BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 24 2012, pages 3306–3315
doi:10.1093/bioinformatics/bts602

Data and text mining

Advance Access publication October 11, 2012

An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data

Yongjun Piao[1], Minghao Piao[1], Kiejung Park[2] and Keun Ho Ryu[1,*]

[1]Department of Electrical and Computer Engineering, Chungbuk National University, Chungbuk, Korea and [2]Division of Bio-Medical informatics, Center for Genome Science, Korea National Institute of Health, Osong, South Korea

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Gene selection for cancer classification is one of the most important topics in the biomedical field. However, microarray data pose a severe challenge for computational techniques. We need dimension reduction techniques that identify a small set of genes to achieve better learning performance. From the perspective of machine learning, the selection of genes can be considered to be a feature selection problem that aims to find a small subset of features that has the most discriminative information for the target.

Results: In this article, we proposed an Ensemble Correlation-Based Gene Selection algorithm based on symmetrical uncertainty and Support Vector Machine. In our method, symmetrical uncertainty was used to analyze the relevance of the genes, the different starting points of the relevant subset were used to generate the gene subsets and the Support Vector Machine was used as an evaluation criterion of the wrapper. The efficiency and effectiveness of our method were demonstrated through comparisons with other feature selection techniques, and the results show that our method outperformed other methods published in the literature.

Availability: By request from the author.

Contact: pyz@dblab.chungbuk.ac.kr; khryu@dblab.cbnu.ac.kr

Received on May 29, 2012; revised on September 18, 2012; accepted on October 2, 2012

1 INTRODUCTION

Recently, there has been increasing interests in changing the emphasis of cancer classification from morphologic to molecular (Xiong *et al*., 2001). Cancers are usually marked by a change in the expression levels of certain genes; thus, the selection of relevant genes for cancer classification is an important task in most cancer gene expression studies. These discriminative genes are very useful in clinical applications, such as in recognizing disease profiles (Yang *et al*., 2006). However, microarray data pose a severe computational challenge because of its high dimensionality and small sample size (Saeys *et al*., 2007). From the perspective of machine learning, the selection of genes is a feature selection problem that aims to find a small subset of features with the most discriminative information for the target.

Feature selection is an important pre-processing step in eliminating irrelevant and redundant features for classification. The growing dimensionality of recorded data demands dimension reduction techniques that identify small sets of features that lead to better learning performance. The objective of feature selection is to provide faster and more effective models, and also to avoid overfitting and the curse of dimensionality. Feature selection methods can be broadly categorized into three types: filter, wrapper and hybrid (Pok *et al.,* 2010; Talavera, 2005). The filter methods use specific evaluation criteria that are independent of a learning algorithm to identify a feature subset from the original feature set. Filter techniques (Liu and Setiono, 1996; Liu *et al.,* 2002) are fast and scale easily to high-dimensional datasets, but they ignore interaction with the classifier. Wrapper methods (Kim *et al.,* 2000; Kohavi and John, 1997) use the classifier to evaluate the performance of each subset with a search algorithm. Wrapper methods tend to find the most suitable feature subset for the learning algorithm, but they are very computationally expensive. Hybrid methods (Kannan *et al.,* 2010; Xie and Wang, 2011) combine the advantages of filter and wrapper techniques. These algorithms aim to achieve the best learning performance with a predetermined learning algorithm and a similar time complexity to filter algorithms (Yu and Liu, 2003).

A feature selection procedure can usually be divided into two steps: subset generation and subset evaluation (Liu and Yu, 2005). The most important issue in generating a feature subset is how to choose the search strategy and the starting point. Complete search, sequential search and random search are the typical search methods used for subset generation. Complete search methods consider every feature subset to be a potential candidate to guarantee finding the optimal result. However, the computational time is intractable when the dimensionality is high. Sequential search methods, such as Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS), sacrifice completeness by applying the greedy hill-climbing approach (Han and Fu, 1996), which adds or removes features one at a time. These algorithms are computationally simpler and faster than a complete search strategy, but they can still lead to local optima. Random search methods, such as random-start hill-climbing and simulated annealing (Doak, 1992), start with a randomly selected subset, and these algorithms help to escape local optima in the search space.

During the past few years, the Support Vector Machine (SVM) has become very popular because of its good performance on high-dimensional data. SVM was developed by Vapnik (1995) to successfully solve the problems of handwritten digit recognition (Adankon and Cheriet, 2009), object recognition (Hanson and Halchenko, 2008), text classification (Zaghloul *et al*., 2009), cancer diagnosis (Akay, 2009) and bioinformatics (Zhang *et al*., 2009).

*To whom correspondence should be addressed.

In this article, we proposed a hybrid feature selection algorithm named Ensemble Correlation-Based Gene Selection (ECBGS) based on symmetrical uncertainty (SU) and SVM for gene selection. Our proposed method combined a filter approach and a wrapper method to remove the redundant features and to find the relevant features from the original feature set. For the original feature set, SU was used as an evaluation criterion for the filter, using the different starting points as a subset generation strategy and SVM as the evaluation learning algorithm of a wrapper. It was observed that the classifier combined with our proposed feature selection method obtained promising classification accuracy with a small gene subset on six gene expression datasets.

2 METHODS

The hybrid model (Deisy *et al*., 2007) attempts to combine the advantages of both filters and wrappers by exploiting their different evaluation criteria in different search strategies. Hybrid models have the advantage of including the interaction with a classification algorithm, while at the same time being far less computationally intensive than wrapper methods. Many hybrid feature selection algorithms have been proposed in the past few years. However, many search algorithms, such as SFS and SBS, ignore feature redundancy during the feature subset generation procedure. Along with irrelevant features, redundant features also affect the speed and accuracy of the classifiers (Yu and Liu, 2003). Furthermore, many hybrid methods are still computationally expensive. Based on these observations, the proposed feature selection method uses SU as the evaluation measure in the filter step to select relevant genes. To use the SU value, all the features need to be discretized. Then, different genes are used as the starting point to generate multiple gene subsets, and the generated subsets are evaluated by the SVM. Finally, we use the best gene subset to train the SVM model. The outline of the classification procedure with ECBGS is shown in Figure 1.

