Correlation-Based Relevancy and Redundancy Measures for Efficient Gene Selection

Kezhi Z. Mao and Wenyin Tang

School of Electrical & Electronic Engineering
Nanyang Technological University
Singapore 639798

Abstract. The gene-label correlation provides an effective measure of the relevancy of a gene. However, this measure evaluates genes on an individual basis, and the gene sets thus obtained may exhibit severe redundancy. In this study, we propose a new correlation heuristic for set-based gene selection, with the goal of alleviating the redundancy problem. The new correlation heuristic consists of two components that account for gene relevancy and redundancy respectively. The relevancy of a gene is evaluated in terms of its correlation with class label on an individual basis, while the redundancy of a gene with respect to a given gene subset is measured by its correlation with a new dimension built upon the gene subset. The new correlation heuristic retains the simplicity of individual gene evaluation and the redundancy handling capacity of set-based gene evaluation. Two different ways of using the relevancy and redundancy measures are presented in this study. One way is the maximization of the ratio of relevancy measure to redundancy measure, and another way is the maximization of the relevancy measure subtracting redundancy measure. Experimental studies on six gene expression problems show that both criteria produce excellent results.

1 Introduction

Gene selection has been an active research area since the birth of the gene microarray technology, and a variety of gene selection algorithms have been proposed. The various gene selection algorithms can be classified into two categories, namely individual gene selection (see for example [8,4,7,15]) and gene subset selection (see for example [14,11,6,10,12,21,20,1]). The two types of gene selection algorithms often serve different purposes. If gene selection is for efficient pattern classification or class prediction, subset-based gene selection should be employed. This is because a gene subset consisting of top individually ranked genes may far from optimal due to the severe redundancy existed. Whatever category a gene selection algorithm belongs to, it involves an evaluation criterion to measure the goodness of an individual gene or a subset of genes. A variety of evaluation criteria have been used in the gene selection algorithms mentioned above, motivated by different considerations. These include t-test, F-test, Fisher ratio, entropy, cross validation error, Bayesian error estimation, loss functions of regression, and support vector machine (SVM) criteria etc.

J.C. Rajapakse, B. Schmidt, and G. Volkert (Eds.): PRIB 2007, LNBI 4774, pp. 230–241, 2007.

Correlation measures have also been used for gene evaluation and selection. To minimize gene redundancy, one available correlation measure is the set-based correlation heuristic proposed by [13], where the merit of a feature subset is evaluated using the ratio of the average feature-label correlation to the average feature-feature correlation. Similar measures were also proposed in [3]. Another correltaion-based algorithm is the two-phase relevancy-redundancy analysis proposed by [19], where relevant genes are first selected through individual relevancy analysis, and redundant genes are then removed through Markov blanket-based redundancy analysis. But our experiment studies show that this algorithm could over-prune and the number of genes finally obtained might be insufficient.

In this study, we propose a new correlation heuristic for forward, *i.e.* bottom-up, gene selection. The new correlation consists of two components accounting for relevancy and redundancy respectively. The relevancy of a gene is evaluated individually in terms of its correlation with class label, while the redundancy of a gene with respect to a given gene subset is measured by its correlation with the output of the classifier built upon the gene subset. This way of evaluating redundancy is an outstanding character of the new correlation heuristic. The rationale lies in the fact that the major discriminative information underlying the gene subset is captured by the classifier, and thus the correlation between the candidate gene and the output of the classifier reflects the redundancy of the candidate gene with respect to the gene subset. Two ways of using relevancy and redundancy measures are presented. One is the ratio of relevancy measure to redundancy measure, and another is the relevancy subtracting redundancy. Through maximizing the two criteria, genes with high relevancy and minor redundancy could be selected.

The new correlation heuristic inherits the simplicity of individual gene evaluation and the redundancy handling capacity of set-based evaluation. Experimental studies show that both criteria produce excellent results.

