Simultaneous Sample and Gene Selection Using T-score and Approximate Support Vectors

Piyushkumar A. Mundra[1], Jagath C. Rajapakse[1,2,3], and D.A.K. Maduranga[1]

[1] Bioinformatics Research Center, School of Computer Engineering, Nanyang Technological University, Singapore
[2] Singapore-MIT Alliance, Singapore
[3] Department of Biological Engineering, Massachusetts Institute of Technology, USA
asjagath@ntu.edu.sg

Abstract. T-score, based on *t*-statistics between samples and disease classes, is a widely used filter criterion for gene selection from microarray data. However, classical T-score uses all the training samples but for both biological and computational reasons, selection of relevant samples for training is an important step in classification. Using a modified logistic regression approach, we propose a sample selection criterion based on T-score and develop a backward elimination approach for gene selection. The method is more stable and computationally less costly compared to support vector machine recursive feature elimination (SVM-RFE) methods.

Keywords: data point selection, gene selection, instance selection, logistic regression.

1 Introduction

Gene selection is a vital step in the analysis of microarray gene-expression data and several approaches have been proposed earlier [1–8]. The methods of gene selection can be broadly categorized into filter, wrapper, or embedded methods. Filter methods are simple and computationally efficient, but have lower performance than the other methods. T-score based on *t*-statistics measuring correlation between input features and output class labels is commonly used as filter criterion for sample classification [1]. Other popular filter methods include Relief [9], correlation based feature selection [10], minimum redundancy maximum relevancy [11]. For more details on filter methods, readers are referred to [2]. However, in classical filter approaches, all the training samples are used for gene ranking while ignoring the relevance and quality of data samples.

On the other hand, popular wrapper and embedded methods include Support Vector Machine Recursive Feature Elimination [5] and its variants [3, 12–16], random forest-RFE [17], elastic net [18] etc. All these methods predominantly use classifier performance in ranking genes. Many classifiers, such as support vector machines (SVM), boosting algorithms, and logistic regression etc. indicate

A. Ngom et al. (Eds.): PRIB 2013, LNBI 7986, pp. 79–90, 2013.
© Springer-Verlag Berlin Heidelberg 2013

that all samples in a training data may not be equally relevant for the classification task [19, 20]. Removal of samples (or data points) that do not provide useful information for classification improves the performance. In microarray analysis, due to heterogeneity of tissues and cell assays, the datasets are inherently multimodal [21] and therefore qualities of samples vary. Using the classical theory of margin of classifier [19], sample points could be classified into three types: within the margin, on the margin, and away from the margin. For a classification task, various theories, including SVM and boosting techniques, suggest that the points on the margin and within the margin are more important than the samples away from the margin. Giving more importance to samples on or within the margin boundary may reduce the error variance in feature selection [22]. Earlier, the importance of selection of sample in active feature selection and dimensionality reduction was demonstrated using *kd*-tree algorithm [23, 24]. A genetic algorithm/k-nearest neighbour based approach was proposed for simultaneous selection of samples and metabolomic features [6]. Similarly, a modified particle swarm algorithm was combined with SVM for simultaneous sample selection and gene ranking [25]. Very recently, sample weighting based gene selection algorithm was proposed where sample weights are determined according to its influence to the estimation of feature relevance [26].

Along with better classification, a method of identification of true markers should be reproducible (stable) with respect to variations of the samples [16, 27]. Instability of a gene ranking casts doubts over computational results and hence does not give confidence for further biological validation. Stability of a gene selection method depends on many factors which includes sample size, treatment to correlative structure and underlying data distribution. However, an improvement in stability should not decrease the accuracy of sample.