2.1 Fast correlation-based filter

There are many filter approaches that have been developed, such as the chi-squared test, mutual information, Pearson correlation coefficients (Li *et al*., 2011), Information gain, the Gain ratio and Relief (Gheyas and Smith, 2010). These methods are fast but lack robustness against interactions among features. Yu and Liu (2004) proposed a Fast Correlation-Based Filter (FCBF) approach to remove the redundant and irrelevant features. SU was used to measure the correlation:

$$IG(X|Y) = H(X) - H(X|Y) \tag{1}$$

$$SU(X, Y) = 2^{*}IG(X|Y)/(H(X) + H(Y)) \tag{2}$$

where $IG(X|Y)$ is the information gain of X after observing variable Y. $H(X)$ and $H(Y)$ are the entropy of variables X and Y. Using SU as the correlation measure, the feature selection procedure can be done by considering the C-correlation and F-correlation.

Definition 1 (C-correlation): The correlation between any feature F_i and the class C is called C-correlation, denoted by $SU_{i,c}$.

Definition 2 (F-correlation): The correlation between any pair of features F_i and F_j ($i \neq j$) is called F-correlation, denoted by $SU_{i,j}$.

FCBF removes irrelevant features by ranking C-correlation. Redundant features could be defined using predominant features and the approximate Markov Blanket. A feature is predominant if it does not have any approximate Markov Blanket in the current set. For two relevant features, F_i and F_j ($i \neq j$), F_j forms an approximate Markov Blanket for F_i if

$$SU_{j,c} \geq SU_{i,c} \text{ and } SU_{i,j} \geq SU_{i,c} \tag{3}$$

where $SU_{i,c}$ is the correlation between the feature and the class, and $SU_{i,j}$ is the correlation between feature i and feature j.

2.2 ECBGS

In general, FCBF identifies a single feature subset for which the discriminative capability is limited for classification purposes (Liu *et al*., 2010). To obtain multiple feature subsets, we use different starting points to remove redundant features in the search procedure, which allow each subset to have its own information and to avoid being trapped in local optima. We call the algorithm ECBGS, and the pseudocode of our algorithm is shown in Figure 2. As in Figure 2, given a dataset D that contains N features (F_1, F_2, ..., F_N) and a class C, the algorithm seeks several feature subsets, whereas each partial set does not include any redundant features. In the first part (lines 2–7), all of the features are sorted in descending order according to the $SU_{i,c}$ value, which is the correlation between the i-th feature and the class. A feature with a higher SU value indicates higher discrimination of this feature compared with other categories, and means that the feature contains useful information for classification. After calculating the SU values for all of the features, a threshold for the results is established. If the SU value of a feature is higher than the threshold, the feature is selected; if not, the feature is not selected. D_{rel} is the selected relevant subset of the original features. In the second part (lines 9–32), a number of feature subsets are derived by splitting the redundant features in the relevant feature subset into several parts [Subset(1), Subset(2), ..., Subset(i)]. During the redundancy analysis, if we remove the redundant features as FCBF does, the selected feature subset cannot guarantee the best prediction for the classification problem. This is because the features that are highly correlated with the class are also highly correlated with each other such that the removed feature subset in FCBF can lead to a more accurate result. Therefore, in our method, we choose the first element of the D_{rem} as a starting point for redundancy analysis (line 13) and repeat the procedure until there are no features in D_{rem}. D_{rem} is the subset of features that are removed in each redundancy analysis step. The details of the redundancy analysis step (lines 19–29) are presented in Yu and Liu (2004). At the first iteration, because there are no features in the D_{rem}, we choose the most relevant feature as the starting point (lines 10–11). After generating a number of feature subsets, each subset is evaluated by SVM, and the subset with the best classification accuracy will be selected for the final input of the

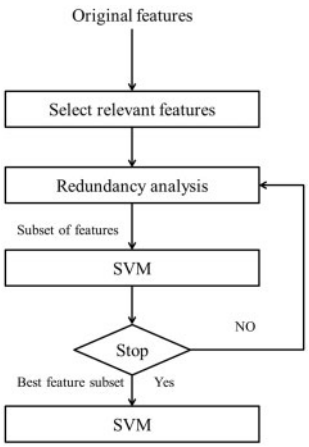

Fig. 1. Procedure of the classification model with ECBGS

```
Algorithm: Ensemble Correlation-Based Gene Selection (ECBGS)
input  : D(F1, F2, ..., FN, C)  // a training data set
         σ                      // a threshold
output : bestSubset

 1  begin
 2      i = 0, Drem = NULL;
 3      for i = 1 to N do
 4          calculate SUi,c for Fi;
 5          if (SUi,c >= σ)
 6            insert Fi to Drel;
 7      end;
 8      do begin
 9          Dtemp = Drel
10          if ( Drem is NULL )
11              spoint = getFirstElement(Drel)
12          else
13              spoint = getFirstElement(Drem)
14          remove  spoint from Drem;
15          //remove features whose ranking is higher
16          //than spoint
17          Dtemp = removeFeatures(Dtemp)
18          fp = spoint;
19          do begin
20              fq = getNextElement(Dtemp, fp);
21              do begin
22                  fqi = fq;
23                  if ( SUp,q >= SUq,c )
24                      insert fq into Drem;
25                      remove fq from Dtemp;
26                      fq = getNextElement(Dtemp, fqi);
27                  else fq = getNextElement(Dtemp, fq);
28              end until (fq == NULL);
29              fp = getNextElement(Dtemp, fp)
29          end until (Fp == NULL);
30          subset[i] = Dtemp;
31          i++;
32      end until (Drem == NULL)
33      bestSubset = getBestSubset(subset);
34  end;
```

Fig. 2. ECBGS algorithm

classification (line 33). If two subsets have identical classification accuracy, the smaller subset will be the final input.