2 Correlation-Based Relevancy and Redundancy Measures for Gene Selection

2.1 Relevancy and Redundancy Measures

Assume there are N training data pairs:

$$\{\mathbf{x}(1), y(1)\}, \{\mathbf{x}(2), y(2)\}, \ldots, \{\mathbf{x}(N), y(N)\}$$

where $y(k)$ denotes the class label of sample k, with value of either $+1$ or -1. $\mathbf{x}(k)$ is the feature vector of sample k consisting of n genes:

$$\mathbf{x}(k) = [x_1(k), x_2(k), \ldots, x_n(k)]$$

The gene-label correlation is defined as the correlation between a gene and the class label:

$$r_{yx_i} = \frac{1}{N-1} \frac{\sum_{k=1}^{N} x_i(k) y(k)}{\sigma_{x_i} \sigma_y} \tag{1}$$

where σ_{x_i} and σ_y denote the standard deviation of gene x_i and class label y respectively. The gene-label correlation reflects the predictive power, or relevancy, of a gene and could be used to identify biologically related genes of certain biological phenomenon of interest. However, the correlation criterion Eqn (1) evaluates genes on an individual basis, without considering correlations between genes. Severe redundancy might exist if it is used to select gene subsets. To achieve good pattern classification results, an ideal gene subset should possess the following properties:

(i) having maximum relevancy;
(ii) having minimum redundancy.

To yield gene subsets with maximum relevancy and minimum redundancy, we can select gene subsets that maximizes the ratio of relevancy measure to redundancy measure or the difference between the two measures [13,3].

In a forward gene selection algorithm, the gene subset is built up step by step, by adding one gene at one step. Assume m genes have already been selected: $s_m = \{x_1, x_2, \ldots, x_m\}$, the objective is to select the next best gene. To select the gene with maximum relevancy and minimum redundancy, we can evaluate and select genes using the following criteria

$$J_1 = \frac{R_{yx_i}}{R_{s_m x_i}} \tag{2}$$

or

$$J_2 = R_{yx_i} - R_{s_m x_i} \tag{3}$$

where R_{yx_i} denotes the relevancy measure of gene x_i, and $R_{s_m x_i}$ denotes redundancy measure of gene x_i with respect to gene subset s_m. The gene with the maximum J_1 or J_2 should be selected.

The relevancy of a gene can be easily measured in terms of its correlation with class label as in Eqn (1) or other measures such as Fisher ratio. The major issue here is how to evaluate the redundancy of x_i with respect to the given subset $\mathbf{s}_m$. In [13] and [3], the redundancy is measured in terms of the average correlation between candidate x_i and those in the gene subset selected s_m. Next, we propose a new approach to redundancy evaluation.

2.2 A New Approach to Redundancy Evaluation

The basic idea of the new way of evaluating redundancy of a candidate gene with respect to gene subset s_m is to project data from the m-dimensional space to a new one-dimensional space using a linear transform, and then measure the redundancy of a candidate gene based on its correlation with the new dimension. Assume the linear transform is given by:

$$z_m(k) = \sum_{j=1}^{m} w_j x_j(k) \tag{4}$$

where w_j, $j = 1, 2 \ldots, m$ are the coefficients of the linear transform. Eqn(4) is such a transform that the major discriminative information underlying the m dimensions, *i.e.* m genes in s_m, is compressed onto z_m. The linear transform that projects data from m-dimensional space to one-dimensional space can be obtained by the support vector machine (SVM) method because the SVM classifier captures the major discriminative power underlying s_m.

