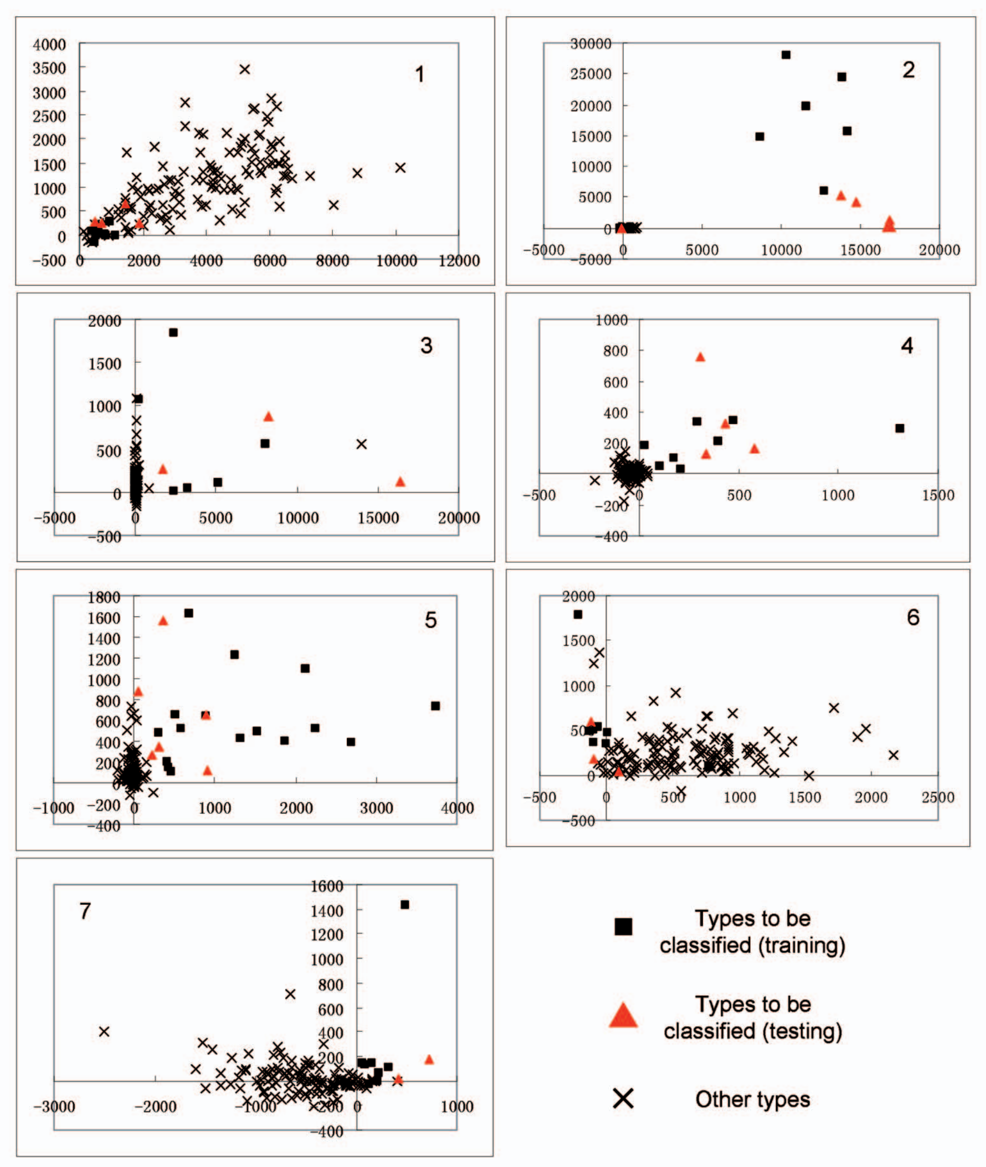


Fig. 10. Plots of gene combinations for binary classifiers 1-7. In each plot, the horizontal axis stands for gene 1 in Table 4 and the vertical axis stands for gene 2 in Table 4 for the corresponding classifier. We used 14 such binary classifiers, each responsible for classifying one of the 14 classes from all the rest. As shown in Figs. 10 and 11, we found a 2-gene combination for each of the 14 classifiers to best separate a given class from the rest. We then combined the 14 binary classifiers into one 14-class classifier, which leads to a slightly better cross-validation accuracy and the same testing accuracy compared to the best reported result, but with only 28 genes rather than all 16,063 genes required by the existing approach.

obtained combination may still lead to good classification. Clustering important genes selected in Step 1 will help us discover such important correlations.