Here, we illustrate the subset generation step of ECBGS using a heart disease dataset that is available in the UCI (University of California at Irvine) machine learning repository. The dataset consists of 13 features denoted as F_1, F_2, ..., F_{13} and 294 instances with five classes. First, the algorithm sorts all of the features in descending order based on SU values between the feature and class. Table 1 shows the SU value of each feature. Then high-ranking features with SU values that are >0 are selected. Here, 0 is the predefined threshold. As a result, F_{11}, F_9, F_3, F_{10}, F_8 and F_2 are selected as the relevant subset. Next, we perform redundancy analysis for this relevant subset. Starting with the first element in the relevant subset, F_{11}, we calculate the SU value between F_{11} and the other features. If the SU value between a feature and F_{11} is greater than the SU value between this feature and class, it will be moved into the removed subset; if not, it will remain in the current subset. As shown in Table 2, the SU value between F_{11} and F_9 is 0.427, which is larger than the SU value between F_9 and the class. Thus, the feature F_9 is considered to be redundant and moved into the removed subset along with F_{10} and F_8. Next, we choose the next element of F_{11} in the relevant subset. The next element of F_{11} is F_3 because the feature F_9 was already moved into the removed subset. We

Table 1. SU value between each feature and the class

	F_{11}	F_9	F_3	F_{10}	F_8	F_2
SU value	0.225	0.211	0.184	0.184	0.07	0.05

The features that have 0 SU value (F_4, F_1, F_{12}, F_{13}, F_5, F_7 and F_6) are not presented in this table. All of the features are sorted in descending order by SU value.

Table 2. The SU value among the features in the first iteration

	F_{11}	F_9	F_3	F_{10}	F_8	F_2
F_{11}	—	0.427	[a]	0.689	0.085	[a]
F_3	—	—	—	—	—	[a]

[a]The SU value between two features is smaller than the SU value respect to the class.

calculate the SU value between F_3 and the other features. From Table 2, we can easily see that there are no features that are redundant with F_3 such that no features will be removed, and the same scenario occurs for F_2. Consequently, subset $\{F_{11}, F_3, F_2\}$ is the subset generated by the first iteration of our algorithm, and subset $\{F_9, F_{10}, F_8\}$ is the removed subset in the first iteration.

In the second iteration, we use the first feature in the removed subset, F_9, as the starting point to remove redundant features. In addition, we do not need to analyze redundancy for the whole relevant subset because the removed features are eliminated by the features that have rankings higher than the starting point during the previous iteration. Thus, we only analyze the redundancy of the features that have rankings lower than the starting point. In this example, subset $\{F_9, F_3, F_{10}, F_8, F_2\}$ is the analyzed subset in the second iteration. As shown in Table 3, F_3, F_{10} and F_8 are redundant with F_9 such that the subset $\{F_9, F_2\}$ will be selected, and subset $\{F_3, F_{10}, F_8\}$ will be moved into the removed subset. We repeat this procedure until there are no features in the removed subset. Thus, the subsets generated from our method are $\{F_{11}, F_3, F_2\}$, $\{F_9, F_2\}$, $\{F_3, F_{10}, F_2\}$ and $\{F_8, F_2\}$. Finally, these four subsets are evaluated by SVM, and the subset that has the highest classification accuracy is selected.

3 RESULTS

3.1 Datasets

To evaluate the effectiveness of our method, we used six publicly available datasets. The datasets share common characteristics, such as a very low sample/dimension ratio.

(1) The Breast_B dataset comprises 49 samples and 1213 genes. Primary breast tumors from the Duke Breast Cancer SPORE frozen tissue bank were selected. Tumors were either positive or negative for both estrogen and progesterone receptors.

(2) The Central Nervous System (CNS) dataset (Pomeroy *et al.*, 2002) is derived from patient samples of embryonal tumors of the central nervous system. The dataset contains 60 samples, including 39 medulloblastoma survivors and 21 treatment failures, with expression profiles of 7129 genes.

Table 3. The SU value among features in the second iteration

	F_9	F_3	F_{10}	F_8	F_2
F_9	—	0.218	0.368	0.10	[a]
F_2	—	—	—	—	—

[a]The SU value between two features is smaller than the SU value respect to the class.

(3) The Leukemia dataset (Golub *et al.*, 1999) was produced in a study that was aimed at building a model to discriminate between acute myeloid leukemia and acute lymphoma leukemia tissues. Gene expression profiles have been constructed from 72 people who have either acute lymphoblastic leukemia or acute myeloid leukemia, and each sample is composed of 7129 gene expression profiles.

(4) The Lymphoma dataset (Alizadeh *et al.*, 2000) comes from a study on diffuse large B-cell lymphoma. The dataset consists of 62 samples and 4026 genes. There are three types of samples in the dataset, with 42 samples of diffuse large B-cell lymphoma, nine observations of follicular lymphoma and 11 cases of chronic lymphocytic leukemia.

(5) The MLL_leukemia dataset (Armstrong *et al.*, 2002) contains three types of leukemia samples compared with the binary-class leukemia dataset. This dataset contains a total of 72 samples in three classes, acute lymphoblastic leukemia, acute myeloid leukemia and mixed-lineage leukemia gene (MLL), which have 24, 28 and 20 samples, respectively. The number of genes is 12 582.

(6) The prostate dataset was first published by Singh *et al.* (2002); it is a two-class classification problem and contains 102 samples and 12 600 genes. One of the tasks addressed by the authors is to build a model that can discriminate between normal prostate and tumorous prostate tissue.

3.2 Parameter settings for the SVM

Selecting the kernel and appropriate parameters plays an important role in SVM classification performance. The radial basis function (RBF) kernel is a commonly used kernel for three reasons (Hsu *et al.*, 2010). First, the RBF kernel can handle nonlinear relationships between class labels and attributes. Second, it has fewer hyperparameters that influence the complexity of the model selection than the Polynomial kernel. Third, the RBF kernel has fewer numerical difficulties. The RBF kernel function is here:

$$K(x, x') = \exp(-||x - x'||^2/\sigma^2) \quad (4)$$

In our experiments, we chose the RBF kernel function, and the parameters C and γ must be optimized for the RBF kernel for each dataset. To determine the best values of C and γ, we conducted a grid-search approach using 10-fold cross validation. A number of pairs of (C,γ) values were attempted, and the pair with the best accuracy was picked in the range of $C \in \{2^{-5}, 2^{-3}, ..., 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, ..., 2^{3}\}$.

3.3 Performance evaluation

To obtain a statistically reliable predictive measurement, we performed 10-fold cross validation on all the datasets. In 10-fold cross validation, each dataset was randomly partitioned into 10 parts. Nine parts were used as the training set, and the remaining one was used as the testing dataset. In cancer classification, it is important to assess both false-positive and false-negative errors, as these two types of errors usually have different consequences (Ma and Huang, 2005). Therefore, we have used several measures to evaluate the effectiveness of our method:

- Classification accuracy = (TP + TN) / (TP + FP + FN + TN)
- Precision = TP / (TP + FP)
- TP rate = TP / (TP + FN)
- FP rate = FP / (FP + TN)
- Area Under Receiver Operating Characteristic Curve (AUC) is a single-value measurement that ranges from 0 to 1.

where true positives (TP) denote the correct classifications of positive examples, true negatives (TN) are the correct classification of negative examples, false positives (FP) denote the incorrect classification of negative examples into the positive class and false negatives (FN) represent the incorrect classification of positive examples into the negative class.

