Gene Selection Using Wilcoxon Rank Sum Test and Support Vector Machine for Cancer Classification

Chen Liao[1], Shutao Li[1], and Zhiyuan Luo[2]

[1] College of Electrical and Information Engineering, Hunan University, Changsha 410082, China
[2] Department of Computer Science, Royal Holloway College, University of London, Egham, Surrey, TW20 0EX, United Kingdom
shutao_li@yahoo.com.cn

Abstract. Gene selection is an important problem in microarray data processing. A new gene selection method based on Wilcoxon rank sum test and Support Vector Machine (SVM) is proposed in this paper. First, Wilcoxon rank sum test is used to select a subset. Then each selected gene is trained and tested using SVM classifier with linear kernel separately, and genes with high testing accuracy rates are chosen to form the final reduced gene subset. Leave-one-out cross validation (LOOCV) classification results on two datasets: Breast Cancer and ALL/AML leukemia, demonstrate the proposed method can get 100% success rate with the final reduced subset. The selected genes are listed and their expression levels are sketched, which show that the selected genes can make clear separation between two classes.

1 Introduction

Cancer turns to be one of the most threatening diseases in the world, which is often caused by abnormal cells that spread and grow unconventionally. Molecular level classification of cancer cells is becoming important now. As a preprocessing step of the classification, gene selection becomes a critical issue [1].

The purpose of gene selection is to eliminate redundant, noisy or irrelevant genes and select the most informative subset of genes to enhance the generalization performance. Traditional statistical methods for classification have been widely used for gene selection. In [1], six gene selection heuristics based on entropy theory, entropy-based, χ^2-statistics, t-statistics, are introduced. In [2], entropy measure and Wilcoxon rank sum test are combined to find relevant genes. Support vector machines (SVMs) also have been extensively utilized to deal with gene selection problems [3-9]. Various methods of recursive feature elimination based on SVM (SVM-RFE) are discussed in [4-9].

In this paper, a new gene selection method is proposed. First, Wilcoxon rank sum test is used to preprocess the dataset to decease the dimensionality. Then each gene is trained and tested using SVM classifier separately, and each gene gets its corresponding accuracy rate. The accuracy rates are ranked in descending order. Finally, the genes with high accuracy rates are selected to form a subset.

Y. Wang, Y. Cheung, and H. Liu (Eds.): CIS 2006, LNAI 4456, pp. 57–66, 2007.

This paper is organized as follows. In the next section, the introduction of Wilcoxon rank sum test is given. In section 3, the basic theory of SVM is introduced. In section 4, a new gene selection method using SVM is proposed. In section 5, the experimental setup and results are shown, and in the last section, the paper is concluded.

2 Feature Selection Using Wilcoxon Rank Sum Test

Because the large dimensionality will increase the complexity and prolong the running time, the dataset is preprocessed by Wilcoxon rank sum test at first. The statistics formula is:

$$s(g) = \sum_{i \in \mathbf{N}_0} \sum_{j \in \mathbf{N}_1} I((\mathbf{x}_j^{(g)} - \mathbf{x}_i^{(g)}) \le 0) \tag{1}$$

where I is the discrimination function, if the logic expression in the bracket is true, the value of I is 1, or else it is 0. $\mathbf{x}_i^{(g)}$ is the expression value of the sample i in the gene g. $\mathbf{N}_0$ and $\mathbf{N}_1$ are the index sets of different classes of samples. $s(g)$ can represent the measurement of the difference between the two classes, when it is closer to 0 or closer to the max value of $n_0 n_1$ (here $n_0 = |\mathbf{N}_0|$, $n_1 = |\mathbf{N}_1|$), the corresponding gene is more important to the classification. As a result, according to (2), the importance degree of gene can be calculated:

$$q(g) = \max(s(g), n_0 n_1 - s(g)) \tag{2}$$

Genes are ranked according to each genes' $q(g)$, and the top p genes are selected to form a new subset.