In the lymphoma data set, we clustered the 100 selected genes into 20 groups using the K-means method [27]. The clustering result is summarized in Table 7. As we mentioned above, (4, 40) is a good combination for the lymphoma data set. We also note that genes 71 and 36 are in the same clusters with genes 4 and 40, respectively, and the two clusters are very small. Therefore, we replaced genes 4 and 40 with genes 71 and 36, respectively (another result using hierarchical clustering is available at http://www.ntu.edu.sg/home5/pg02317674, which also confirmed the high correlation between genes 4, 40 and genes 71, 36, respectively). From the plot of (71, 36) (Fig. 12), we find it is also a good combination.

This work is the first attempt to test the classification ability of gene combinations in similar applications. Following this idea, we solved this problem with much smaller numbers of genes compared with the previously published methods. In fact, after finding the optimal gene combinations, this problem becomes a comparably easy pattern recognition problem, which can be seen from the plots of gene combinations.

In the SRBCT data set, the three genes, 5, 12, 30, are the genes IGF2, AF1q, and CD99, respectively. It has been reported that CD99 is a marker for EWS [28], [29]. In addition, IGF2 is also reported to be indispensable for the formation of medulloblastoma and RMS [30], [31]. As for the genes where no relations between them and the related cancers were reported, our work calls for the further investigation of such relations.