In addition, some datasets such as Breast_B, Lymphoma and MLL_Leukemia have the multi-class problem, which refers to the input data being divided into more than two categories. To solve this problem, we used the one-against-one approach that constructs N(N-1)/2 binary classifiers for an N class dataset. The posterior probabilities provided by individual binary classifiers were combined using the pairwise coupling method (Hastie and Tibshirani, 1998).

As mentioned previously, the FCBF method cannot always identify the best feature subset for high-dimensional data, which was demonstrated by comparing the prediction accuracy of the FCBF feature selection algorithm with our method. Table 4 exhibits the classification accuracy of the subsets that are selected from each iteration of our method on six datasets with a 60–40% training-test partition, and the best results in each row are shown in bold letters. The subset selected from the second iteration of our method is denoted as Set 2, the subset selected from the third iteration is denoted as Set 3 and so on. Because our method uses the same starting point as the FCBF initially, the subset selected by the first iteration of our method will be the same as the subset selected using FCBF. As shown in Table 4, although the FCBF algorithm has the same accuracy as other subsets on the MLL_Leukemia datasets, it has a lower accuracy than other subsets in most cases, which proves that our method outperforms the FCBF.

Tables 5–10 summarize the results of the classification achieved by ECBGS on different values of the relevance threshold. The performance of the classifier was evaluated by precision, TP rate, FP rate and AUC. For each dataset, the number of selected genes is listed as the '#Genes' row. With the increasing of threshold, there are no genes to be selected in some datasets. In this case, we cannot measure the performance of the classifier, as there are no inputs for the classifier.

Table 4. The classification accuracy (%) of FCBF and our method

Dataset	FCBF	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7
1	90	**95**	85	70	85	85	75
2	75	75	70.83	**83.33**	70.83	**83.33**	75
3	86.21	**93.10**	**93.10**	**93.10**	**93.10**	**93.10**	**93.10**
4	92.11	**94.74**	92.11	92.11	**94.74**	92.11	92.11
5	**100**	**100**	**100**	**100**	**100**	**100**	**100**
6	95.12	92.68	95.12	90.24	92.68	92.68	**97.65**

The training-test data partition is 60–40%, and each subset was evaluated using SVM. Set 2, Set 3, . . . , Set 7 are the feature subsets generated in the second, third, . . . , seventh iteration of our algorithm, respectively. The bold values indicate the highest classification accuracy.

Table 5. The precision, TP rate, FP rate and AUC of SVM trained using ECBGS varying the relevance threshold on the Breast_B dataset

Breast_B																	
Threshold	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
#Genes	39	39	39	39	38	27	17	10	5	4	1	1	0	0	0	0	0
Precision	0.94	0.94	0.94	0.94	0.94	0.96	0.86	0.86	0.86	0.86	0.15	0.15	NA	NA	NA	NA	NA
TP rate	0.94	0.94	0.94	0.94	0.94	0.96	0.86	0.86	0.86	0.86	0.39	0.39	NA	NA	NA	NA	NA
FP rate	0.03	0.03	0.03	0.03	0.03	0.02	0.07	0.07	0.07	0.05	0.04	0.04	NA	NA	NA	NA	NA
AUC	0.98	0.98	0.98	0.98	0.98	0.98	0.94	0.93	0.93	0.93	0.58	0.58	NA	NA	NA	NA	NA

The third row shows the number of selected genes corresponding to the relevance threshold.

Table 6. The precision, TP rate, FP rate and AUC of SVM trained using ECBGS varying the relevance threshold on the CNS dataset

CNS																	
Threshold	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
#Genes	36	36	36	36	23	7	4	1	0	0	0	0	0	0	0	0	0
Precision	0.90	0.90	0.90	0.90	0.88	0.74	0.44	0.44	NA	NA	NA	NA	NA	NA	NA	NA	NA
TP rate	0.90	0.90	0.90	0.90	0.88	0.73	0.67	0.67	NA	NA	NA	NA	NA	NA	NA	NA	NA
FP rate	0.13	0.13	0.13	0.13	0.16	0.48	0.44	0.67	NA	NA	NA	NA	NA	NA	NA	NA	NA
AUC	0.88	0.88	0.88	0.88	0.86	0.63	0.50	0.50	NA	NA	NA	NA	NA	NA	NA	NA	NA

The third row shows the number of selected genes corresponding to the relevance threshold.

Table 7. The precision, TP rate, FP rate and AUC of SVM trained using ECBGS varying the relevance threshold on the Leukemia dataset

Leukemia																	
Threshold	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
#Genes	59	59	59	59	27	6	1	0	0	0	0	0	0	0	0	0	0
Precision	0.90	0.90	0.90	0.90	0.90	0.85	0.76	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
TP rate	0.90	0.90	0.90	0.90	0.90	0.85	0.76	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
FP rate	0.11	0.11	0.11	0.11	0.15	0.23	0.29	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AUC	0.90	0.90	0.90	0.90	0.88	0.81	0.74	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

The third row shows the number of selected genes corresponding to the relevance threshold.

This situation is denoted as 'NA' in the table. From Table 5, we can know that the choice of the appropriate relevance threshold is important, as it will directly affect the performance of the classifier. The classifiers show the best performance when the threshold is in the range of 0.15–0.25 on Breast_B, CNS, Leukemia and Prostate dataset. However, for the Lymphoma and MLL_Leukemia dataset, the appropriate threshold is much larger than other datasets (0.75 and 0.4 for Lymphoma

Table 8. The precision, TP rate, FP rate and AUC of SVM trained using ECBGS varying the relevance threshold on the lymphoma dataset

Lymphoma																	
Threshold	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
#Genes	89	89	89	89	89	86	86	79	75	69	61	54	37	31	23	9	4
Precision	1	1	1	1	1	1	1	1	1	0.99	1	0.99	0.99	1	1	1	0.96
TP rate	1	1	1	1	1	1	1	1	1	0.98	1	0.98	0.98	1	1	1	0.95
FP rate	0	0	0	0	0	0	0	0	0	0.01	0	0.01	0.01	0	0	0	0.01
AUC	1	1	1	1	1	1	1	1	1	0.99	1	0.99	0.99	1	1	1	0.98

The third row shows the number of selected genes corresponding to the relevance threshold.