Recently, predictive performance, stability and functional interpretability of 32 gene selection methods were analysed on 4 breast cancer datasets and results indicate that a simple Student's t-test (similar to T-score) performs the best [28]. However, the issue of relevant samples still persists. In our previous work, we decomposed T-score into two parts corresponding to relevant samples and non-relevant samples to show the importance of sample selection in T-score. And thereby a support vector based t-score recursive feature elimination (SVt-RFE) algorithm was proposed for feature selection [29, 30]. However, this algorithm uses SVM to select the samples and hence is computationally expensive. It also suffers from low stability. In this paper, we propose a gene selection method to improve stability and computational complexity of SVt-RFE and SVM-RFE methods without compromising on the performance of classification. To do so, we propose an efficient sample selection criterion to identify relevant samples by incorporating a modified logistic regression model, similar to SVM, using T-score as the selection criterion. A backward elimination approach is then proposed to iteratively select the relevant genes and achieve better classification accuracy than the existing methods. Our analysis indicates that the proposed algorithm improves the stability compared to SVM based approaches.

2 Method

Suppose $D = \{x_{ij}\}_{i=1,j=1}^{n,m}$ denotes a microarray dataset of m gene expression samples obtained on n genes where x_{ij} is the expression of gene i gathered in sample j. The vector $x_j = (x_{ij})_{i=1}^{n}$ denotes gene expressions on sample j and $x_i = (x_{ij})_{j=1}^{m}$ denotes the expressions of gene i across all the samples. Let two-class classification of sample j be $y_j \in \ell = \{+1, -1\}$ taking values +1 or -1 for cancerous (+ class) or benign (- class), respectively.

2.1 T-score

T-score is a ranking measure based on t-statistic between gene expressions and class labels. For gene i, T-score (r_i) is given by

$$r_i = \frac{\left|\mu_i^+ - \mu_i^-\right|}{\sqrt{\frac{m^+(\sigma_i^+)^2 + m^-(\sigma_i^-)^2}{m^+ + m^-}}} \tag{1}$$

where superscript + and − denotes positive and negative classes, respectively. The m^+, μ_i^+ and σ_i^+ represents the number of samples, the mean and standard deviation of expression values of gene i in samples of the positive class respectively. Similarly, m^-, μ_i^- and σ_i^- are defined for negative class. Higher the ranking value, more important the gene for separation of the classes is [1].

T-score is an easy and fast measure to compute as it assumes independence among genes and normality of data. However, many a times this method gives a stable gene subset which performs poor in classification compared to wrapper and embedded methods because it does not take into account the characteristics of the classifier in the ranking of genes. One way to improve the performance of this criterion is to select relevant samples when computing the T-score [29].

2.2 Efficient Sample Selection Technique

The margin of separation of SVM is defined by the support vectors or the samples on the margin. The support vectors are the samples that in fact define the discriminant function. Use of only the support vectors for gene selection was earlier demonstrated in support vector machine recursive feature elimination (SVM-RFE) method [29]. In this section, an efficient method to select samples (approximate support vectors) is proposed for gene selection. *Relevant samples* refers to those on and within the margin of separation. Using SVM, determining the margin of separation in two-class sample classification has a computational complexity of $O(\max(n, m)m^2)$. This becomes even more costly for SVM-RFE as each iteration needs retraining the SVM. Therefore, there is a need for a simpler model selecting samples on and within the margins, which is computationally inexpensive and gives a good biological interpretability.

An approximate loss function for SVM using concepts of logistic regression was proposed by Zhang *et al.* [31]. This function uses a sequence of smooth

functions for iterations to uniformly converge to SVM objective function. The approximate loss function L is given by

$$\mathcal{L}(x, y : w) = \frac{1}{\lambda}\ln\left(1 + \exp\left(-\lambda\left(yw^T x - 1\right)\right)\right) \tag{2}$$

where λ is a tuning parameter and w denotes the weights determining the discriminant. Instead of using a standard 0-1 loss function in SVM, the use of (2) leads to the following penalized objective function:

$$\begin{aligned}\mathcal{L}_P(x, y : w) = &\sum_{j=1}^{m}\frac{1}{\lambda}\ln\left(1 + \exp\left(\lambda\left(1 - y_j w^T x_j\right)\right)\right) \\ &+ \eta\left\|w\right\|^2\end{aligned} \tag{3}$$

where η denotes the sensitivity parameter.