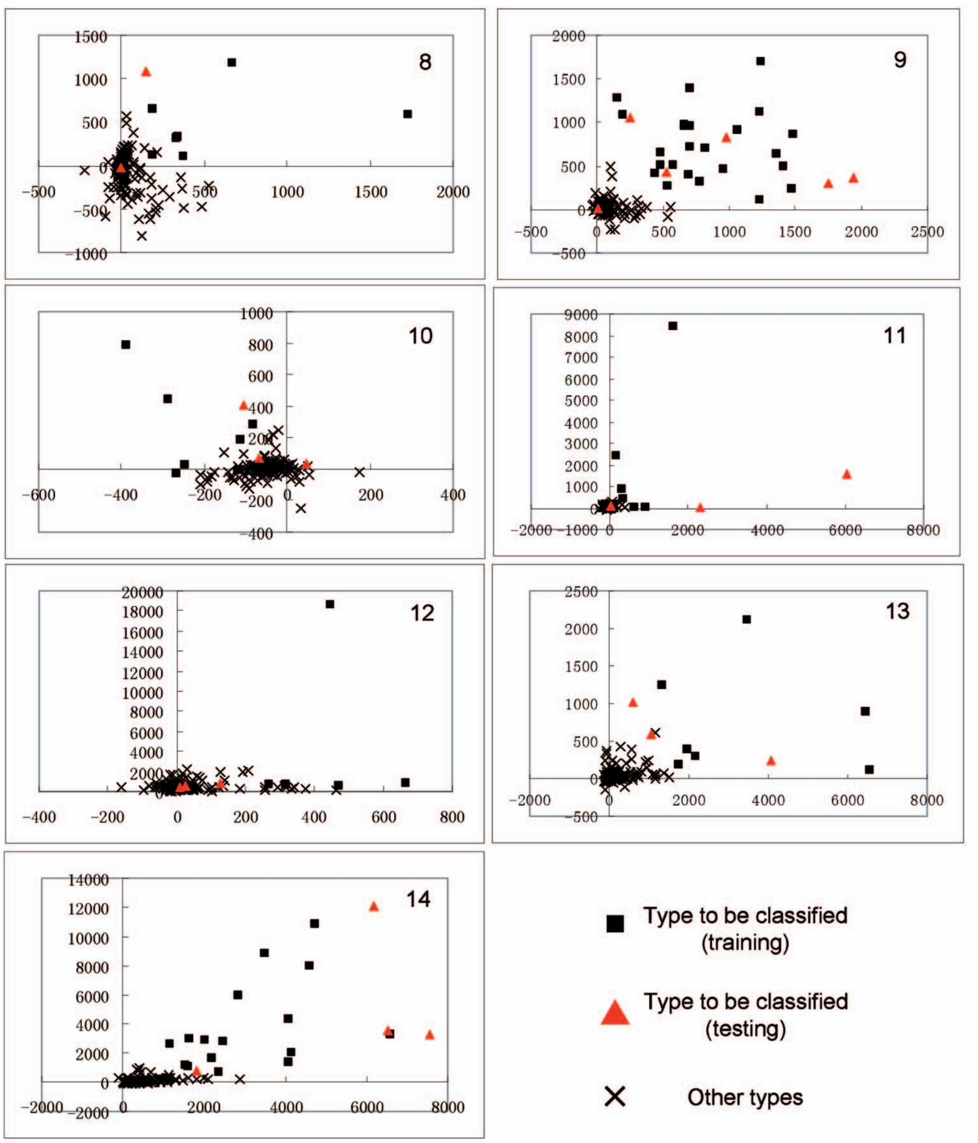


Fig. 11. The same as Fig. 10, for binary classifiers 8-14.

TABLE 4
The Best 5-Fold Cross-Validation Accuracy and the Genes Used by the 14 Binary Classifiers

Classifier Number	Cancer Type Concerned	Gene 1 (Accession)	Gene 2 (Accession)	Best 5-fold Cross Validation Accuracy
1	Breast	RCAA397825_at	RC_AA195229_s_at	97.22%
2	Prostate	X07730_at	RC_AA176975_s_at	99.31%
3	Lung	M24461_at	X02419_rna1_s_at	97.92%
4	Colorectal	X83228_at	U51095_at	99.31%
5	Lymphoma	M27394_s_at	RC_AA233620_at	100%
6	Bladder	RC_AA135095_at	M92449_at	97.92%
7	Melanoma	RC_AA176812_at	U65092_at	98.61%
8	Uterus	M61906_at	U19718_at	98.61%
9	Leukemia	S72024_s_at	RC_AA410338_at	100%
10	Renal	U18088_s_at	M64082_at	98.61%
11	Pancreas	U31449_at	J04040_at	98.61%
12	Ovary	RC_AA456055_at	X54667_s_at	97.22%
13	Mesothelioma	RC_AA419609_at	M62402_at	99.31%
14	Brain	AA093923_at	N76496_at	100%

TABLE 5
The Confusion Matrix for the 5-Fold Cross-Validation Results of the GCM Data Set

	Predicted Type														
	B	P	L	CR	Ly	BL	M	U	Le	R	PA	Ov	Ms	C	n
B	**6**						1					1			8
P	1	**6**	1												8
L	1		**6**									1			8
CR				**8**											8
Ly					**16**										16
BL						**6**	1	1							8
M						1	**6**					1			8
U	1							**6**		1					8
Le									**24**						24
R			2						1	**5**					8
PA						1					**6**	1			8
Ov						1		1		1		**5**			8
Ms							1						**7**		8
C						1								**15**	16

Rows delineate the actual types and columns delineate the predicted types.

TABLE 6
The Confusion Matrix for the Testing Results of the GCM Data Set

	Predicted Type														
	B	P	L	CR	Ly	BL	M	U	Le	R	PA	Ov	Ms	C	n
B	**3**							1							4
P		**5**	1												6
L			**4**												4
CR				**4**											4
Ly					**6**										6
BL	2					**1**									3
M							**2**								2
U	1							**1**							2
Le	1				1				**4**						6
R	1						1			**1**					3
PA						1					**2**				3
Ov						2						**2**			4
Ms													**3**		3
C														**4**	4

Rows delineate the actual types and columns delineate the predicted types.