Table 9. The precision, TP rate, FP rate and AUC of SVM trained using ECBGS varying the relevance threshold on the MLL_Leukemia dataset

MLL_Leukemia																	
Threshold	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
#Genes	117	117	117	117	113	105	89	55	53	31	30	14	11	5	3	1	0
Precision	1	1	1	1	1	1	1	0.99	1	0.99	0.99	0.96	0.95	0.55	0.55	0.52	NA
TP rate	1	1	1	1	1	1	1	0.99	1	0.99	0.99	0.96	0.94	0.69	0.68	0.56	NA
FP rate	0	0	0	0	0	0	0	0.01	0	0.01	0.01	0.02	0.03	0.19	0.20	0.28	NA
AUC	1	1	1	1	1	1	1	0.99	1	0.99	0.99	0.99	0.98	0.78	0.78	0.66	NA

The third row shows the number of selected genes corresponding to the relevance threshold.

Table 10. The precision, TP rate, FP rate and AUC of SVM trained using ECBGS varying the relevance threshold on the Prostate dataset

Prostate																	
Threshold	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
#Genes	76	76	76	64	46	30	18	9	9	5	2	1	0	0	0	0	0
Precision	0.97	0.97	0.97	0.98	0.98	0.96	0.96	0.96	0.94	0.94	0.91	0.79	NA	NA	NA	NA	NA
TP rate	0.97	0.97	0.97	0.98	0.98	0.96	0.96	0.96	0.94	0.94	0.91	0.70	NA	NA	NA	NA	NA
FP rate	0.03	0.03	0.03	0.02	0.02	0.04	0.04	0.04	0.06	0.06	0.09	0.32	NA	NA	NA	NA	NA
AUC	0.97	0.97	0.97	0.98	0.98	0.96	0.96	0.96	0.94	0.94	0.91	0.69	NA	NA	NA	NA	NA

The third row shows the number of selected genes corresponding to the relevance threshold.

and MLL_Leukemia dataset, respectively). It is reasonable because the genes in the Lymphoma and MLL_Leukemia dataset are more relevant than others. Moreover, the number of relevant genes in the Lymphoma and MLL_Leukemia dataset is much bigger than other datasets.

In addition, we compared the accuracies of SVM using the features selected by the Gain Ratio (GR), Information Gain (IG), ReliefF and our method. Furthermore, we used the SFS and SBS search strategies in the subset generation step for these three feature selection algorithms. These classification accuracies were obtained through 10-fold cross validation. Table 11 shows classification accuracy of four feature selection algorithms on six datasets. In the majority of the datasets, the accuracy is higher than other methods, and the classifier with our proposed feature selection algorithm is found to result in the best prediction average accuracy, which was 95.71%. The other methods were found to be 89.97, 91.89, 89.54, 93.89, 89.46 and 92.66%. To catch the detailed characteristics of ECBGS, we also made the comparison of accuracy obtained from nine different partitions of training and test datasets. Figure 3 presents the classification accuracy on different sizes of training and test data. From Figure 3, one can easily observe that the prediction performance of the classification model constructed from the feature subset that is generated using our method is better than other models in most cases. Moreover, even when the training data set is small, our method shows high prediction accuracy, whereas other methods primarily make poor predictions. It is also obvious that, at least for one training-test partition, the classification accuracy of our method is 100% on all of the tested datasets. For example, the classification accuracy is 100% when the training-test partition is 70–30 or 85–15 on the Breast_B dataset (Fig. 3a), as well as when the training-test partition is 80–20 on the prostate dataset (Fig. 3f).

Table 12 shows the running time for each feature selection algorithm. For each method, the parameter of the filter part is

Table 11. 10-fold cross validation classification accuracy (%) of four feature selection methods

Dataset	GR + SVM		IG + SVM		ReliefF+SVM		ECBGS
	SFS	SBS	SFS	SBS	SFS	SBS	
1	87.76	93.88	87.76	**97.96**	93.88	91.84	95.92
2	75	73.33	73.33	90.00	65	83.33	**90.00**
3	86.11	87.5	90.28	90.28	90.28	87.50	**90.28**
4	100	100	98.31	95.16	100	100	**100**
5	95.83	98.61	94.44	95.83	94.44	97.22	**100**
6	95.10	98.04	93.14	94.12	93.14	96.08	**98.04**
Average	89.97	91.89	89.54	93.89	89.46	92.66	**95.71**

The last row is the total average accuracy of four methods on six datasets. The bold values indicate the highest classification accuracy.

set to 0 throughout the experiments. From Table 12, it is clear that ECBGS is significantly faster than the other three algorithms. Moreover, the running time of these three algorithms is extremely expensive because the number of selected features in the filter part is >2000 on the lymphoma, MLL_Leukemia and prostate datasets. This result verifies that ECBGS is suitable for high-dimensional microarray data analysis, saving a significant amount of time.

In addition to these feature selection algorithms, we can also make a comparison with the results of other methods published in the literature. Table 13 presents the best classification accuracy of other methods. From Table 13, we can see that the classification accuracy achieved by our method is higher than other methods and the number of selected genes is significantly smaller than other methods except on the Breast_B dataset. Comparing the results between the Breast_B and Prostate dataset, it is found that greater amount of genes and samples can result in smaller models than minor amount of genes and samples. This finding is not surprising, as our method is based on the information theory of entropy. Thus, for the dataset that has larger number of samples, the importance of the genes is easier to be captured.