Setting the partial derivation of (3) with respect to each gene i to zero,

$$\begin{aligned}w &= \sum_{j=1}^{m}\frac{1}{2\eta}\frac{\exp\left(\lambda\left(1 - y_j w^T x_j\right)\right)}{1 + \exp\left(\lambda\left(1 - y_j w^T x_j\right)\right)} y_j x_j \\ &= \sum_{j=1}^{m}\alpha_j x_i y_i\end{aligned} \tag{4}$$

Like in SVM, the multiplication factor α_j to $y_j x_j$ incorporates the margin information while computing weights. For example, if margin $y_j w^T x_j$ is greater than one, the multiplication factor becomes zero for large value of λ. In a sense, it rejects the contribution of that particular sample point. Hence, based on this property and considering that $\frac{1}{2\eta}$ is a multiplicative factor, we propose following approximation of support vectors:

$$\alpha_j = \frac{\exp\left(\lambda\left(1 - y_j w^T x_j\right)\right)}{1 + \exp\left(\lambda\left(1 - y_j w^T x_j\right)\right)} \tag{5}$$

With respect to SVM-RFE, the standard 0-1 Loss function gives following SVM weight vector [5, 19]

$$w = \sum_{j=1}^{m}\alpha_j^* y_j x_j \tag{6}$$

Comparing (4),(5) and (6), we can represent the SVM induced weight to a particular sample point α_j^* with α_j.

2.3 T-score with Sample Selection (T-SS)

The margin of a data point is defined as the distance from the data point to the discriminant boundary. The margin of jth data sample is given by the term $y_j w^{\mathrm{T}} x_j$. Zhang *et al.* proposed a gradient descent algorithm to determine the

Algorithm 1. Gene ranking using T-score and sample selection

Begin
Gene set $S = \{i\}_{i=1}^{n}$, data D, and ranked list $R = [\,]$;
set λ; $\epsilon = 0.001$
repeat
  Find the set of samples $M \subset D$ using (5) with $\alpha_j > \epsilon$
  if $|M| < 2$ **then**
    $M = D$
  end if
  Compute the ranking r_i using samples in M
  Select the gene $i^* = \arg\min\{r_i\}$
  Update $R = [R; i^*]\,; S = S \setminus \{i^*\}$
until all genes are ranked
end : output R

margin of separation [31]. In order to simplify the computations, we propose to use T-score of each gene as the selection criteria and thereby remove the optimization step in (3). With this idea, we propose an algorithm for simultaneous sample and gene selection, which is described in Algorithm 1.

Sample points are selected using (5) with a small threshold ϵ. Let M_ℓ denote the set of selected sample points in class ℓ. With a given λ value, the samples are selected using the margin information based on T-score. Using only the selected samples, genes are ranked with T-score. A gene with the least absolute score is then removed from the gene set and the whole process is iterated again until all genes are ranked. In other words, the proposed method selects genes in backward elimination manner while selecting the relevant samples. The T-score with sample selection method fails whenever there is less than two relevant data points. In such cases, we revert to all the sample points and compute the ranking scores in that iteration using all training samples.

The margin is determined by using the T-score of individual gene. It has a direct relation with log-odds ratio if the data is normally distributed, which is given by

$$\log \frac{P\,(+|x)}{P\,(-|x)} = x^T \Sigma^{-1} \left(\mu^+ - \mu^-\right) + w_0 \tag{7}$$

