SVM-RFE with Relevancy and Redundancy Criteria for Gene Selection

Piyushkumar A. Mundra[1] and Jagath C. Rajapakse[1,2]

[1] Bioinformatics Research Center, School of Computer Engineering,
Nanyang Technological University, 50 Nanyang Avenue,
Singapore 639798
[2] Singapore-MIT Alliance, N2-B2C-15, 50 Nanyang Avenue, Singapore
asjagath@ntu.edu.sg

Abstract. This paper introduces a novel gene selection method incorporating mutual information in the support vector machine recursive feature elimination (SVM-RFE). We incorporate an additional term of mutual information based minimum redundancy maximum relevancy criteria along with feature weight calculated by SVM algorithm. We tested proposed method on colon cancer and leukemia cancer gene expression dataset. The results show that the proposed method performs better than the original SVM-RFE method. The selected gene subset has better classification accuracy and better generalization capability.

Keywords: Gene selection, mutual information, minimum redundancy, maximum relevancy, SVM-RFE, cancer classification.

1 Introduction

DNA-microarray has emerged as a very powerful method to analyze gene expression of cells. This high throughput technology enables simultaneous monitoring of expression level of thousand of genes and hence results in a vast pool of data. Detecting differences among the gene expressions can be very useful in disease diagnosis and distinction of specific tumor type. Most of gene expression datasets contain small number of samples and very high number of genes. For accurate classification, it is extremely imperative to select relevant genes. Because it is possible that totally irrelevant genes are selected, the classifier still produces very high classification accuracy.

Broadly, two approaches of gene selections appear in machine learning and bioinformatics literature: the filter and wrapper methods [1-2]. Filter methods are purely based on the statistical correlations and independent of the classifier used. They evaluate the goodness of the feature subset only by intrinsic characteristic of the data. Based on the relation of each single gene with class labels by the calculation of simple statistical measures computed from the empirical distribution, feature ranking is performed. Some of the statistical measures are Shannon-entropy, Euclidean distance, Kolmogorov-dependence, t-score, P-metric, mutual information etc [3].

On the other hand, wrapper methods rank features based on their effect on the classification accuracy. In this method, the feature selected will be highly dependent

J.C. Rajapakse, B. Schmidt, and G. Volkert (Eds.): PRIB 2007, LNBI 4774, pp. 242–252, 2007.

on the classification algorithm used. It is claimed by many authors that wrapper approach obtains better subset of predictive genes than filter approach [3]. Both filter and wrapper methods have their own advantages and disadvantages. Filters are usually simple and have less computational cost but fail to provide a small subset of genes. Wrappers are more computationally complex and gene subset selected may not generalize with other classification algorithm as it is highly dependent on the classification algorithm used in feature ranking. Different wrapper approaches are proposed by various authors [4-10].

Support Vector Machine - Recursive Feature Elimination (SVM-RFE) is one of the most successful wrapper method based algorithm in the feature (gene) ranking and hence reduction in the dimensionality of the dataset [10]. Multiple SVM-RFE (MSVM-RFE) has shown improvement on the classification accuracy over SVM-RFE [7]. Similarly like SVM-RFE, [6] presented Recursive Cluster Elimination (RCE) algorithm. Though SVM-RFE is very powerful method, it does not ensure to select the genes which are maximally relevant to the class and at the same time possesses minimum redundancy among them as feature selected are highly dependent on the weights derived from SVM algorithm.

Maximum gene relevancy and minimum gene redundancy is very important for gene selection as it can result in more balanced coverage of the feature space, capturing broad characteristics of the dataset and improvement in the classification accuracy. Minimum Redundancy Maximum Relevancy (MRMR) algorithm was proposed for maximizing gene relevancy and minimizing the gene redundancy [11-13]. They had ranked all the genes and according to information theoretic criteria and selected the top ranked genes for the classification. Another approach in degree of differential prioritization (DDP) criteria was proposed to strike the balance between relevancy and redundancy [14].

In this paper, we propose a novel hybrid approach to incorporate MRMR criteria in SVM-RFE algorithm itself. Mutual information based additional term will be added along with the SVM-ranking criteria. This additional term will be useful in achieving maximum relevant and minimum redundant gene subset without sacrificing on classification accuracy. The resulting gene subset may represent whole gene expression dataset broadly and may have better generalization capabilities.