3 Support Vector Machines

SVM has appeared as an extensively used classifier of statistical learning theory. The training set is supposed to be $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, with each input $\mathbf{x}_i \in R^m$ and $y_i \in \{\pm 1\}$. The SVM maps $\mathbf{x}$ to $\mathbf{z} = \varphi(\mathbf{x})$ in a Hilbert space F by a nonlinear map $\varphi : R^m \to F$. The dimensionality of F is quite high in most conditions. When the data is linearly separable in F, a separation hyperplane $(\langle \mathbf{w}, \varphi(\mathbf{x}) \rangle + b)$ is constructed by the SVM, and the separation margin of the hyperplane between the positive and negative examples is maximized. By minimizing $\|\mathbf{w}\|$, the w for the optimal hyperplane is gained, and the solution can be presented as $\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \varphi(\mathbf{x}_i)$ for some certain $\alpha_i \ge 0$. The vector of α_i's, $\Lambda = (\alpha_i, ..., \alpha_N)$, can be gained by solving the following quadratic programming problem:

$$\text{maximize } W(\Lambda) = \Lambda^{\mathrm{T}} 1 - \frac{1}{2} \Lambda^{\mathrm{T}} Q \Lambda \tag{3}$$

with respect to Λ, subject to the constraints $\Lambda \geq 0$ and $\Lambda Y = 0$. Here, $Y^{\mathrm{T}} = (y_1, ..., y_N)$ and Q are symmetric matrixes with elements

$$Q_{ij} = y_i y_j \left\langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \right\rangle \tag{4}$$

For those α_i greater than zero, the relevant training examples should lie along the margins of the decision boundary, and these are defined as the support vectors.

However, due to the high dimensionality of F and $\varphi(\mathbf{x}_i)$ and $\varphi(\mathbf{x}_j)$ in (2), the way is not so practical. So a critical characteristic of the SVM, and of kernel methods in general, plays an important role here. It is that one can gain $\left\langle \varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j) \right\rangle$ in (2) without calculating $\varphi(\mathbf{x}_i)$ and $\varphi(\mathbf{x}_j)$ explicitly first, and this is realized via using kernel function. The kernel methods supply wonderful tools to process and compare many types of data, and supply state-of-the-art performance in many cases. Here some kinds of kernel functions are introduced as follows:

① Linear kernel

$$k_L(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T \mathbf{x}_2 \tag{5}$$

where $\mathbf{x}_1$ is the value of the independent variable for which one seeks an estimate, and $\mathbf{x}_2$ are the values of the independent variable in the data.

② Polynomial kernel

$$k_P(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1^T \mathbf{x}_2)^d \tag{6}$$

where d is the degree of the polynomial. The kernel k_P of degree 2 is corresponding to a feature space spanned by all products of two variables, that is, $\{\mathbf{x}_1^2, \mathbf{x}_1\mathbf{x}_2, \mathbf{x}_2^2\}$.

③ Gaussian RBF kernel

$$k_G(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\left\|\mathbf{x}_1 - \mathbf{x}_2\right\|^2 / 2\sigma^2) \tag{7}$$

where σ is a parameter. The Gaussian kernel is one of the most popular utilized kernels in practice due to its capacity to produce nonparametric classification functions [3].

4 Feature Selection and Classification Using SVM

The gene selection and classification using SVM criteria is illustrated as follows.

Let x_{ij} be the measurement of the expression level of the jth gene for the ith sample, where j=1, 2, …, n, $\mathbf{X}_{reduced}$=(x_{ij})m,n denotes the expression levels of the genes selected by Wilcoxon rank sum test, i.e.,

$$\mathbf{X}_{reduced} = \begin{array}{c} \begin{matrix} \text{gene 1} & \text{gene 2} & & \text{gene } n \end{matrix} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \end{array}$$

Here we assume $\mathbf{x}_1, \cdots, \mathbf{x}_m$ are the m samples, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \cdots, x_{in}]$. Let $\boldsymbol{Y}=[y_1, \ldots, y_m]^{\mathrm{T}}$ denote the class labels of m samples.

Step 1: training each gene and obtaining relevant accuracy rates

If there are k samples in the training set, and m-k samples in the test set, n SVM classifiers with linear kernel are trained using the k training samples. For example, to the jth SVM, the input vector is $\mathbf{x}_j = \left[x_{1j}, x_{2j}, \cdots, x_{kj}\right]$, the output is $\mathbf{y} = [y_1, y_2, \cdots, y_k]$. Then the training accuracies of all the n genes can be obtained. The genes with high accuracy rates are more informative than those with low ones.