The redundancy of x_i with respect to gene subset s_m is measured using the correlation between x_i and z_m. The rationale of the new way of evaluating redundancy can be explained from the point of view of variable selection in multiple regression. Assume the regression of class label on the m features in s_m is as Eqn (4), then the resultant regression error is given by:

$$e(k) = y(k) - z_m(k) \tag{5}$$

The variable to be selected next should have maximum correlation with the regression error. Assume

$$\begin{aligned}
\mathbf{y} &= [y(1), y(2), \ldots, y(N)]^T \\
\mathbf{e} &= [e(1), e(2), \ldots, e(N)]^T \\
\mathbf{z_m} &= [z_m(1), z_m(2), \ldots, z_m(N)]^T \\
\mathbf{x}_i &= [x_i(1), x_i(2), \ldots, x_i(N)]^T
\end{aligned}$$

The correlation between x_i and e, denoted by r_{ex_i} is given by:

$$\begin{aligned}
r_{ex_i} &= \frac{1}{N-1} \frac{\mathbf{x}_i^T \mathbf{e}}{\sigma_e \sigma_{x_i}} \\
&= \frac{1}{N-1} \frac{\mathbf{x}_i^T \mathbf{y} - \mathbf{x}_i^T \mathbf{z}_m}{\sigma_e \sigma_{x_i}}
\end{aligned} \tag{6}$$

where σ_e denote the standard deviation of error signal e. If genes, class label and sample projections on the new dimension are normalised to zero mean and unit standard deviation, Eqn (6) can be written as

$$r_{ex_i} = \frac{1}{\sigma_e} [r_{yx_i} - r_{z_m x_i}] \tag{7}$$

where r_{yx_i} and $r_{z_m x_i}$ denotes the correlations between x_i and class label and the output of the classifier respectively. To ensure the minimum regression error after adding the new feature, selection of the new feature should be based on maximization of r_{ex_i}. A comparison of Eqn (7) with Eqn (3) shows that if the correlation between x_i and class label is used to evaluate the relevancy of x_i, then the redundancy $R_{s_m x_i}$ can be measured using the correlation between gene x_i and the output of the classifier built upon s_m.

The heuristic J_1 and J_2 can be rewritten as:

$$J_1 = \frac{|\mathbf{y}^T \mathbf{x}_i|}{|\mathbf{z}^T \mathbf{x}_i|} \tag{8}$$

$$J_2 = |\mathbf{y}^T \mathbf{x}_i| - |\mathbf{z}^T \mathbf{x}_i| \tag{9}$$

where $|.|$ denotes the absolute value. This is because the correlations can take both positive or negative values.

J_2 actually can be modified by putting a weighting element on the redundancy measure:

$$J_2^* = |\mathbf{y}^T \mathbf{x}_i| - \lambda |\mathbf{z}^T \mathbf{x}_i| \tag{10}$$

where λ denotes the weighting element.

The main characteristic of the present study is that the redundancy of a gene with respect to a gene subset selected is measured using the correlation between the gene and a new dimension built upon the gene subset. An important issue here is how to create the new dimension. As analysed above, the correlation measure Eqn (3) is equivalent to regression error based feature evaluation when the role of the previously selected features is controlled. This suggest that we may control the effect of the previously selected gene subsets when a new dimension is created after a new gene is added. This is briefly described below. A new dimension, named z_2, is first created using x_1 and x_2. Selection of the third gene is based on the correlation criteria where the redundancy of a candidate gene is measured using the correlation between the candidate gene and z_2. After the 3^{rd} gene, say x_3 is selected, a new dimension z_3 is created using x_3 and z_2. In this process, the creation of a new dimension is always done in a 2-dimensional space. And the creation can be based on different approaches such as support vector machine (SVM).

Due to small sample size and very high dimensionality in gene expression data, the training data could be mapped to the class label. Thus, the redundancy measure would approaches the relevancy measure and a zero value of the criterion would be obtained. To overcome this problem, the new dimension created at each step is rotated by an angle. Taking z_{m-1}, x_i and z_m as an example, where z_m is created by z_{m-1} and x_i.