The rest of the paper is organized as follows: in Section II, we will review Minimum Redundancy and Maximum Relevancy (MRMR) criteria and SVM-RFE. In Section III, we will propose hybrid algorithm of MRMR with SVM-RFE. Section IV will discuss the experimental procedure to test the algorithm on various gene expression datasets and results. Finally, in Section V we analyzed the results and conclude the paper.

2 Method

2.1 Minimum Redundancy Maximum Relevancy (MRMR) Criteria

Here, this criteria attempts to find the subset of genes having maximal relevancy to target class and least redundant among themselves. If a gene is expressed randomly or uniformly in different class, its relevancy to the respective class will be zero, i.e. its

mutual information with the class will be zero. Strongly expressed gene in one class will have larger mutual information with the respective class.

Let $S = \{x_i: i = 1,2,\ldots,n\}$ be subset of dataset where i is the gene and x_i is expression of gene i. Each x_i can be represented as $(x_{i1}, x_{i2},\ldots, x_{iJ})$, where x_{ij} is the expression of i^{th} gene in j^{th} sample. If target classes are $c = \{c_1,c_2,\ldots,c_K\}$, where K is number of class, then $I(c;x_i)$ will quantify the relevance of gene i to the classification task. Thus the by maximizing the total relevance of all genes in subset S should be the maximum relevance criteria:

$$\max\ V_i,\quad V_i = \frac{1}{|S|}\sum_{x_i \in S} I(c;x_i) \tag{1}$$

Here, $|S|$ is number of genes in subset S.

It is possible that the features selected from the maximum relevancy criteria are highly redundant. Removal of features from this redundant set will not affect the discrimination power. The 'minimum redundancy' means select the features in such a way that they are mutually maximally dissimilar in the subset. Mutual information can also be used as a measure to find similarity between two features. Hence, following redundancy removal criteria can be used to achieve mutually exclusive features:

$$\min\ W_i,\quad W_i = \frac{1}{|S|^2}\sum_{x_i, x_j \in S} I(x_i, x_j) \tag{2}$$

It is necessary to optimize both maximum relevancy and minimum redundancy criteria to get the best feature subset. To achieve so, we will need a single objective function which can describe both the criteria. Such simplest objective criteria can be written as,

$$\max\left(\frac{V_i}{W_i}\right) \quad or \quad \max(V_i - W_i) \tag{3}$$

In present work, we have used the quotient objective criteria with SVM-RFE.

2.2 SVM-RFE

To select the genes for accurate cancer classification, SVM-RFE algorithm was proposed by [10]. The algorithm produces nested subset of the genes by backward elimination, starting with all the features and removing one feature in every iteration. Here, the feature removal is based on the SVM ranking criteria, the i^{th} feature with the smallest ranking score $c_i = (w_i)^2$ is eliminated, where w_i is the corresponding weight of i^{th} feature calculated from SVM.

The reason for choosing $c_i = (w_i)^2$ as ranking criteria is the feature removed by this criteria will have least change in the objective function. The objective function in the SVM-RFE is $J = \frac{1}{2}\|w\|^2$. Optimal Brain Damage (OBD) algorithm [15] has explained

this effect. It approximates the change in objective function caused by removing the feature by second order Taylor series expansion of the objective function,

$$\Delta J(i) = \frac{\partial J}{\partial w_i}\Delta w_i + \frac{\partial^2 J}{\partial w_i^2}(\Delta w_i)^2 \tag{4}$$

At the optimum, first derivative can be neglected and using $J = \frac{1}{2}\|w\|^2$, equation (4) becomes

$$\Delta J(i) = (\Delta w_i)^2 \tag{5}$$

The SVM-Recursive feature elimination procedure can be described as follows:

Start: Ranked feature set R = [ ] and selected feature subset S = [1, 2,..., n]
Repeat until all features are ranked
a) Train linear SVM with feature set S in input variable

b) Compute the weight vector

$$w = \sum_i \alpha_i y_i x_i$$

c) Compute the ranking score of features

$$c_i = (w_i)^2$$

d) Select the feature with smallest ranking score

$$e = \arg\min(c)$$

e) Update R = [e,R]; S = S – [e]

Output: Ranked feature set R.