Step 2: ranking the accuracies and selecting the most informative ones

The accuracy rates of all the genes are ranked in descending order, and the top N genes with highest training accuracies are chosen to form the final reduced subset. Now the new gene dataset is as follows.

$$\mathbf{X}_{final} = \begin{array}{c} \begin{matrix} \text{gene } i & \text{gene } j & & \text{gene } N \end{matrix} \\ \begin{bmatrix} x_{1i} & x_{1j} & \cdots & x_{1N} \\ x_{2i} & x_{2j} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{mi} & x_{mj} & \cdots & x_{mN} \end{bmatrix} \end{array}$$

Step 3: classifying with support vector machine

The new training subset with selected genes is used to train one SVM. And the testing set is tested by the classifier to get the corresponding testing accuracy rate.

5 Experimental Setup and Results

5.1 Experimental Setup

The selection and classification performance of the proposed method is evaluated by two benchmark datasets. The Breast Cancer dataset contains 7129 genes and 38 samples, including 18 ER+ (estrogen receptor) samples and 20 ER- samples [10]. The Leukemia dataset contains 7129 genes too and 72 samples, which contain 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloid leukemia (AML) [11].

Because of the small number of samples, leave-one-out cross validation (LOOCV) is utilized to test the accuracy of the proposed method. Firstly, the LOOCV procedure removes one sample to form a testing set, and the remaining samples are used for gene selection and classifier construction. Finally, the constructed classifier is tested by using the removed sample. After all samples have been left out and tested in turn, the final classification error is obtained by the fraction of errors over the total number of training samples.

5.2 Experimental Results

The classification results of the proposed method by using different parameter values are shown in Table 1. The p in the Wilcoxon rank sum test is set to 300, which means that the top 300 genes are selected to form a new subset. The C_1 is the penalty

Table 1. The classification performance (%) with different parameters C_1,C_2 and N

C_1	C_2	N	Linear kernel		Polynomial kernel		Gaussian kernel	
			B.C.	L.	B.C.	L.	B.C.	L.
1	1	20	94.7	95.8	92.1	94.4	94.7	94.4
		30	92.1	94.4	94.7	94.4	92.1	93.1
		40	94.7	93.1	92.1	91.7	94.7	93.1
		50	97.4	94.4	97.4	91.7	92.1	95.8
	10	20	**100.0**	93.1	97.4	93.1	**100.0**	95.8
		30	92.1	90.2	94.7	91.7	94.7	94.4
		40	94.7	94.4	92.1	93.1	94.7	93.1
		50	97.4	93.1	97.4	94.4	94.7	94.4
	100	20	**100.0**	91.7	**100.0**	88.9	**100.0**	93.6
		30	92.1	88.9	94.7	91.7	94.7	91.7
		40	94.7	94.4	92.1	93.1	94.7	91.7
		50	97.4	93.1	97.4	94.4	94.7	91.7
10	1	20	92.1	97.2	92.1	95.8	92.1	94.4
		30	97.4	97.2	97.4	**98.6**	94.7	95.8
		40	**100.0**	97.2	97.4	**98.6**	94.7	95.8
		50	94.7	97.2	97.4	95.8	94.7	95.8
	10	20	92.1	95.8	92.1	93.1	92.1	97.2
		30	97.4	**98.6**	97.4	**98.6**	97.4	97.2
		40	**100.0**	**98.6**	97.4	97.2	**100.0**	97.2
		50	94.7	95.8	97.4	95.8	97.4	97.2
	100	20	92.1	94.4	92.1	93.1	92.1	94.4
		30	97.4	97.2	97.4	**98.6**	97.4	**98.6**
		40	**100.0**	**98.6**	97.4	97.2	**100.0**	**98.6**
		50	94.7	95.8	97.4	95.8	97.4	95.8
100	1	20	92.1	94.4	92.1	95.8	89.5	94.4
		30	94.7	97.2	97.4	**98.6**	92.1	94.4
		40	94.7	97.2	92.1	**98.6**	94.7	95.8
		50	97.4	97.2	97.4	**98.6**	94.7	95.8
	10	20	92.1	97.2	92.1	97.2	94.7	95.8
		30	94.7	**98.6**	97.4	**98.6**	94.7	97.2
		40	97.4	95.8	92.1	97.2	94.7	97.2
		50	97.4	97.2	97.4	95.8	97.4	97.2
	100	20	92.1	97.2	92.1	97.2	92.1	97.2
		30	94.7	**98.6**	97.4	**98.6**	94.7	**98.6**
		40	94.7	95.8	92.1	97.2	94.7	97.2
		50	97.4	97.2	97.4	95.8	97.4	95.8