Here, w_0 is a bias term and was computed using

$$w_0 = log \frac{\pi^+}{\pi^-} - \frac{1}{2}\mu^{+T} \Sigma^{-1} \mu^+ + \frac{1}{2}\mu^{-T} \Sigma^{-1} \mu^- \tag{8}$$

where π^+ and π^- represent the prior probabilities of respective classes; the Σ represents the covariance matrix. As we assume independence among genes, in (7) and (8), the covariance matrix becomes $\Sigma = \sigma^2 \mathrm{I}$, where I is the identity matrix. In computing a sample margin $yw^{\mathrm{T}}x$, the bias term is included in w. The weights w are normalized before computing the margin of a sample:

$$w = \frac{w}{\|w\|} \tag{9}$$

Table 1. Details on Benchmark Gene Expression Datasets

Dataset	No. of Genes	No. of + Samples	No. of - Samples
Colon	2000	22	40
Leukemia	7129	38	34
Breast	7129	25	47

3 Experiments and Results

3.1 Datasets and Preprocessing

To evaluate the performance of the proposed method, we performed extensive experiments on three benchmark microarray gene expression datasets, namely, colon [32], leukemia [33], and breast cancer [34]. The details of these widely used datasets for evaluating gene ranking methods are given in Table 1.

All the training datasets were normalized to zero mean and unit variance based on gene expressions of a particular gene to implement T-score, SVM-RFE, SVt-RFE, and T-SS. The datasets were normalized using the parameters from the corresponding training dataset only.

3.2 Parameter Estimation

The parameter λ was determined from a set of $\{1, 3, 5, 7, 10\}$ and selected for the best classification accuracy with the selected genes from Algorithm 1. For algorithms like SVt-RFE and SVM-RFE, the selection of training data points depends on the sensitivity η of the linear SVM, which was determined from the finite set $\left\{2^{-20}, 2^{-19}, \ldots, 2^{15}\right\}$, giving the maximum Matthew's correlation coefficient (MCC[1]) on 10-fold cross-validation.

In recursive elimination, we gradually removed genes in each of the iteration. To increase the speed of the numerical simulations with SVt-RFE, SVM-RFE, and T-SS, the following step-wise strategy was employed:

$$\text{No. of genes removed} = \begin{cases} 100 \text{ if } n' \geq 10000 \\ 10 \quad \text{if } 1000 \leq n' < 10000 \\ 1 \quad n' < 1000 \end{cases} \tag{10}$$

where n' is the number of genes in the gene set.

3.3 Performance Evaluation

With five-fold external cross-validation for 20 times, we obtained $B = 100$ sets of gene ranking lists for each dataset. The gene ranking was obtained using T-score, SVM-RFE, SVt-RFE, and T-SS. The test validation was performed using

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

corresponding test set of a gene ranking list. We tested gene subsets starting from the top ranked genes and then successively adding one gene at a time in test subset till the total number of genes in subset equals 100. The performance measures such as accuracy, sensitivity, and specificity were averaged over those 100 trials. The cardinality of the gene subset giving the minimum average test error was reported as the number of genes corresponding to the best classification performance. We also performed pair-wise one-sided t-test to determine if the performance of the T-SS is significantly better over the other methods.

3.4 Stability Analysis

In this section, a similarity based approach is taken to compute the stability of gene selection, which is measured by the average over all pair-wise similarity comparisons among all the ranked gene lists obtained by the method over different subsets of training samples [35].

Let $\left\{D^b\right\}_{b=1}^{B}$ be a set of B sub-samplings of the dataset of the same size and R^b be the b-th rank list of genes. The stability $\mathcal{S}_{\mathcal{D}}$ of the method over the dataset D is given by

$$\mathcal{S}_D = \frac{2}{B(B-1)} \sum_{b=1}^{B} \sum_{b'=b+1}^{B} \mathcal{S}(R^b, R^{b'}) \tag{11}$$

where $\mathcal{S}(R^b, R^{b'})$ is a similarity measure between the gene rank lists R^b and $R^{b'}$ for top n^* genes in both lists. One of the popular measure to find similarities between two gene lists is a Kuncheva index [35] given by

$$\mathcal{S}(R^b, R^{b'}) = \frac{|R^b \bigcap R^{b'}| - {n^*}^2/n}{n^* - {n^*}^2/n} \tag{12}$$

where n denotes total number of genes in a dataset and n^* is the set of the top genes. Kuncheva index has a range between $[-1, 1]$ with large value indicating large number of common genes between the subsets. A negative Kuncheva index denotes an overlap between two subsets by chance. The term $(n^*)^2/n$ corrects for a bias due to chance of selecting common features between two randomly chosen subsets.