$$z_m(k) = w_{m1} z_{m-1}(k) + w_{m2} x_i(k) \tag{11}$$

Taking the z_{m-1} as an reference, the angle of the new dimension is given by:

$$\alpha = \arctan\left(\frac{w_{m2}}{w_{m1}}\right) \tag{12}$$

After a few genes are selected, the sample projections on z_{m-1} are very close to class labels, and play dominant role in creating z_m. Thus, the value of w_{m1} has a much greater amplitude than w_{m2}, and the angle becomes very small. Hence we have:

$$\alpha \approx \frac{w_{m2}}{w_{m1}} \tag{13}$$

To rotate the new axis, we can reduce the value of w_{m1} to w_{m1}/γ, where $\gamma > 1$. Thus, the new angle is given by:

$$\beta \approx \gamma \frac{w_{m2}}{w_{m1}} = \gamma\alpha \tag{14}$$

The new dimension is usually obtained by optimizing certain criterion. The transform obtained is therefore optimal in the sense of maximum separating margin in support vector machine, maximum class separability in Fisher's linear discriminant analysis, and minimum regression error in least mean square estimation etc. The rotation introduce with deteriorate the optimality, and is therefore can be regarded as a regularization.

Criterion J_1 and J_2^* consist of two components. One component accounts for the relevance of the gene, and another component accounts for the redundancy of the gene with respect to gene subset s_m. The relevance is measured on an individual basis, while the redundancy is measured on a set basis. The merit of this way of evaluating a candidate gene is that it retains the simplicity of individual gene evaluation and the capacity of redundancy handling of set-based gene evaluation.

2.3 The Correlation Criteria-Based Gene Selection Algorithm

The procedure of forward gene selection based on the correlation J_1 and J_2^* is summarized below:

(i) Normalise data including class label to zero mean and unit standard deviation.
(ii) Evaluate the correlation between class label and each of the n genes in the candidate gene pool: $x_1, x_2, \ldots, x_n$. Identify the gene that has the maximum correlation measure, say x_j, add it to the gene subset and remove it from the candidate gene pool. Let $z = x_j$.
(iii) Evaluate the correlation between z and each of the $n-1$ genes in the candidate gene pool, and calculate J_1 or J_2^* using Eqn (8) or (10). Identify the gene having the maximum measure, say x_i, add it to the gene subset and remove it from the candidate gene pool.
(iv) Train the linear SVM classifier using the genes in the gene subset selected and denote the decision value of classifier for the training samples as z. Normalise z to zero mean and unit standard deviation. Evaluate the correlation between z and each of the $n-2$ genes in the candidate gene pool, and calculate J_1 or J_2^* using Eqn (8) or (10). Identify the gene having the maximum measure, say x_k, add it to the gene subset and remove it from the candidate gene pool.
(v) Step (iv) is repeated until a stopping criterion, say the number of genes selected, is satisfied.

To identify the $m+1^{th}$ gene from a candidate gene pool of $n-m$ genes at step $m+1$, the computations involved include training a linear classifier such as a linear support vector machine (SVM) once and performing $n-m$ vector product in N-dimensional space, where N is the training sample size. Apparently, the computational complexity of the proposed method is very limited.

3 Experimental Studies

In the experiment, the performance of the proposed correlation heuristic was studied. For comparison purpose, the two-phase relevancy-redundancy analysis proposed in [19] and set-based correlation heuristic proposed in [13] were also studied. In addition, the recursive feature elimination (RFE) algorithm [12], which is often considered as a benchmark algorithm, was also studied.

The performance of these gene selection algorithms was evaluated in terms of classification error rate. The study in [2] revealed that error estimation based on cross validation including leave-one-out and repeated k-fold cross validation may exhibit excessive variability. In this study, .632+ bootstrapping [5] was used. In the bootstrap testing, 200 replica were generated to estimate the error rate, and the splits of training and test data in the 200 replica were kept identical during the testing of the gene selection algorithms.

Six gene expression datasets were used to test the performance of the proposed algorithm. The eight datasets are summarized in Table 1:

Table 1. Datasets description

Datasets	Original sources	Genes
Leukaemia	[8]	7129
Breast cancer (ER)	[18]	7129
Breast cancer (LN)	[18]	7129
Lung cancer	[9]	12533
CNS tumour	[16]	7129
Breast cancer	[17]	24481

Each of these datasets was standardized to zero mean and unit standard deviation across genes. Since the dimensionality (*i.e.* the number of genes) of gene expression data is very high, and most of these genes are irrelevant to the discriminant task, a pre-selection procedure was employed to reduce the number of candidate genes to 1000 based on Fisher's ratio, which is an individual gene ranking criterion. All the experiments and comparisons in this work were conducted on the pre-selected data.