factor of the SVM classifier with linear kernel for gene selection. The C_2 is the penalty factor of the SVM classifier used for final classification. The B.C. is the abbreviation of Breast Cancer dataset, and L. is the abbreviation of Leukemia dataset. For polynomial kernel, the degree is set to 2, and in Gaussian kernel, the σ is set to 0.1. The size of the selected genes N is set to 20, 30, 40 or 50.

For the Breast Cancer data, the proposed method can reach to the best accuracy rate 100%, with 20 selected genes and linear SVM classifier. For the Leukemia dataset, the highest accuracy rate is 98.61%, with at least 30 genes selected.

From Table 1, we can also conclude that the linear kernel classifier can give the same or even better classification performance than the polynomial and Gaussian kernels. To the gene selection phase, the penalty factor C_1 of the linear classifier should not be too large, such as over l00, or else the classification accuracy will decrease.

Table 2 shows the performance of the other methods on the two datasets as reported in the literature. All these methods use LOOCV and so their classification accuracies can be directly compared. As can be seen, the proposed method, attain the best classification accuracy (of 100%) on the Breast Cancer dataset. On the Leukemia dataset, the proposed method also outperforms all other methods except for the JCFO(Joint Classifier and Feature Optimization) [13] with linear kernel.

Table 2. Classification accuracies (%) obtained by the various methods as reported in the literature

Classifier	Breast Cancer	Leukemia
Support Vector Machine (linear kernel) [13]	97.4	94.4
Relevance Vector Machine (linear kernel) [14]	94.7	94.4
Relevance Vector Machine (no kernel) [14]	89.5	97.2
Sparse probit regression(linear kernel) [15]	97.4	97.2
Sparse probit regression(no kernel) [15]	84.2	97.2
JCFO (linear kernel) [12]	97.4	**100.0**
Proposed method	**100.0**	98.6

In the 72-fold cross-validation cycle, we conduct SVM-based gene selection and classification operations, as described in the section 4. There is no guarantee that the same subset of genes will be selected in each of the 72 cycles in 72-fold cross-validation. However, the most informative genes tend to be selected more consistently than others across cycles. So we select a minimal set of genes by collecting the genes with the highest picked frequencies during the 72 fold.

On the leukemia data, our method selected 25 genes (Table 3) from the microarray gene expression data. Using the selected genes, the LOOCV classification is performed. The training and testing accuracies were 100% and 100% respectively

using SVM classifier with Gaussian kernel($\sigma = 0.0001, C = 1$). The selected genes are also belong to the top 33 genes used by Ben-Dor et al[14].

For Breast Cancer dataset, there are 21 genes which are chosen in all 38 LOOCV circulations. The 21 genes selected from Breast Cancer are listed in Table 4. When we use the new reduced subset to perform the LOOCV classification using SVM, the accuracy rates can reach to 100% using SVM classifier with Gaussian kernel($\sigma = 0.0001, C = 10$).

We show in Fig.1 and Fig.2 the expression values of the selected 25 and 21 genes subset from leukemia and breast cancer dataset, respectively. In Fig.1 and Fig.2, the columns represent different genes and the rows represent expression levels in different samples. In Fig.1, the left 25 columns are AML patients and the right 47 columns are ALL patients. In Fig.2, the left 18 columns are ER+ patients and the right 20 columns are ER- patients. From the figures, we can see that the selected genes can make very clear separation between two classes.