3.5 Redundancy Analysis

Apart from stability and performance in classification, we further evaluate gene selection methods for their ability to select non-redundant genes. We use the absolute value of Pearson's correlation coefficient to estimate the redundancy among top-ranked genes in a given dataset. In a gene rank list R^b, we first measure a pair-wise correlation coefficient of top n^* genes, resulting in a $n^* \times n^*$ correlation matrix with each element representing pair-wise similarity. Using the upper triangular matrix, we obtained average of absolute pair-wise correlations, which represents redundancy among those n^* top-ranked genes in rank list R^b.

Table 2. Performance of T-score, SVM-RFE, SV*t*-RFE, and T-SS on Benchmark *Cancer* Datasets

Dataset	Method	T-score	SVM-RFE	SV*t*-RFE	T-SS
Colon	# Genes	83	97	**32**	91
	Accuracy	86.53 ± 9.00	83.47 ± 9.37	86.08 ± 9.44	$\mathbf{87.12 \pm 9.59}$
	Significance	...	$p < 0.001$	...	
	Sensitivity	80.10 ± 19.08	74.35 ± 18.97	78.25 ± 17.84	$\mathbf{80.90 \pm 18.44}$
	Significance	...	$p < 0.001$	$p < 0.05$	
	Specificity	90.00 ± 10.05	88.50 ± 11.47	90.37 ± 10.02	$\mathbf{90.50 \pm 10.22}$
	Significance	...	$p < 0.001$	...	
Leukemia	# Genes	65	**43**	63	49
	Accuracy	96.36 ± 4.72	96.65 ± 4.14	$\mathbf{97.01 \pm 4.09}$	97.00 ± 4.10
	Significance	$p < 0.05$	...	...	
	Sensitivity	94.40 ± 11.04	95.00 ± 9.16	95.20 ± 9.04	$\mathbf{95.40 \pm 8.46}$
	Significance	...	...	...	
	Specificity	97.43 ± 4.83	97.54 ± 4.52	$\mathbf{97.99 \pm 4.18}$	97.88 ± 4.50
	Significance	...	...	...	
Breast	# Genes	**57**	97	87	99
	Accuracy	86.17 ± 11.78	87.67 ± 11.02	87.97 ± 11.35	$\mathbf{89.30 \pm 10.77}$
	Significance	$p < 0.001$	$p < 0.01$	$p < 0.05$	
	Sensitivity	88.80 ± 13.13	$\mathbf{91.20 \pm 13.43}$	89.80 ± 13.48	89.20 ± 12.85
	Significance	...	...	...	
	Specificity	83.45 ± 19.07	84.15 ± 17.99	86.05 ± 18.82	$\mathbf{89.45 \pm 15.42}$
	Significance	$p < 0.001$	$p < 0.001$	$p < 0.01$	

This value is averaged over the total number of gene rankings, i.e., number of bootstrapped trials (B).

Mathematically, the average redundancy among top n^* genes over B trials can be given by,