Recently, much research has been performed on analyzing gene expression data for cancer classification using various gene selection methods. The Lymphoma dataset has been cross validated by many authors. Dettling and Bühlmann (2003) modified the boosting classifiers and applied Wilcoxon's two sample tests to select discriminative genes; comparing the results between our study and their proposed method, our best performing results are better than their results using about the same number of genes. Liu *et al.* (2010) proposed an ensemble gene selection method to analyze the gene expression data. The classification results for the Lymphoma dataset are identical to ours. However, the performance on the CNS and Prostate dataset are not better than ours (though they used leave one out cross validation). Moreover, our method returns a much smaller set of genes on these three datasets, and our method does not need to predefine the number of genes that will be selected. For the Prostate dataset, Díaz-Uriarteb *et al.* (2006) reported 0.061 error rates with 18 genes; Yang *et al.* (2006) proposed two gene selection methods which were not affected by the unbalanced sample class sizes. Their results with K-nearest-neighbor and SVM for the MLL_Leukemia and Prostate dataset show lower performance and they used larger set of genes (56 for the MLL_Leukemia and 8 for the Prostate dataset). The Breast_B dataset was firstly analyzed by West *et al.* (2001) using Bayesian approach. Dettling and Bühlmann (2002) also used Breast_B dataset and reported 4.08% of test error rates with 10 genes, which was better than West *et al.* (2001). In comparison with our method, although their method used a smaller subset of genes, our method made prediction more accurately. Finally, Hsu *et al.* (2011) introduced a hybrid feature selection method based on Information Gain and F-score. They achieved 98.6% of accuracy using SVM with 70 features on the Leukemia dataset. However, this method seems to be difficult to be used for gene selection because they select too much genes. Therefore, our proposed method is an ideal candidate for gene selection in cancer classification problem.

4 DISCUSSION

In this study, we used SVM classification method to analyze the gene expression data. A lot of research has been shown that SVM is the most effective classifier in performing accurate cancer diagnosis from gene expression data (Statnikov *et al.*, 2005; Statnikov *et al.*, 2008). SVM is interesting (Abeel *et al.*, 2010) because the number of parameters to be estimated essentially depends on the number of samples rather than on the number of features, which is particularly relevant with very small sample-to-feature ratios. Moreover, SVM has many mathematical features that make them attractive for gene expression analysis (George and Raj, 2011), including their flexibility in choosing a similarity function, sparseness of solution when dealing with large datasets, the ability to handle large feature spaces and the ability to identify outliers.

We have first examined the performance of our method using SVM on six microarray datasets in terms of precision, TP rate, FP rate and AUC. In ECBGS, there is a parameter, the relevance threshold σ. Different settings of σ will directly affect the number of selected genes. The closer σ is set to 1, the smaller the number of selected genes is. From the experiments, we found that the larger number of genes does not always lead to better performance. Therefore, it is very important to choose the appropriate

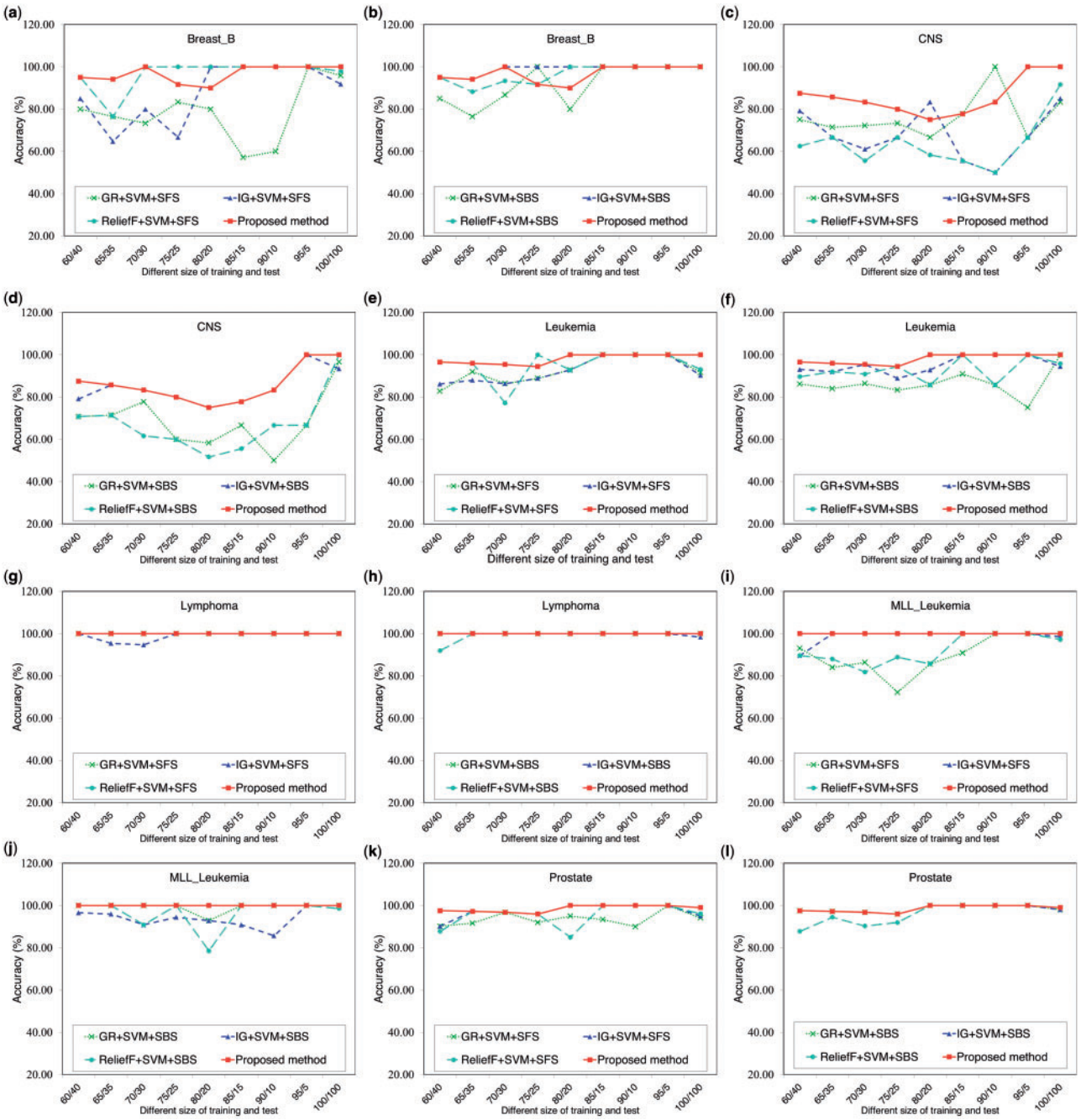

Fig. 3. Classification accuracy of different sizes of training and test datasets. (**a**) and (**b**) indicate the results on the Breast_B dataset, (**c**) and (**d**) indicate the results on the CNS dataset, (**e**) and (**f**) indicate the result on Leukemia dataset, (**g**) and (**h**) indicate the result on Lymphoma dataset, (**i**) and (**j**) indicate the results on MLL_Leukemia dataset, (**k**) and (**l**) indicate the results on Prostate dataset. We performed two experiments for each dataset: one applied the SBS strategy, and the other one applied the SFS strategy for three feature selection methods. As in the figure, when the training dataset is small, the proposed method shows high prediction accuracy, and other methods primarily make poor predictions. Moreover, the proposed method outperforms other methods in most cases