To sum up, we proposed a simple yet very effective 2-step method for gene selection and gene expression data classification. We applied our method to four well-known

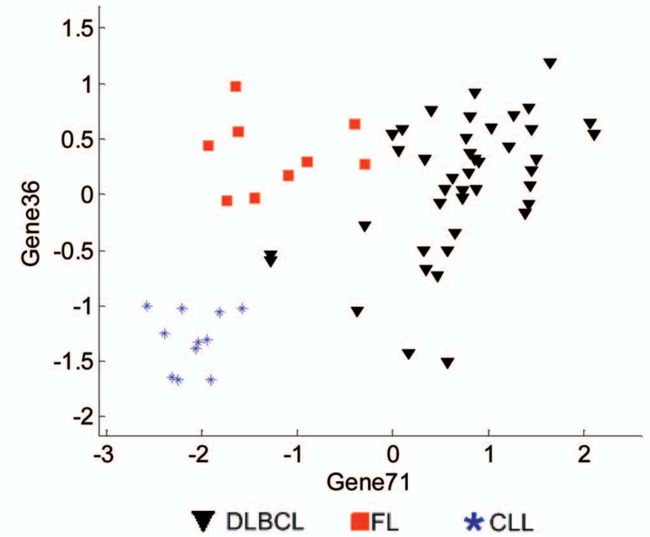


Fig. 12. Expression levels of gene 71 and gene 36 in the lymphoma data set. Here, all of the genes are labeled according to their TS ranks. Genes 71 and 36 are highly correlated with genes 4 and 40, respectively. The combination (4, 40) separates the three lymphoma subtypes very well, as shown in Fig. 3b. This plot shows that its correlated combination (71, 36) also separates the three subtypes well.

TABLE 7
The K-Means Clustering Result for the 100 Selected Genes in the Lymphoma Data Set: The 20 Clusters and the Genes Included in Each Cluster

Cluster Number	Total Number of Genes in the Cluster	Genes Included in the Cluster
1	12	5, 10, 23, 26, 38, 39, 49, 50, 51, 52, 70, 77
2	2	85, 87
3	1	83
4	3	2, 58, 68
5	3	4, 71, 75
6	7	25, 33, 53, 56, 93, 95, 97
7	6	7, 8, 21, 32, 36, 40
8	3	24, 73, 80
9	3	5, 29, 35
10	2	89, 91
11	1	90
12	1	82
13	1	13
14	1	37
15	13	3, 15, 19, 22, 30, 41, 42, 44, 47, 57, 60, 65, 79
16	37	6, 11, 12, 15, 16, 17, 18, 20, 27, 28, 31, 34, 43, 45, 46, 48, 54, 59, 61, 62, 63, 64, 66, 67, 69, 72, 74, 76, 79, 84, 86, 88, 94, 96, 98, 99, 100
17	1	81
18	1	92
19	1	78
20	1	55

The genes in each cluster may be highly correlated.

microarray data sets, i.e., the lymphoma, the SRBCT, the liver cancer, and the GCM data sets. The results in all of the data sets indicate that our method can find minimum gene subsets that can ensure very high prediction accuracy. In addition, although the TS [4], [11] and the CS [8] based approaches have been proven to be effective in selecting important genes for reliable prediction, they are not perfect. To find minimum gene subsets that ensure accurate predictions, we must also consider the cooperation between genes.

5 Supplemental Information

The supplemental information, such as full gene lists, figures, and results that are not included in this paper, can be obtained at http://www.ntu.edu.sg/home5/pg02317674.

Acknowledgments

The authors would like to thank the associate editor and the three anonymous reviewers for their valuable comments which helped improve the paper significantly.

References

[1] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," *Science,* vol. 270, pp. 467-470, 1995.

[2] J.M. Khan et al., "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," *Nature Medicine,* vol. 7, pp. 673-679, 2001.

[3] J. Deutsch, "Evolutionary Algorithms for Finding Optimal Gene Sets in Microarray Prediction," *Bioinformatics,* vol. 19, pp. 45-52, 2003.

[4] R. Tibshirani, T. Hastie, B. Narashiman, and G. Chu, "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proc. Nat'l Academy of Sciences USA,* vol. 99, pp. 6567-6572, 2002.