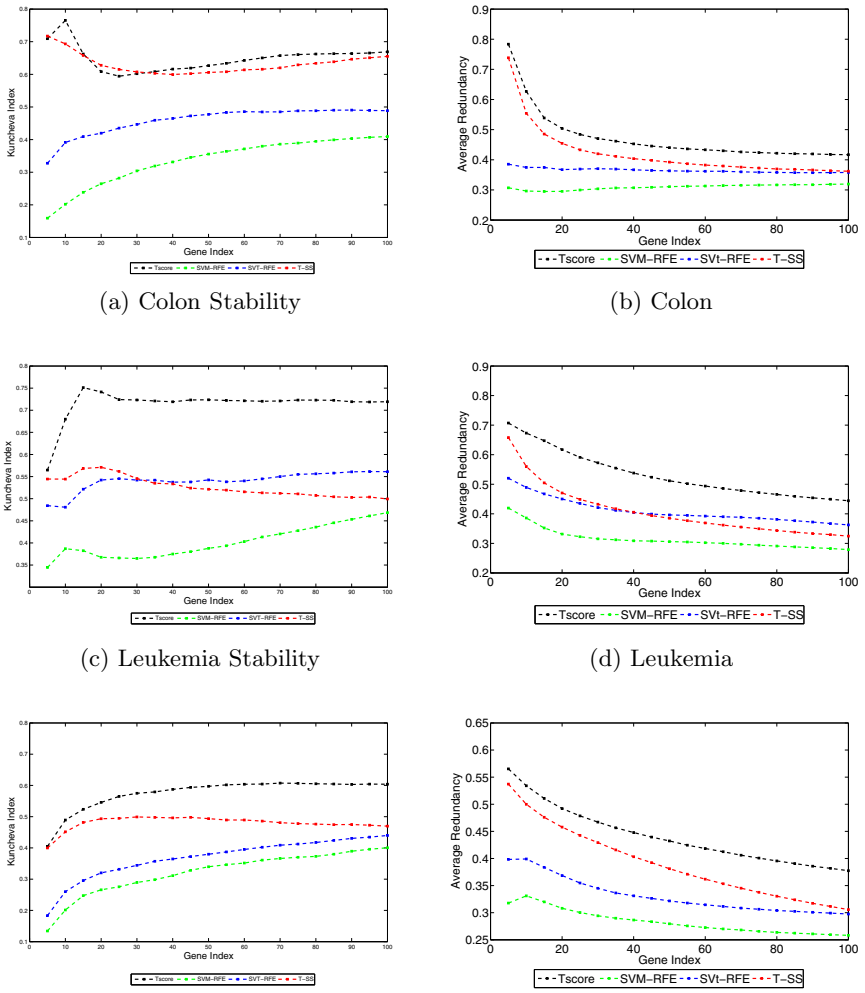
$$\overline{Q} = \frac{1}{B} \sum_{b=1}^{B} \frac{2}{n^* (n^* - 1)} \sum_{i=1}^{n^*-1} \sum_{i'=i+1}^{n^*} |\rho (x_i, x_{i'})| \tag{13}$$

where $|\rho (x_i, x_{i'})|$ is absolute value of Pearson's correlation coefficient between expression values of gene i and i'. In a given dataset, the redundancy analysis is performed over top 100 genes, obtained from various ranking methods.

3.6 Results

A comparison of classification performances of T-score, SVM-RFE, SV*t*-RFE, and T-SS is shown in Table 2. The p-values shown gives the statistical significance of superior performance of T-SS over the other methods. The stability and redundancy plots are depicted in Figure 1.

As seen, the performance of the proposed method is significantly better than the gene ranking by T-score and SVM-RFE methods in at least two datasets. For



(a) Colon Stability (b) Colon

(c) Leukemia Stability (d) Leukemia

(e) Breast Stability (f) Breast

Fig. 1. Plots of Stability and Redundacy against the number of selected genes on benchmark expression datasets

breast cancer dataset, the proposed algorithm outperforms all the other methods. Importantly, the stability plots shows that the proposed method is more stable than SVM-RFE and SVt-RFE methods for top-ranked genes. Comparing redundancy, T-score gave highly redundant top-ranked genes while SVM-RFE returned the least redundant genes. Genes from T-SS and SVt-RFE methods have intermediate redundancy. The numbers of genes selected by T-SS were higher in most cases.

4 Discussion and Conclusion

This paper proposed a sample selection criterion using a modified logistic regression loss function and a backward elimination based gene ranking algorithm. The method selects sample points iteratively before ranking genes using the T-score in each iteration. The performance was evaluated on a number of benchmark datasets and results showed not only promise in the classification results but also the superior stability of the method.