Table 3. Leukemia data: most significant genes

Gene ID	Gene description
760	CYSTATIN A
804	Macmarcks
1144	SPTAN1 Spectrin, alpha, non-erythrocytic 1 (alpha-fodrin)
1685	Termianl transferase mRNA
1779	MPO Myeloperoxidase
1829	PPGB Protective protein for beta-galactosidase (galactosialidosis)
1834	CD33 CD33 antigen (differentiation antigen)
1882	CST3 Cystatin C (amyloid angiopathy and cerebral hemorrhage)
1928	Oncoprotein 18 (Op18) gene
2020	FAH Fumarylacetoacetate
2111	ATP6C Vacuolar H+ ATPase proton channel subunit
2121	CTSD Cathepsin D (lysosomal aspartyl protease)
2288	DF D component of complement (adipsin)
2354	CCND3 Cyclin D3
3252	GLUTATHIONE S-TRANSFERASE, MICROSOMAL
3320	Leukotriene C4 synthase (LTC4S) gene
4196	PRG1 Proteoglycan 1, secretory granule
4328	PROTEASOME IOTA CHAIN
4377	ME491 gene extracted from H. sapiens gene for Me491/CD63 antigen
4847	Zyxin
6041	APLP2 Amyloid beta (A4) precursor-like protein protein 2
6185	SELL Leukocyte adhesion protein beta subunit
6281	MYL1 Myosin light chain (alkali)
6376	PFC Properdin P factor, complement
6855	TCF3 Transcription factor 3 (E2A immunoglobulin enhancer binding factors E12/E47)

Table 4. Breast cancer data: most significant genes

Gene ID	Gene description
495	Human mRNA for KIAA0068 gene
715	Human mRNA for KIAA0187 gene
1505	Human, plasminogen activator inhibitor-1 gene
1512	Human uroporphyrinogen III synthase mRNA
2542	Human Ca2-activated neutral protease large subunit
3087	Human heat-stable enterotoxin receptor mRNA
3823	Human protein kinase (JNK2) mRNA
4220	Human chitotriosidase precursor mRNA
4414	Human Bloom syndrome protein (BLM) mRNA
4445	Human retinal protein (HRG4) mRNA
4473	Homo sapiens Trio mRNA
5188	Human clone 23721 mRNA sequence
5433	Human sarcolipin (SLN) mRNA
5444	Homo sapiens sin3 associated polypeptide p18 (SAP18)
5524	Human mRNA for raf oncogene
5639	Human mRNA for D-amino acid oxidase (EC 1.4.3.3)
5859	Human mRNA for ribonuclease/angiogenin inhibitor
5914	Human mRNA for corticotropin-releasing factor binding protein (CRF-BP)
6247	H.sapiens mRNA for protein kinase C mu
6419	H.sapiens PrP gene, exon 2
6951	un-named-transcript-1 from H.sapiens cdc25 gene promoter region.