3 SVM-RFE with MRMR Criteria

The final subset obtained from the SVM-RFE algorithm may contain many redundant genes. Many biologically important genes may have lost because of less weight compare to these redundant features. We propose to integrate MRMR criteria with the weight criteria in SVM-RFE. MRMR criteria will make the subset less redundant and SVM-RFE weight will make sure the selected genes are useful in classification. The final selected gene subset will represent best mutually exclusive genes. The final dataset obtained will represent the whole dataset better than obtained by SVM-RFE alone.

The detailed SVM-RFE algorithm with MRMR criteria is discussed below:

Start: Ranked feature set R = [ ] and selected feature subset S = [1, 2,..., n]

Repeat until all features are ranked
a) Train linear SVM with feature set S in input variable and calculate the weight of each vector $w_{i,svm}$

$$w_{i,svm} = \sum_i \alpha_i y_i x_i$$

b) Calculate the class relevancy of each feature and mutual information among features using equation (2)

$$w_{i,MI} = \frac{I(c; x_i)}{\frac{1}{|S|}\sum_{x_i, x_j} I(x_i; x_j)}$$

c) Compute the weight vector

$$w_i = \frac{|w_{i,svm}|}{\max(w_{svm})} + \frac{w_{i,MI}}{\max(w_{MI})}$$

d) Compute the ranking score of features

$$c_i = (w_i)^2$$

e) Select the feature with smallest ranking score

$$e = \arg\min(c)$$

f) Update $R = [e,R]$; $S = S - [e]$

Output: Ranked feature set R.

4 Experiments

4.1 Data

To evaluate the performance of MRMR based SVM-RFE, experiments were carried out on two most popular gene expression dataset, leukemia cancer dataset [16] and colon cancer dataset [17]. For the present study, we had only taken available training data of the leukemia dataset. These datasets were obtained from http://ligarto.org/rdiaz/Papers/rfVS/randomForestVarSel.html [9]. Both the dataset were further divided in two separate training and testing dataset. The details of the dataset are given in Table 1.

Table 1. Sizes of training and test sets, number of gene in two gene expression dataset

Dataset	Training Samples	Testing Samples	Total Number of Genes
Colon	42	20	2000
Leukemia	24	14	3051

4.2 Preprocessing

The dataset was randomly divided into training and testing set with maintaining the class ratio in both the sets. Training dataset was normalized to zero mean and unit variance. These continuous datasets were directly used in SVM-RFE after normalization.

It is difficult to find the mutual information of two continuous features. Hence for the simplicity of calculating mutual information, training dataset was discretized. Discretization will also help in the noise reduction. Mean (μ) and standard deviation (σ) of each individual gene expression variable was used to discretize the observation. Following criteria is then used to categorize the data: Data larger than μ + σ/2 will be changed to state 2 ; Data in between μ + σ/2 and μ - σ/2 will be transformed to state 0 ; Data smaller than μ - σ/2 will be transformed to state -2.

4.3 Parameter Estimation

SVMs performances depend upon its two critical hyperparameters, the kernel function and the regularization parameter *C*. It is imperative to select these parameters carefully. In present study, linear SVMs were used, which require only *C* parameter to tune. *C* values were chosen from finite set $\{2^{-20},\ldots,2^{0},\ldots,2^{15}\}$. This set was used for both recursive feature elimination (both from SVM-RFE and hybrid of mutual information and SVM-RFE) and performance evaluation.

To estimate the prediction generalization error, CV can be used. The resulting estimate of generalization error is often used as model selection criteria. Model that has the smallest generalization error are chosen. In *k*-fold CV, the data instances are divided into *k* – mutual folds with equal size. Model is trained with *k*-1 folds and tested on omitted fold. This average testing error, calculated by testing on each fold, represents the generalization error estimate. Another important variant of *k*-fold CV is 'Leave-one-out' method. In this method, *k* equals to the number of data instances. Classifier is built with all samples except one and tested on the omitted sample.

As sample size is small and class imbalance prevalent in most of the dataset, we used Matthew's Correlation Coefficient (MCC) with 10 fold cross validation. After each 10 fold CV, we summed the true positive (TP), true negative (TN), false positive (FP) and false negative (FN). These values were used to calculate MCC[1] parameter. MCC will vary between -1 to 1. Higher the MCC value means classifier has high sensitivity and specificity.