For selection of samples, the approaches involving standard SVM, such as SVt-RFE, are computationally expensive and involves computational complexity of the order of $O(\max(n, m)m^2)$ [36]. If a single gene is removed in each iteration, we need to train SVM for n number of times. On the other hand, selecting samples on the margin with T-score by using an approximation to SVM lost function, the speed is improved. A standard T-score have computational complexity of order of $O(nm)$ [36], so our proposed algorithm is approximate to this complexity compared to SVM-RFE and SVt-RFE.

In two-class classification, as the standard T-score ranks genes independently, it has the highest stability and no penalization for redundancy in gene selection among the other methods tested in the experiments. As SVM-RFE does not treat genes independently and penalizes for redundant genes [37], it is less stable and robust to the variations of training samples. The proposed method not only performs better in classification but retains independence among genes while ranking. This leads to better stability compared to SVM based approaches. Following [37], the proposed sample selection criterion may induce some penalization for the redundant genes. This is evident with reduced redundancy compared to T-score method.

In conclusion, the proposed method is a simple yet efficient criterion for sample selection. Simultaneous sample and gene selection algorithms significantly outperform both T-score and SVM-RFE methods on at least two benchmark datasets. Along with better classification, the proposed method was computationally efficient and highly stable. This suggests that sample selection indeed plays an important role in gene selection. As future of this work, one may want to penalize for redundancy among genes in the cost function as it would improve the stability and performance of tissue classification.

Acknowledgments. This work is supported by a AcRF Tier 2 grant MOE2010-T2-1-056 (ARC 9/10), Ministry of Education, Singapore.

References

1. Inza, I., Larranaga, P., Blanco, R., Cerrolaza, A.: Filter versus wrapper gene selection approaches in dna microarray domains. Artificial Intelligence Medicine 31, 91–103 (2004)
2. Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H., Nowe, A.: A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9(4), 1106–1119 (2012)

3. Mundra, P.A., Rajapakse, J.C.: Svm-rfe with mrmr filter for gene selection. IEEE Transactions on Nanobioscience 9(1), 31–37 (2010)
4. Rajapakse, J.C., Mundra, P.A.: Multiclass gene selection using pareto-fronts. IEEE/ACM Transactions on Computational Biology and Bioinformatics (accepted, 2013)
5. Guyon, I., Weston, J., Barhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422 (2002)
6. Cavill, R., Keun, H., Holmes, E., Lindon, J., Nicholson, J., Ebbels, T.: Genetic algorithms for simultaneous variable and sample selection in metabonomics. Bioinformatics 25(1), 112–118 (2009)
7. Chakraborty, S.: Simultaneous cancer classification and gene selection with bayesian nearest neighbor method: An integrated approach. Computational Statistics & Data Analysis 53(4), 1462–1474 (2009)
8. Hapfelmeier, A., Ulm, K.: A new variable selection approach using random forests. Computational Statistics & Data Analysis 60, 50–69 (2013)
9. Kira, K., Rendell, L.A.: A feature selection problem: traditional methods and a new algorithm. In: Proc. of the 10th National Conference on Artificial Intelligence, pp. 129–134 (1992)
10. Wang, Y., Tetko, I., Hall, M., Frank, E., Facius, A., Mayer, K., Mewes, H.: Gene selection from microarray data for cancer classification - a machine learning approach. Computational Biology and Chemistry 29, 37–46 (2005)
11. Ding, C., Peng, H.: Minimum redundancy feature selection from microarray gene expression data. J Bioinformatics Computational Biology 3, 185–205 (2005)
12. Tang, Y., Zhang, Y.Q., Huang, Z.: Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. IEEE Trans on Computational Biology and Bioinformatics 4(3), 365–381 (2007)
13. Tang, Y., Zhang, Y.Q., Huang, Z., Hu, X., Zhao, Y.: Recursive fuzzy granulation for gene subset extraction and cancer classification. IEEE Trans on Information Technology in Biomedicine 12(6), 723–730 (2008)
14. Kai-Bo, D., Rajapakse, J., Wang, H., Azuaje, F.: Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Trans Nanobioscience 4, 228–234 (2005)
15. Yoon, S., Kim, S.: Adaboost-based multiple svm-rfe for classification of mammograms in ddsm. BMC Medical Informatics and Decision Making 9(S1), 693–708 (2009)
16. Abeel, T., Helleputte, T., Van de Peer, Y., Sayes, Y., et al.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26(3), 392–398 (2010)
17. Diaz-Uriarte, R., Andres, S.: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3 (2006)
18. Zou, H., Hastie, T.: The regularization and variable selection via the elastic net. J. Royal Stat. Society B 67, 301–320 (2005)
19. Vapnik, V.N.: Statistical Learning Theory. Wiley-Interscience Publications (1998)
20. Freund, Y., Schapire, R.: A short introduction to boosting. J. Japanese Society for Artificial Intelligence 14(5), 771–780 (1999)
21. Clarke, R., Ressom, H., Wang, A., Xuan, J., et al.: The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature Reviews Cancer 8, 37–49 (2008)
22. Han, Y., Yu, L.: A variance reduction framework for stable feature selection. In: Proc. of the 10th IEEE International Conference on Data Mining (2010)

