A Hybrid of SVM and SCAD with Group-specific Tuning Parameters in Identification of Informative Genes and Biological Pathways

Muhammad Faiz Misman[1], Weng Howe Chan[1], Mohd Saberi Mohamad[1*],
Safaai Deris[1]

[1]Artificial Intelligence and Bioinformatics Research Group, Faculty of Computing,
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
faizmisman@gmail.com, whchan2@live.utm.my,{saberi,safaai}@utm.my

Abstract. Advancements in pathway-based microarray classification approach leads to a new era of genomic research. However, this approach is limited by issues regarding the quality of the pathway data as these data are usually curated from biological literatures and in specific biological experiment (e.g. lung cancer experiment), context free pathway information collection process takes place leading to the presence of uninformative genes in the pathways. Many methods in this approach neglect these limitations by treating all genes in a pathway as significant. In this paper, we propose a hybrid of support vector machine and smoothly clipped absolute deviation with group-specific tuning parameters (gSVM-SCAD) to select informative genes within pathways before the pathway evaluation process. Experiments conducted on gender and lung cancer datasets shows that gSVM-SCAD obtains significant results in identifying significant genes and pathways, and in classification accuracy.

Keywords: Pathway analysis, smoothly clipped absolute deviation, support vector machines, gene selection

1 Introduction

Incorporation of prior pathway data into microarray analysis has become a popular research area in bioinformatics due to the advantages in providing further biological interpretation compare to single gene microarray analysis. Such advantages further spurred the development of various approaches to identify informative genes and pathways that contribute to the certain cellular processes. The goal of the pathway-based microarray analysis is to identify significant pathways and also genes within the pathways that contribute to the phenotypes of interest. This is in contrast to single gene microarray analysis that identifies only the significant genes. Two most common approaches in pathway-based microarray analysis are enrichment analysis approaches (EA) and machine learning approaches (ML) [1].

However, there are some challenges in pathway-based microarray analysis such as the quality of pathway data where some of the uninformative genes maybe included into pathways while informative genes being excluded [2]. In order to deal with this challenge, researchers attempt to make improvement by removing unaltered genes in pathways [1] and include additional functional interpretation in EA approaches [3]. While for ML approaches, gene selection methods have been included to select informative genes within a pathway before the classification model building instead of including all the genes within a pathway into the model building [2, 4, 5].

However, EA approaches considered all genes within pathways as equally important [1]. Alternatively, ML can select the only important genes within a pathway by including the gene selection method. In contrast to EA, ML aim to identify both relevant genes and pathways that related to the phenotypes of interest. Therefore, ML can bring more insight in a biological perspective. ML is used in this research due to its advantages. However, there are arguments against incorporating gene selection methods in ML where informative genes may be discarded [1]. This is due to the nature of microarray data where it can impose sparseness and biasness on the penalty function that act as the gene selection method in evaluating the informative genes [6]. Therefore, the efficient and robust gene selection technique is needed in order to deal effectively with the problems arise in pathway-based microarray analysis.

Following the good results obtained from support vector machines (SVM) in classifying gene expression data, hybrid of SVM with smoothly clipped absolute deviation (SCAD) penalty was produced, named as SVM-SCAD [7]. SCAD provides nearly unbiased coefficient estimation and select the important genes consistently compared to other popular penalty function such as least absolute shrinkage and selection operator (LASSO) [8]. SVM-SCAD had proved its ability in selecting the informative genes and the method is comparable to LASSO penalty function. Hence, in order to identify both significant genes and pathways that related to phenotypes of interests, this paper proposed an improved of SVM-SCAD with group-specific tuning parameters, termed as gSVM-SCAD.

2 SVM-SCAD and the proposed method (gSVM-SCAD)

2.1 SVM-SCAD

Given a data set *{(xi,yi)}, yi ϵ {-1,1}* is the sample tissue with possible two classes y_i = *-1* and y_i = *1* for each data set used in this paper, while $x_i = (x_{i1}, \dots ,x_{id}) \epsilon R^d$ represents the input vector of expression levels of d genes of the *i*-th sample tissue. SVM is a large margin classifier which separates classes of interest by maximizing the margin between them using the kernel function [7, 9]. This has been widely used especially in microarray classification area [10]. SVM distinguish input variables into its classes by a margin of

$$min_{\beta,c} \sum [1 - y_i f(x_i)]_+ + pen_\lambda(\beta) \tag{1}$$

$[1-y_i f(x_i)]+$ is the SVM convex hinge loss function where

$$[1 - y_i f(x_i)]_+ \leq \lambda f(x_i) + (1 - \lambda) f(y_i)$$

while penλ(β) is the penalty function with parameters λ, where β = (β1,…., βi) are the coefficients of the hyperplane, while c is the intercept of the hyperplane. Hinge loss function is a commonly used loss function in SVM in order to keep the fidelity of the resulting model to the data set [11]. However, the standard SVM can suffer from irrelevant data, since all the variables are used for constructing the classifier [7]. This is due to the usage of the L2 penalty in a soft-thresholding function for the common SVM. The detailed applications of L2 penalty in a soft-thresholding function and its drawbacks in identifying noises can be obtained from [7].