The experimental results on the 6 datasets are shown in Figures 1-6 respectively. On each dataset, 4 algorithms were tested, including the recursive feature elimination (RFE), correlation-based feature selection (CFS), and the two new correlation heuristics Eqn (8) and Eqn (9), named as CH1 and CH2 respectively. In the experimental study, the weight λ on the redundancy measure in criterion J_2^* was set to 2, and the weight on slack variable in RFE was set to a wide range of values, as small as 0.001 and as great as 100, but the results were almost identical.

Across the 6 problems, the two-phase relevancy-redundancy analysis produced gene subsets consisting of just a few genes since the Markov blanket principle removed most of the candidate genes while the other 3 algorithms used the number of genes selected as the stopping criterion. As shown in Figures 1-6, the

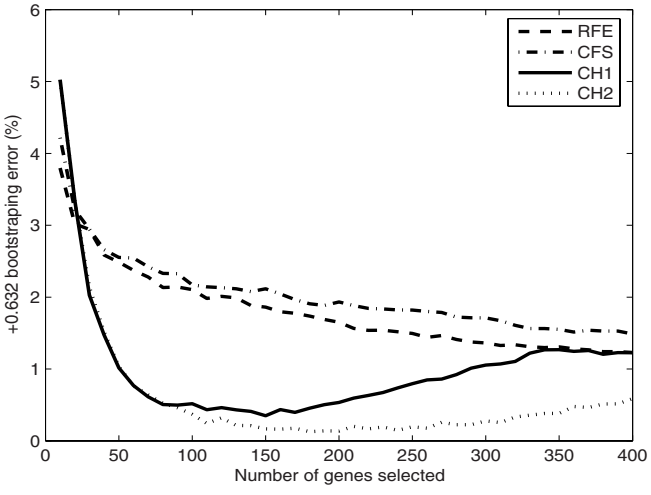