To increase the speed of the numerical simulations with both SVM-RFE and proposed hybrid method, we eliminate *m* features each time when number of features *n* is large in recursive feature subset *S*. If $n > 10000$, we choose $m = 100$, if $1000<n<10000$, *m* will be 10, and if $n < 1000$, $m = 1$.

4.4 Testing

It is necessary to check the validation accuracy of the classifiers as many times classifier fits training data extremely good but their prediction accuracy on unseen data may be very poor. However, the training and testing set of gene expression data are small and test error may not represent the true validation accuracy due to "unfortunate" partition of training and testing sets. To avoid such situation, we merge the training and testing datasets and then partition the total samples again in training

[1] $$MCC = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

and testing sets by random sampling. This process is performed 100 times and for each time, classifier is trained on the training set and tested on the corresponding testing set. The test error, sensitivity and specificity were computed for these 100 trials.

The feature ranking is carried out using only the training data. The goodness of feature subset is evaluated using linear SVM classifier trained with ranked genes as input variables. No data discretization was done for testing. For each gene expression dataset and method, we test feature subsets with number of genes ranging from 1 to 100. We take the gene subset with the least average test error as the best feature subset. This gene set is used to calculate the performance of the each method in terms of sensitivity and specificity on each gene expression dataset.

To compare the results of the proposed algorithm with MRMR filtering, we ranked the features using program available at http://research.janelia.org/peng/proj/mRMR/index.htm [13]. We obtained top 100 features from the training dataset for both the gene expression data. The classification accuracy of gene subset is evaluated with SVM algorithm.

In all feature selection methods and testing the classifier, we had used LIBSVM – 2.83 software [18].

4.5 Results

We applied proposed hybrid of mutual information and SVM-RFE on Colon cancer and Leukemia cancer gene expression dataset. To compare our method, we also tested with SVM-RFE and MRMR method. The results are shown in the Tables 2 and 3. The results are shown in terms of number of genes, overall accuracy of the classifier and class-wise accuracy (sensitivity and specificity). In Figs 1-2, average test error of linear SVM classifier on selected gene subsets with SVM-RFE, MRMR + SVM and hybrid method is plotted.

From Table 2-3, it is clear that proposed hybrid of mutual information and SVM-RFE performs better than the SVM-RFE in both Colon cancer and Leukemia gene expression dataset. Apart from accuracy, number of genes in the best subset of both dataset is also small compare to SVM-RFE. Comparing with MRMR method, hybrid method performs better in colon cancer dataset and results are comparable in the Leukemia dataset.

Table 2. Performance of SVM classifier with feature selection by SVMRFE, hybrid of SVM and MI based RFE and MRMR method on Colon Cancer gene expression dataset

	Number of genes	Accuracy (%)	Sensitivity (%)	Specificity (%)
MRMR - SVM	13	87.7 ± 7.12	86.2 ± 8.32	89.13 ± 8.32
SVM-RFE	74	88.5 ± 5.97	85.6 ± 12.19	90.19 ± 7.48
SVM-RFE with MRMR	51	89.3 ± 6.71	85.98 ± 12.49	91.44 ± 7.91