[5] A.A. Alizadeh et al., "Distinct Types of Diffuse Large b-Cell Lymphoma Identified by Gene Expression Profiling," *Nature,* vol. 403, pp. 503-511, 2000.

[6] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu, "Class Predicition by Nearest Shrunken Centroids with Applications to DNA Microarrays," *Statistical Science,* vol. 18, pp. 104-117, 2003.

[7] Y. Lee and C.K. Lee, "Classification of Multiple Cancer Types by Mulitcategory Support Vector Machines Using Gene Expression Data," *Bioinformatics,* vol. 19, pp. 1132-1139, 2003.

[8] S. Dudoit, J. Fridlyand, and T.P. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *J. Am. Statistical Assoc.,* vol. 97, pp. 77-87, 2002.

[9] X. Chen et al., "Gene Expression Patterns in Human Liver Cancers," *Molecular Biology of the Cell,* vol. 13, pp. 1929-1939, 2002.

[10] C. Ambroise and G.J. McLachlan, "Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data," *Proc. Nat'l Academy of Sciences USA,* vol. 99, pp. 6562-6566, 2002.

[11] J. Devore and R. Peck, *Statistics: The Exploration and Analysis of Data,* third ed. Duxbury Press, 1997.

[12] Y. Lai, B. Wu, L. Chen, and H. Zhao, "Statistical Method for Identifying Differential Gene-Gene Coexpression Patterns," *Bioinformatics,* vol. 20, pp. 3146-3155, 2004.

[13] P. Broet, A. Lewin, S. Richardson, C. Dalmasso, and H. Magdelenat, "A Mixture Model-Based Strategy for Selecting Sets of Genes in Multiclass Response Microarray Experiments," *Bioinformatics,* vol. 20, pp. 2562-2571, 2004.

[14] Y. Frayman, L. Wang, and C. Wan, "Cold Rolling Mill Thickness Control Using the Cascade-Correlation Neural Network," *Control and Cybernetics,* vol. 31, pp. 327-342, 2002.

[15] Y. Frayman and L. Wang, "Data Mining Using Dynamically Constructed Fuzzy Neural Networks," *Lecture Notes in Artificial Intelligence,* vol. 1394, pp. 122-131, 1998.

[16] V. Vapnik, *Statistical Learning Theory.* Wiley, 1998.

[17] M.P. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Ares Jr., and D. Haussler, "Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines," *Proc. Nat'l Academy of Sciences USA,* vol. 97, pp. 262-267, 2000.

[18] O. Troyanskaya et al., "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics,* vol. 17, pp. 520-525, 2001.

[19] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J.A. Olson Jr., J.R. Marks, and J.R. Nevins, "Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles," *Proc. Nat'l Academy of Sciences USA,* vol. 98, pp. 11 462-11 467, 2001.

[20] E. Freyhult, P. Prusis, M. Lapinsh, J.E. Wikberg, V. Moulton, and M.G. Gustafsson, "Unbiased Descriptor and Parameter Selection Confirms the Potential of Proteochemometric Modelling," *BMC Bioinformatics,* vol. 6, no. 50, 2005.

[21] S. Dudoit, M.J.V.D. Laan, S. Keles, A.M. Molinaro, S.E. Sinisi, and S.L. Teng, "Loss-Based Estimation with Cross-Validation: Application to Microarray Data Analysis and Motif Finding," Univ of California Berkeley Division of Biostatistics Working Paper Series, no. 137, 2003, http://www.bepress.com/ucbbiostat/paper137.

[22] A. Barrier, M.J.V.D. Laan, and S. Dudoit, "Prognosis of Stage II Colon Cancer by Non-Neoplastic Mucosa Gene Expression Profiling," Univ. of California Berkeley Division of Biostatistics Working Paper Series, no. 179, 2003, http://www.bepress.com/ucbbiostat/paper179.

[23] C.C. Chang and C.J. Lin, "A Comparison of Methods for Muti-Class Support Vector Machines," *IEEE Trans. Neural Network,* vol. 13, pp. 415-425, 2002.