A penalty function is usually used as a variable selection in the statistics, and in bioinformatics it is called as gene selection. SCAD is different from other popular penalty functions such as LASSO, also called the L1 penalty [8], this is because SCAD provides nearly unbiased coefficient estimation when dealing with large coefficients. This is contrary to other penalty functions that usually increase the penalty linearly as the coefficient increases [6]. SCAD penalty has the form of

$$pen_{\lambda}(\beta) = \sum_{j=1}^{d} P_{\lambda}(\beta_j) \tag{2}$$

where $P_{\lambda}(\beta_j)$ is a penalty function with tuning parameter λ for β_j. For providing nearly unbiased, sparsity, and continuity estimate of β [7], the continuous differentiable penalty function is defined as

$$pen_{\lambda}(\beta_j) = \begin{cases} \lambda|\beta|, & if\ |\beta| \le \lambda \\ -(|\beta|^2 - 2a\lambda|\beta| + \lambda^2)/(2(a-1)), & if\ \lambda < |\beta| \le a\lambda \\ ((a+1)\lambda)/2, & if\ |\beta| > a\lambda \end{cases} \tag{3}$$

where a and λ are tuning parameters with a > 2 and λ > 0 [6]. For a tuning parameter a, previous research suggested the parameter a = 3.7 due to the minimal achievement in a Bayes risk [6]. Therefore, in this research a = 3.7 is used while λ is a tuning parameter obtained using generalized approximate cross validation (GACV) tuning parameter selection methods (as discussed latter).

In order to surmount the limitations of the SVM due to its inability to distinguish between noise and informative data, SVM-SCAD was proposed by replacing the L2 penalty in function (1) with (2), which takes the form

$$min_{\beta,c} \frac{1}{n} \sum [1 - y_i f(x_i)]_+ + \sum_{j=1}^{d} P_{\lambda}(\beta_j) \tag{4}$$

In order to select the informative genes, SVM-SCAD have to minimize the function (4) using the successive quadratic algorithm (SQA) and repeated for kth times until convergence, where k = 1,...,n. During the procedure, if $\beta_j^k < \epsilon$, the gene is considered as uninformative. Where β is the coefficient for the gene j in the kth iteration and ϵ is a preselected small positive thresholding value with ϵ = yi - f(xi).

2.2 Tuning Parameter Selection Method

In SCAD there are two tuning parameters namely *a* and *λ* that play an important role in determining an effective predictive model. The tuning parameter selection in SVM-

SCAD is used to estimate the nearly optimal λ in order to identify the effective predictive model for SCAD. In this paper, the generalized approximate cross validation (GACV) [12] is used to select the nearly optimal λ. The formula as given below:

$$GACV_\lambda = \frac{1}{n}\sum_{i=1}^{n}[1 - y_i f(x_i)]_+ + DF_\lambda \tag{5}$$

where n is a total number of samples, DF_λ is a degree of freedom where

$$DF_\lambda = \frac{1}{n}\left[\sum_{y_i f(x_{i\lambda})<--1} 2\,\frac{\alpha_{\lambda i}}{2n\lambda} \cdot \|K(.,x_i)\|_{Hk}^2 \;+\; \sum_{y_i f(x_{i\lambda})\epsilon[-1,1]} \frac{\alpha_{\lambda i}}{2n\lambda} \cdot \|K(.,x_i)\|_{Hk}^2\right]$$

where $\frac{\alpha_{\lambda i}}{2n\lambda} = \frac{f(x_{i\lambda})\,[y_i] - f(x_{i\lambda})[x]}{y_i - x}$ and $\|K(.,x_i)\|_{Hk}^2$ is the reproducing kernel hilbert space (RKHS) with SVM reproducing kernel *K*. If all samples in microarray data are correctly classified, then $y_i f(x_{i\lambda}) > 0$ and sum following 2 in DF_λ does not appear and $DF_\lambda = K(0,0)/n\gamma^2$ where γ is the hard margin of an SVM [12]. The nearly optimal tuning parameter λ is obtained by minimizing the error rate from the GACV.