Fig. 1. Comparison of RM and RRM with RFE in Leukaemia problem

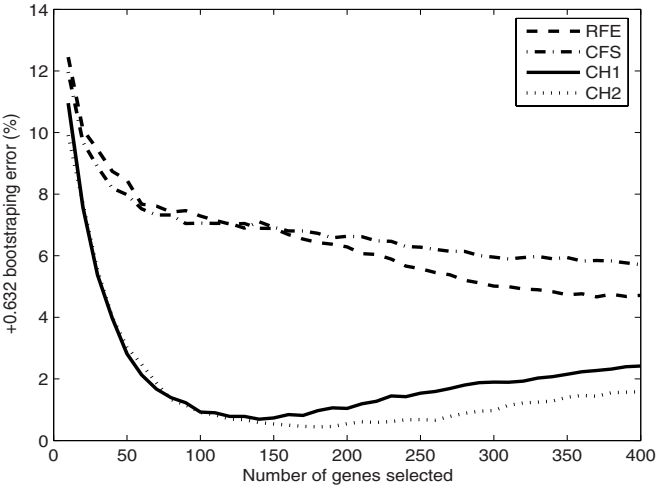

Fig. 2. Comparison of RM and RRM with RFE in Breast Cancer (ER) problem

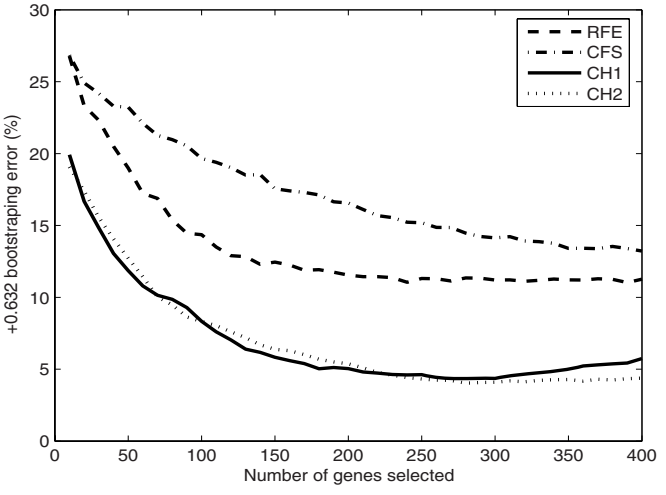

Fig. 3. Comparison of RM and RRM with RFE in Breast Cancer (LN) problem

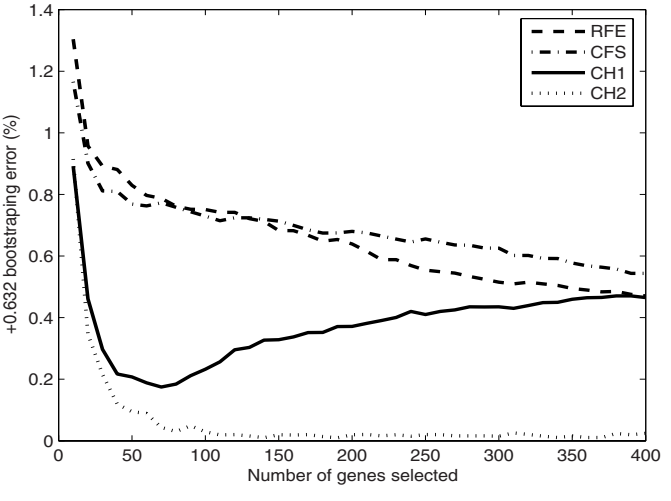

Fig. 4. Comparison of RM and RRM with RFE in Lung Cancer problem

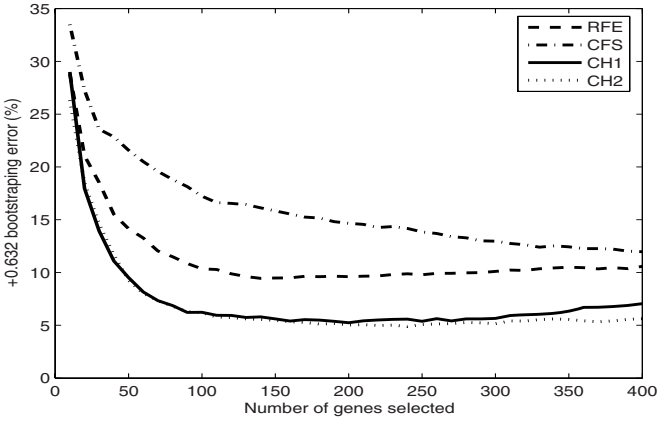

Fig. 5. Comparison of RM and RRM with RFE in CNS Tumor problem

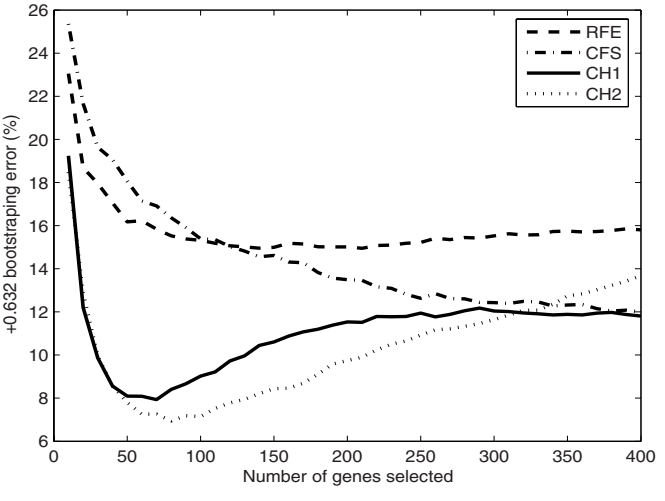

Fig. 6. Comparison of RM and RRM with RFE in Breast Cancer problem

RFE algorithm outperform the CFS algorithm in all the 6 problems. However, CH1 and CH2 outperform both CFS and RFE substantially. The results of CH2 are a bit inferior to those of CH1, this is probably because the introduction of the weight element λ improves the adaptability and flexibility of the correlation heuristic.