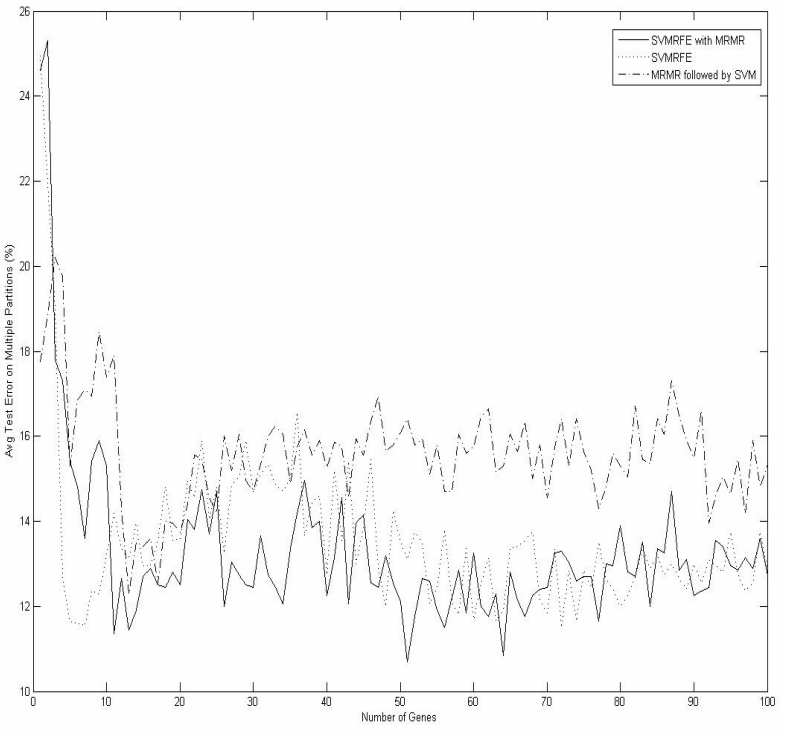

Fig. 1. Average misclassification error of gene subset selected by SVMRFE, hybrid of SVM and MI based RFE and MRMR method on 100 random testing with Colon Cancer gene expression dataset

Table 3. Performance of SVM classifier with feature selection by SVMRFE, hybrid of SVM and MI based RFE and MRMR method on Leukemia Cancer gene expression dataset

	Number of genes	Accuracy (%)	Sensitivity (%)	Specificity (%)
MRMR - SVM	74	97.6 ± 4.19	99.36 ± 3.78	97.04 ± 5.5
SVM-RFE	44	97.13 ± 4.67	100 ± 0	96.13 ± 6.23
SVM-RFE with MRMR	21	97.27 ± 4.75	99.25 ± 2.5	96.63 ± 5.92

5 Discussion and Conclusion

As seen the Table 2-3, the results are better than SVM-RFE both in terms of small number of genes and better classification accuracy. Results are comparable with MRMR based feature ranking. MRMR is basically a type of filter method. Mostly,

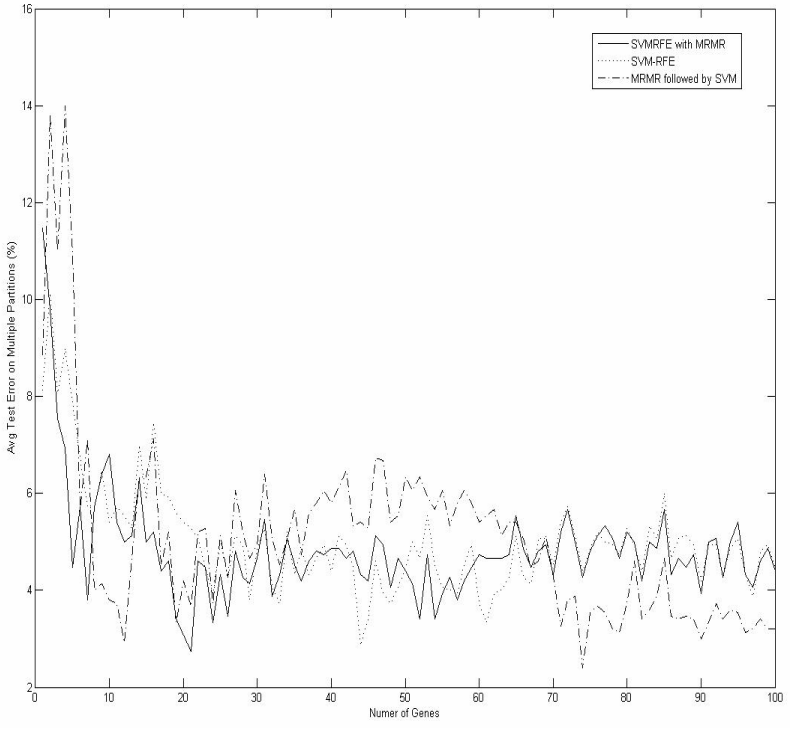

Fig. 2. Average misclassification error of gene subset selected by SVMRFE, hybrid of SVM and MI based RFE and MRMR method on 100 random testing with Leukemia Cancer gene expression dataset

filter method gives large gene subset for comparable classification accuracy with wrapper method. Small number of gene in subset will produce inferior classification accuracy in filter approach. The results with colon cancer and leukemia gene expression dataset show exactly the same nature.