23. Liu, H., Motoda, H., Yu, L.: A selective sampling approach to active feature selection. Artificial Intelligence 159, 49–74 (2004)
24. Pechenizkiy, M., Puuronen, S., Tsymbal, A.: The impact of sample reduction on PCA-based feature extraction for supervised learning. In: Proc. of the 21st ACM Symposium on Applied Computing, pp. 553–558 (2006)
25. Shen, Q., Mei, Z., Ye, B.X.: Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification. Computers in Biology and Medicine 39, 646–649 (2009)
26. Lei, Y., Yue, H., Berens, M.: Stable gene selection from microarray data via sample weighting. IEEE Transactions on Computational Biology and Bioinformatics 9(1), 262–272 (2012)
27. Somol, P., Novovicova, J.: Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Transactions on Pattern Analysis and machine intelligence 32(11), 1921–1939 (2010)
28. Haury, A.C., Gestraud, P., Vert, J.P.: The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. Plos One 6(12), e28210 (2011)
29. Mundra, P.A., Rajapakse, J.C.: Gene and sample selection for cancer classification with support vectors based t-statistic. Neurocomputing 73(13-15), 2353–2362 (2010)
30. Mundra, P.A., Rajapakse, J.C.: Support vector based T-score for gene ranking. In: Chetty, M., Ngom, A., Ahmad, S. (eds.) PRIB 2008. LNCS (LNBI), vol. 5265, pp. 144–153. Springer, Heidelberg (2008)
31. Zhang, J., Jin, R., Yang, Y., Hauptmann, A.: Modified logistic regressionl an approximation to svm and its applications in large-scale text categorization. In: Proceedings of 20th International Conference on Machine Learning, ICML 2003 (2003)
32. 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)
33. 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)
34. West, M., Blanchette, C., Dressman, H., et al.: Predicting the clinical status of human breast cancer by using gene expression profiles. Proceedings of National Academy of sciences 98(20), 11462–11467 (2001)
35. Kuncheva, L.: A stability index for feature selection. In: Proceedings of the 25th IASTED International Conference on Artificial Intelligence and Applications, pp. 390–395 (2007)
36. Guyon, I., Elisseeff, A.: An introduction to feature extraction. In: Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. (eds.) Feature Extraction, Foundations and Applications. STUDFUZZ, pp. 1–25. Springer, Heidelberg (2006)
37. Li, F., Yang, Y.: Analysis of recursive gene selection approaches from microarray data. Bioinformatics 21(19), 3741–3747 (2005)