4 Conclusions

In this study, we have proposed a new correlation heuristic for efficient gene selection, where relevancy and redundancy components of a gene are considered explicitly in merit evaluation. Two formulae have been presented by different way of combining the two components. The proposed correlation heuristic retains the simplicity of individual gene evaluation and the capacity of redundancy handling of set-based gene evaluation. Experimental studies have shown that the correlation heuristic produces gene subsets leading to excellent classification accuracy.

References

1. Braga-Neto, U., Dougherty, E.R.: Bolstered error estimation. Pattern Recognition 37(6), 1267–1281 (2004a)
2. Braga-Neto, U.M., Dougherty, E.R.: Is cross-validation valid for small-sample microarray classification? Bioinformatics 20(3), 374–380 (2004b)
3. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. In: Proceedings of 2nd IEEE Computer Society Bioinformatics Conference. IEEE Computer Society Press, Los Alamitos (2003a)
4. Dudoit, S., Fridyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97, 77–87 (2002)
5. Efron, B., Tibshirani, R.: Improvements on cross-validation: the.632+ bootstrap method. Journal of the American Statistical Association 92(438), 548–560 (1997)
6. Fan, L., Yang, Y.: Analysis of recursive gene selection approaches from microarray data. Bioinformatics 21(19), 3741–3747 (2005)
7. Furlanello, C., Serafini, M., Merler, S., Jurman, G.: Entropy-based gene ranking without selection bias for the predictive classification of microarray data. BMC Bioinformatics 4(54) (2003)
8. 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 monitoring. Science 286, 531–537 (1999)
9. Gordon, G.J., Jensen, R.V., Hsiao, L.-L., Gullans, S.R., Blumenstock, J.E., Ramaswamy, S., Richards, W.G., Sugarbaker, D.J., Bueno, R.: Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Research 62 (2002)
10. Guan, Z., Zhao, H.: A semiparametric approach for marker gene selection based on gene expression data. Bioinformatics 21(4), 529–536 (2005)
11. Gui, J., Li, H.: Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics 21(13), 3001–3008 (2005)
12. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46(1-3), 389–422 (2002)
13. Hall, M.: Correlation-based feature selection for discrete and numeric class machine learning. In: Proceedings of Seventeenth International Conference on Machine Learning, San Francisco, CA, USA (2000)

14. Li, Y., Campbell, C., Tipping, M.: Bayesian automatic relevance determination algorithms for classifying gene expression data. Bioinformatics 18(10), 1332–1339 (2002)
15. Liu, X., Krishnan, A., Mondry, A.: Entropy-based gene selection for cancer classification using microarray data. BMC Bioinformatics 6(76) (2005)
16. Pomeroy, S.L.: Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415 (2002)
17. van't Veer, Dai, H., van de Vijver, He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 415 (2002)
18. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J.A., Marks, J.R., Nevins, J.R.: Predicting the clinical status of human breast cancer by using gene expression profiles. Proc. Natl. Acad. Sci. USA 98(20), 11462–11467 (2001)
19. Yu, L., Liu, H.: Efficient feature selection via analysis of relevance and redundancy. Journal of Machine Learning Research 5 (2004)
20. Zhang, H.H., Ahn, J., Lin, X., Park, C.: Gene selection using support vector machines with non-convex penalty. Bioinformatics 22(1), 88–95 (2006)
21. Zhou, X., Mao, K.Z.: Ls bound based gene selection for dna microarray data. Bioinformatics 21(8), 1559–1564 (2005)