threshold for improving classification accuracy. However, choosing a proper threshold is difficult because the distinct values for each dataset are different. Through the experiments on six datasets, we found that the threshold was closely related with the 'degree' of the relevance of the genes. For example, in the Lymphoma dataset, there are >300 genes with SU value that is >0.5. In contrast, there are only 37 genes that satisfy the threshold 0.5 in Breast_B dataset. Thus, the appropriate threshold for the dataset that has higher 'degree' of the relevance has to be larger than the one for the dataset that has lower 'degree'.

Table 12. Running time (s) for each feature selection algorithm

Dataset	ECBGS	GR + SVM		IG + SVM		ReliefF + SVM	
		SFS	SBS	SFS	SBS	SFS	SBS
1	7	2959	20 090	14 885	18 735	12 471	$>10^5$
2	231	286	4129	234	1994	10 837	$>10^5$
3	243	1041	8823	608	12 083	10 807	$>10^5$
4	90	$>10^5$	$>10^5$	$>10^5$	$>10^5$	$>10^5$	$>10^5$
5	809	$>10^5$	$>10^5$	$>10^5$	$>10^5$	$>10^5$	$>10^5$
6	863	$>10^5$	$>10^5$	$>10^5$	$>10^5$	$>10^5$	$>10^5$

The running time for GR + SVM, IG + SVM and ReliefF + SVM is much >100 000 s on the No. 4, 5 and 6 datasets.

Table 13. The best accuracy (%) and the number of selected genes obtained with our method and other approaches in the literature. The bold letters indicate the results of our proposed method

Dataset	Authors	Accuracy (%)	Number of features
1	Dettling and Bühlmann	98.39	10
	Proposed method	**100**	**17**
2	Kannan *et al.*	97.78	374
	Liu *et al.*	98.33	30
	Tan *et al.*	88.33	74
	Yeh *et al.*	77.31	58
	Proposed methods	**100**	**20**
3	Fujibuchi and Kato	97.8	170
	Cho and Ryu	94.1	30
	Cho and Won	97.1	50
	Hsu *et al.*	98.6	70
	Proposed method	**100**	**5**
4	Dettling and Bühlmann	98.39	10
	Lee *et al.*	99.8	50
	Liu *et al.*	100	30
	Proposed method	**100**	**9**
5	Yang *et al.*	97.2	56
	Yang *et al.*	98.61	782
	Kannan *et al.*	100	108
	Proposed method	**100**	**6**
6	Yang *et al.*	95.1	8
	Liu *et al.*	96.08	30
	Díaz-Uriarteb *et al.*	99.4	18
	Yang *et al.*	96.08	343
	Proposed method	**100**	**5**

The bold values indicate the results of our proposed method.

Based on this observation, we can make some general recommendations based on the experiments. (i) The threshold could be selected as the mean of all the SU value of the genes respect to the class or (ii) decide the threshold as the following equation:

$$\sigma = (SU_{\max} - SU_{\min}) * 0.7 \quad (5)$$

where $SU_{\max}$ indicates the SU value of the most relevant gene respect to the class, and $SU_{\min}$ refers to the lowest one.

We have also compared ECBGS to the FCBF algorithm. The result shows that the proposed ECBGS is able to generate the most meaningful and discriminative genes in most cases. However, if the number of features is <50, FCBF tends to be more effective; it is because ECBGS produces the feature subsets by removing top informative features. If the datasets have a small number of features with a lack of useful information, removing some informative features will result in less discriminative or meaningless subsets to train the classifier. Furthermore, we found that relevant and non-redundant features are selected before repeating our method more than 10 times.

We have also made a comparison between the proposed method and other feature selection methods in terms of classification accuracy and speed. One interesting observation is that our method is still more powerful than other methods even when small data are given. It indicates that ECBGS is more appropriate than others for analyzing small datasets, such as gene expression data. Moreover, our method is significantly faster than other feature selection methods. Additionally, when the number of selected features is >2000, the computational costs of the other three feature selection methods are very expensive.

The selection of discriminant genes is a common task for cancer classification. Research in Biology and Medicine may benefit from the examination of the top ranking genes to confirm recent discoveries in cancer research, or suggest new avenues to be explored (Guyon *et al.*, 2002). Recently, several gene selection approaches (Jirapech-Umpai and Aitken, 2005; Li *et al.*, 2004) have been proposed to solve the cancer classification problem. In contrast to these methods, the prediction accuracy of our method is competitive with a small subset of genes.

5 CONCLUSIONS

In this work, we proposed an ensemble gene selection algorithm based on SU and SVM for cancer classification. To select multiple gene subsets, we used different starting points during the redundancy analysis step. In this way, we selected more informative genes than FCBF. We found that between two redundant genes, the less relevant gene makes a poor prediction; however, a combination of genes of this type can sometimes produce a competitive result. During the experiments, we used six freely accessible benchmark datasets from the Internet to meet our objective, which was to evaluate and investigate the performance of our method using the classifiers trained from both 10-cross validation and different sizes of datasets. The results show that the classification model with our proposed gene selection algorithm has higher prediction accuracy, and that our method can still achieve high accuracy when the number of training instances is small. Compared with other methods published in the literature, our method yields good results. However, for different datasets, the relevance threshold is different under the context of classification performance. Therefore, how to determine the relevance threshold in a self-adaptive matter will be focused on our future work. Moreover, we believe that our mechanism is also applicable to other feature selection problems and can be expanded to other classifications of disease states.

ACKNOWLEDGEMENTS

The authors are grateful to anonymous referees for their valuable comments.

Funding: The National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-0000478) and the Korea Biobank Project (4851-307) of the Korea Centers for Disease Control and Prevention.

Conflict of Interest: none declared.

REFERENCES

Abeel,T. *et al.* (2010) Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, **26**, 392–398.

Adankon,M. and Cheriet,M. (2009) Model selection for the LS-SVM. Application to handwriting recognition. *Pattern Recognit.*, **42**, 3264–3270.