[24] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander, and T. Golub, "Multiclass Cancer Diagnosis Using Tumor Gene Expression Signature," *Proc. Nat'l Academy of Sciences USA,* vol. 98, pp. 15 149-15 154, 2000.

[25] T. Li, C. Zhang, and M. Ogihara, "A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Based on Gene Expression," *Bioinformatics,* vol. 20, pp. 2429-2437, 2004.

[26] A.C. Tan, D.Q. Naiman, L. Xu, R.L. Winslow, and D. Geman, "Simple Decision Rules for Classifying Human Cancers from Gene Expression Profiles," *Bioinformatics,* vol. 21, pp. 3896-3904, 2005.

[27] T. Mitchell, *Machine Learning.* McGraw-Hill, 1998.

[28] I.M. Ambros, P.F. Ambros, S. Strehl, H. Kovar, H. Gadner, and M. Salzer-Kuntschik, "MIC2 Is a Specific Marker for Ewing's Sarcoma and Peripheral Primitive Neuraoectodermal Tumor. Evidence for a Common Histogenesis of Ewing's Sarcoma and Peripheral Primitive Neuroectodermal Tumors from MIC2 Expression and Specific Chromosome Aberration," *Cancer,* vol. 67, pp. 1886-1893, 1991.

[29] H. Kovar et al., "Overexpression of the Pseudoautosomal Gene MIC2 in Ewing's Sarcoma and Peripheral Primitive Neuroectodermal Tumor," *Oncogene,* vol. 45, pp. 1067-1070, 1990.

[30] S. Zhan, D.N. Shapiro, and L.J. Helman, "Activation of an Imprinted Allele of the Insulin-Like Growth Factor II Gene Implicated in Rhabdomyosarcoma," *J. Clinical Investigation,* vol. 94, pp. 445-448, 1994.

[31] H. Hahn et al., "Pached Target Igf2 Is Indispensable for the Formation of Medulloblastoma and Rhabdomyosarcoma," *J. Biological Chemistry,* vol. 275, pp. 28 341-28 343, 2000.

Lipo Wang is the author or coauthor of more than 60 journal publications, 12 book chapters, and 90 conference presentations. His research interests include computational intelligence, with applications to data mining, bioinformatics, and optimization. He holds a US patent in neural networks. Dr. Wang has authored two monographs and edited 16 books. He was a keynote/panel speaker for several international conferences. Dr. Wang is an associate editor for the *IEEE Transactions on Neural Networks* (2002-present), *IEEE Transactions on Evolutionary Computation* (2003-present), and *IEEE Transactions on Knowledge and Data Engineering* (2005-present). He is an area editor of the *Soft Computing* journal (2002-present). He is/was an editorial board member of five additional international journals. Dr. Wang is the vice president-technical activities of the IEEE Computational Intelligence Society (2006-2007) and served as chair of the Emergent Technologies Technical Committee (2004-2005). He has been on the Governing Board of the Asia-Pacific Neural Network Assembly since 1999 and served as its president in 2002/2003. He was the founding chair of both the IEEE Engineering in Medicine and Biology Chapter Singapore and IEEE Computational Intelligence Chapter Singapore. Dr. Wang serves/served as general/program chair for 11 international conferences and as a member of the steering/advisory/organizing/program committees of more than 100 international conferences. He is a senior member of the IEEE.

Feng Chu received the BEng degree from Zhejiang University, Hangzhou, China, and the MEng degree from Huazhong University of Science and Technology, Wuhan, China, in 1995 and 2002, respectively. Since 2002, he has been pursuing the PhD degree at Nanyang Technological University in Singapore. Since 2005, he has also worked for Siemens Pte Ltd in Singapore as a research and development engineer. His research interests include computational intelligence, data mining, and their applications, e.g., bioinformatics, computational finance, etc. He is a member of the IEEE.

Wei Xie received the BEng degree in communication engineering in 2002 with first class honors and the MEng degree in information engineering in 2004 from Nanyang Technological University. He is now working at the Institute for Infocomm Research, Singapore. His research interests include biomedical signal processing, image processing, and pattern recognition. He is a student member of the IEEE.