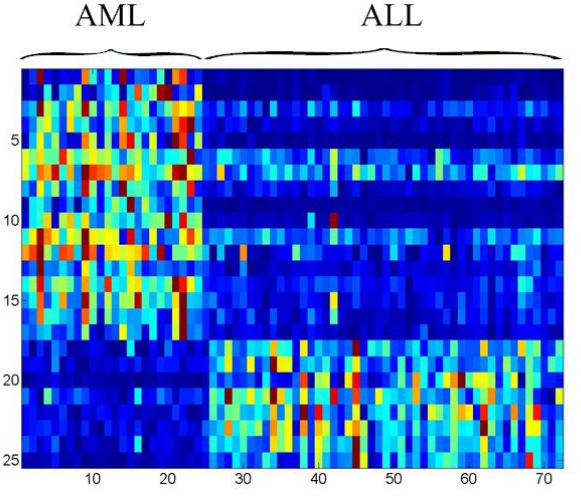

Fig. 1. Selected 25 genes from Leukemia dataset

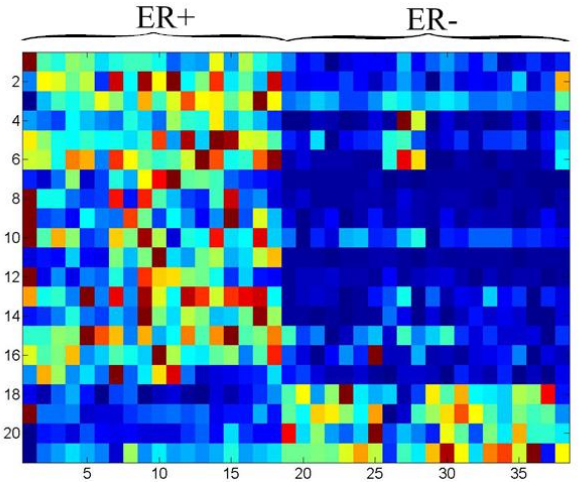

Fig. 2. Selected 21 genes from Brest cancer dataset

6 Conclusions

A new Wilcoxon rank sum test and SVM based gene selection scheme is proposed in this paper. Our method has been tested on Breast cancer and Leukemia data. The effects of different parameter settings on the classification performance are analyzed. And the subsets of the most informative genes are listed and their expression levels are sketched, which show that the selected genes can make clear separation between two classes. The comparisons with other existed methods show that the presented method outperforms most of others.

Acknowledgement

This paper is supported by the National Nature Science Foundation of China (No. 6040204), Program for New Century Excellent Talents in University, and the Excellent Youth Foundation of Hunan Province (06JJ1010).

References

1. Liu, H., Li, J., Wong, L.: A Comparative Study on Feature Selection and Classification Methods using Gene Expression Profiles and Proteomic Patterns. Genome Information 13, 51–60 (2002)
2. Liu, H., Li, J., Wong, L.: Selection of Patient Samples and Genes for Outcome Prediction. In: Proceedings of the IEEE Computational Systems Bioinformatics Conference, Stanford, pp. 382–392. IEEE Computer Society Press, Los Alamitos (2004)
3. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge (2000)

4. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46, 389–422 (2002)
5. Zhang, X., Wong, W.: Recursive Sample Classification and Gene Selection Based on SVM: Method and Software Description. In: Technical Report, Department of Biostatistics, Harvard School of Public Health, USA (2001)
6. Furlanello, C., Serafini, M., Merler, S., Jurman, G.: An Accelerated Procedure for Recursive Feature Ranking on Microarray Data. Neural Networks 16, 641–648 (2003)
7. Tang, Y., Zhang, Y., Huang, Z.: FCM-SVM-RFE Gene Feature Selection Algorithm for Leukemia Classification from Microarray Gene Expression Data. In: Proceedings of the IEEE International Conference on Fuzzy Systems, pp. 97–101. IEEE Computer Society Press, Los Alamitos (2005)
8. Duan, K., Rajapakse, J.C.: A Variant of SVM-RFE for Gene Selection in Cancer Classification with Expression Data. In: Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, San Diego, pp. 49–55. IEEE Computer Society Press, Los Alamitos (2004)
9. Duan, K., Rajapakse, J.C., Haiying, W., Azuaje, F.: Multiple SVM-RFE for Gene Selection in Cancer Classification with Expression Data. IEEE Transactions on Nanobioscience 4, 228–233 (2005)
10. West, M., Blanchette, C., Dressman, H., et al.: Predicting the Clinical Status of Human Breast Cancer Using Gene Expression Profiles. In: Proceedings of the National Academy of Science, vol. 98, pp. 11462–11467 (2001)
11. Golub, T., Slonim, D., Tamayo, P., et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 28, 531–537 (1999)
12. Krishnapuram, B., Carin, L., Hartemink, A.: Gene expression analysis: Joint Feature Selection and Classifier Design. In: Schölkopf, B., Tsuda, K., Vert, J-P (eds.) Kernel Methods in Computational Biology, Schölkopf, B, pp. 299–317. MIT Press, Cambridge, MA (2004)
13. Ben-Dor, A., Bruhn, L., Friedman, N. et al.: Tissue Classification with Gene Expression Profiles. Journal of Computational Biology 7, 559–583 (2000)
14. Li, Y., Campbell, C., Tipping, M.: Bayesian Automatic Relevance Determination Algorithms for Classifying Gene Expression Data. Bioinformatics 18, 1332–1339 (2002)
15. Figueiredo, M., Jain, A.: Bayesian Learning of Sparse Classifiers. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, Wisconsin, pp. 35–41 (2001)