From the result table, we observe that standard deviations of all performance measures (accuracy, sensitivity and specificity) over 100 times training and testing are large. In other words, the variability of single test is large and such test results are not fair performance reference due to possible 'unfortunate' partitioning. When dataset is small, the risk of 'unfortunate' partitioning increases.

In our proposed hybrid method, advantage of both continuous data (in SVM feature weighting) and discrete data (for MRMR) is encoded. Hence, we believe that this ranking method is noise tolerant without affecting the continuous nature of the gene expression data.

As seen from the Figure 1 and 2, hybrid method gave small classification error than SVMRFE and MRMR filtering in most part of gene subset tested. It means this method has better generalization capability. The proposed hybrid method selects the genes based on their effect on classification accuracy and make sure that they are least

redundant among themselves. As the gene subset selected is best representing the whole dataset and least redundant, better generalization was expected and hence seen in the results. We also believe that the gene subset selected by this method should give similar classification accuracy with other classifiers.

Finally in conclusion, the gene subset selected by this method represents the broader class characteristics than SVMRFE. It will insure better screening of the dataset and better representation of whole dataset in small number of most relevant and least redundant gene subset.

References

1. Blum, A., Langley, A.: Selection of relevant features and examples in machine learning. Artif. Intell. 97, 245–271 (1997)
2. Kohavi, R., John, G.: Wrappers for feature subset selection. Artif. Intell. 97, 273–324 (1997)
3. Inza, I., Larranaga, P., Blanco, R., Cerrolaza, A.: Filter versus wrapper gene selection approaches in DNA microarray domains. Arti. Intelli. Medicine 31, 91–103 (2004)
4. Rakotomamonjy, A.: Variable selection using SVM criteria. J. Mach. Learn. Res (Special Issue on Variable Selection) 3, 1357–1370 (2003)
5. Ruiz, R., Riquelme, J., Aguilar-Ruiz, J.: Incremental wrapper-based gene selection from microarraydata for cancer classification. Patter. Recog. 39, 2383–2392 (2006)
6. Yousef, M., Jung, S., Showe, L., Showe, M.: Recursive Cluster Elimination (RCE) for Classification and Feature Selection from Gene Expression Data. BMC Bioinfo. 8, 144 (2007)
7. Kai-Bo, D., Rajapakse, J.C., Wang, H., Azuaje, F.: Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data. IEEE Trans. Nanobio. 4, 228–234 (2005)
8. Rajapakse, J.C., Kai-Bo, D., Yeo, W.K.: Proteomic Cancer Classification with Mass Spectrometry Data. Am. J. Pharmacogenomics 5, 281–292 (2005)
9. Diaz-Uriarte, R., Andres, S.: Gene Selection and classification of microarray data using random forest. BMC Bioinfo. 7, 3 (2006)
10. Guyon, I., Weston, J., Barhill, S., Vapnik, V.: Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46, 389–422 (2002)
11. Ding, C., Peng, H.: Minimum Redundancy Feature Selection from Microarray Gene Expression Data. J. Bioinfo. Compu. Bio. 3, 185–205 (2005)
12. Ding, C., Peng, H.: Minimum Redundancy Feature Selection from Microarray Gene Expression Data. In: Proceed. Second IEEE Comp. System. Bioinfo. Conferen., pp. 523–529. IEEE Computer Society Press, Los Alamitos (2003)
13. Peng, H., Long, F., Ding, C.: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans. Patt. Anal. Machi. Intell. 27, 1226–1237 (2005)
14. Ooi, C., Chetty, M., Teng, S.: Differential prioritization between relevance and redundancy in correlation-based feature selection techniques for multiclass gene expression data. BMC Bioinfo. 7, 320–339 (2006)
15. LeCun, Y., Denker, J., Solla, S., Howard, R., Jackel, L.: Optimal Brain Damage. In: Touretzky, D. (ed.) Advances in Neural Information Processing Systems II, pp. 598–605. Morgan Kaufmann, San Mateo, CA (1990)

16. Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., Lander, E.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression. Science 286, 531–537 (1999)
17. Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., Levine, A.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. PNAS 96, 6745–6750 (1999)
18. Chang, C., Lin, C.: LIBSVM: A Library for Support Vector Machines (2001), http://www.csie.ntu.edu.tw/ cjlin/libsvm