Akay,M.F. (2009) Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Syst. Appl.*, **36**, 3240–3247.

Alizadeh,A. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Armstrong,S.A. *et al.* (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.*, **30**, 41–47.

Cho,S. and Ryu,J. (2002) Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features. *Proc. IEEE*, **90**, 1744–1753.

Cho,S. and Won,H. (2007) Cancer classification using ensemble of neural networks with multiple significant gene subsets. *Appl. Intell.*, **26**, 243–250.

Deisy,C. *et al.* (2007) Efficient dimensionality reduction approaches for feature selection. In: *International Conference on Computational Intelligence and Multimedia Applications*. Sivakasi, Tamil Nadu, pp. 121–127.

Díaz-Uriarteb,R. *et al.* (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

Dettling,M. and Bühlmann,P. (2003) Boosting for tumor classification with gene expression data. *Bioinformatics*, **19**, 1061–1069.

Doak,J. (1992) An evaluation of feature selection methods and their application to computer security. *Technical report*. Department of Computer Science, University of California at Davis.

Fujibuchi,W. and Kato,T. (2007) Classification of heterogeneous microarray data by maximum entropy kernel. *BMC Bioinformatics*, **8**, 267–277.

George,G. and Raj,V. (2011) Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile. *Int. J. Comput. Sci. Eng. Surv.*, **2**, 3.

Gheyas,I. and Smith,L. (2010) Feature subset selection in large dimensionality domains. *Pattern Recognit.*, **43**, 5–13.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Guyon,I. *et al.* (2002) Gene selection for cancer classification using support vector machines. *Machine. Learn.*, **46**, 389–422.

Han,J. and Fu,Y. (1996) Attribute-oriented induction in data mining. In:Fayyad,U.M. *et al.* (ed.) *Advances in Knowledge Discovery sand Data Mining*. AAAI Press/The MIT Press, Cambridge, MA, pp. 339–421.

Hanson,S. and Halchenko,Y. (2008) Brain reading using full brain support vector machines for object recognition: there is no 'face' identification area. *Neural Comput.*, **20**, 486–503.

Hastie,T. and Tibshirani,R. (1998) Classification by pairwise coupling. *Ann. Statist.*, **26**, 451–471.

Hsu,C.W. *et al.* (2010) *A Practical Guide to Support Vector Classification*. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (25 September 2012, date last accessed).

Hsu,H.H. *et al.* (2011) Hybrid feature selection by combining filters and wrappers. *Expert Syst. Appl.*, **38**, 8144–8150.

Jirapech-Umpai,T. and Aitken,S. (2005) Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. *BMC Bioinformatics*, **6**, 148.

Kannan,S. *et al.* (2010) A novel hybrid feature selection via symmetrical uncertainty ranking based local memetric search algorithm. *Knowl. Based Syst.*, **23**, 580–585.

Kim,Y. *et al.* (2000) Feature selection for unsupervised learning via evolutionary search. In: *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, USA, pp. 365–369.

Kohavi,R. and John,G.H. (1997) Wrappers for feature subset selection. *Artif. Intell.*, **97**, 273–324.

Lee,J. *et al.* (2002) An extensive comparison of recent classification tools applied to microarray data. *Comput. Stat. Data Anal.*, **48**, 77–87.

Li,P. *et al.* (2011) QSE: a new 3-D solvent exposure measure for the analysis of protein structure. *Proteomics*, **11**, 3793–3801.

Li,T. *et al.* (2004) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, **20**, 2429–2437.

Liu,H. *et al.* (2010) Ensemble gene selection for cancer classification. *Pattern Recognit.*, **43**, 2763–2772.

Liu,H. *et al.* (2002) Feature selection with selective sampling. In: *Proceedings of the Nineteenth International Conference on Machine Learning, 2002*. Sydney, Australia, pp. 395–402.

Liu,H. and Setiono,R. (1996) A probabilistic approach to feature selection—a filter solution. In: *Proceedings of the Thirteenth International Conference on Machine Learning*. Bari, Italy, pp. 319–327.

Liu,H. and Yu,L. (2005) Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng.*, **17**, 491–502.

Ma,S. and Huang,J. (2005) Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics*, **21**, 4356–4362.

Pomeroy,S.L. *et al.* (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436–442.

Pok,G. *et al.* (2010) Effective feature selection framework for cluster analysis of microarray data. *Bioinformation*, **4**, 385–389.

Singh,D. *et al.* (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **2**, 203–209.

Statnikov,A. *et al.* (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, **21**, 631–643.

Statnikov,A. *et al.* (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, **9**, 319.

Tan,A.C. and Gilbert,D. (2004) Ensemble machine learning on gene expression data for cancer classification. *Bioinformatics*, **20**, 3583–3593.

Talavera,L. (2005) An evaluation of filter and wrapper methods for feature selection in categorical clustering. In: *Proceedings of 6th International Symposium on Intelligent Data Analysis*. Madrid, Spain, pp. 440–451.

Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY.

West,M. *et al.* (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.

Xie,J. and Wang,C. (2011) Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases. *Expert Syst. Appl.*, **38**, 5809–5815.

Xiong,M. *et al.* (2001) Feature (Gene) selection in gene expression-based tumor classification. *Mol. Genet. Metab.*, **73**, 239–247.

Yang,K. *et al.* (2006) A stable gene selection in microarray data analysis. *BMC Bioinformatics*, **7**, 228.

Yang,C.H. *et al.* (2009) IG-GA: a hybrid filter/wrapper method for feature selection of microarray data. *J. Med. Biol. Eng.*, **30**, 23–28.

Yeh,J.Y. (2008) Applying data mining techniques for cancer classification on gene expression data. *Cybern. Syst. Int. J.*, **39**, 583–602.

Yu,L. and Liu,H. (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*. pp. 856–863.

Yu,L. and Liu,H. (2004) Efficient feature selection via analysis of relevance and redundancy. *J. Mach. Learn. Res.*, **5**, 1205–1224.

Saeys,Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Zaghloul,W. *et al.* (2009) Text classification: neural networks vs support vector machines. *Ind. Manag. Data Syst.*, **109**, 708–717.

Zhang,L. *et al.* (2009) A novel representation for apoptosis protein subcellular localization prediction using support vector machine. *J. Theor. Biol.*, **259**, 361–365.