2.3 The Proposed Method (gSVM-SCAD)

In SVM-SCAD, the magnitude of penalization of β is determined by two tuning parameters a and λ. Since a has been setup as 3.7 [6], there is only one parameter λ left that play an important role. In order to incorporate pathway or set of gene data, the gSVM-SCAD used group-specific parameters λ_j estimation. In this paper, there are *k* groups of genes where *k = 1...n,* each gene able to be in one or more pathways. We grouped the genes based on their pathway information from the pathway data. In order to provide the group-specific tuning parameters, we modified (2) to the form of

$$pen_{\lambda k}(\beta_j) = \sum_{j=1}^{d} P\lambda_k(\beta_j) \tag{6}$$

where by allowing each pathway to have its own parameter λ_k as in (6) instead of general λ in (2)*,* the genes within pathways can be selected and classified more accurately.

Table 1 illustrates the procedure of gSVM-SCAD. The procedure consisted of 3 main steps. In the first step, the genes in microarray data are selected and grouped based on their prior pathway information from the pathway data. This process repeated for each pathway in the pathway data and there is a possibility that some genes are not involved in any pathways. From this step, the new sets of gene expression data are produced to be evaluated by the SVM-SCAD. In step 2, each pathway is evaluated using the SVM-SCAD. This procedure is started with the tuning parameter selection (step 2.1) where in this research, the grid search is applied. According to previous research, the best λ can be obtained in the range of 0 < λ < 2, therefore the grid search ranges from 0.001 to 0.009, 0.01 to 0.09, and 0.1 to 1 are used. The GACV is used to estimate the error for each tuning parameter value from the grid search. The nearly optimal tuning parameter produces the minimum GACV error. In step 2.2, the genes in the pathway are evaluated using SVM-SCAD and the informative genes within pathway is selected and selected while the non-informative genes are excluded from the pathway. In step 2.3, the informative genes obtained are classified between phenotypes of interests using an SVM. The classification error from the selected genes for each pathway is calculated using 10-fold CV in step 3. Biological validation for top pathways is conducted using the information from the biological research databases.

Table 1. The gSVM-SCAD procedure

```
Input: GE: Gene expression data , PD: Pathway data , TP : Tuning parameter, λ
Output: SP: Significant pathways , IG: Informative genes
Begin
Step 1: Grouping genes based on their pathway information
            For $j=1$ to max number of pathways in PD do
                Find genes from GE that related to the pathway
                Select and assign the related genes as a one group
           End-for
   Step 2: Evaluate the pathways
     For $j=1$ to max number of pathways in PD do
         Step 2.1: Estimation of TP using a GACV
            For TP = 0.001 to 0.009 ,0.01 to 0.09 and 0.1 to 1 do
                    $GACV_\lambda = \frac{1}{n}\sum_{i=1}^{n}[1 - y_i f(x_i)_\lambda]_+ + DF_\lambda$
           End-for
  $\lambda = argmin_\lambda\{GACV(\lambda)\}$ // best TP produces minimum GACV error
        Step 2.2: Select the informative genes using the SVM-SCAD
            Let $\beta^k$ as the estimate of $\beta$ at step $k$ where $k = 0, \dots, n$
            The value of $\beta^0$ set by an SVM
          While $\beta^k$ not converge do
                 Minimizing the $\frac{1}{n}\Sigma[1-y_i f(x_i)]_+ + \Sigma^d_{j=1} P\lambda_k(\beta_j)$
                 $k = k + 1$
            If $\beta_j^k \leq \epsilon$ then
               The gene $j$ considered as non-informative and discarded
            End-if
        End-while
      Step 2.3: Classify the selected genes using an SVM
   Step 3: Calculate the classification error using a 10-fold CV
  End-for
End
```

There are several main differences between gSVM-SCAD and other current methods in ML approaches. Firstly, it provides the genes selection method to identify and select the informative genes that are related to the pathway and the phenotype of interest which provides more in biological aspect. Secondly, the penalty function SCAD is more robust when dealing with a high number of genes, and it selects important genes more consistently than other popular LASSO (L1) penalty function [13]. And lastly, with group-specific tuning parameters, the gSVM-SCAD provides more flexibility in choosing the best λ for each pathway so that each pathway can be assessed more accurately. Therefore, by selecting the informative genes within pathway, the gSVM-SCAD can be seen as the best method in dealing with pathway data quality problems in pathway-based microarray analysis.

3 Experimental data sets

The performance of the gSVM-SCAD is tested using two types of data, gene expression and biological pathway data. The role of biological pathway data is as a metadata or prior biological knowledge. Both gene expression and pathway data are the same as those used in previous research done by Pang and colleagues [14].

3.1 Gene expression data sets

In gene expression data, it consists of m samples and n gene expression levels. The first column of the data represents the name of genes while the next column represents the gene expression levels. The data set forming a matrix of $E = \{e_{i,j}\}_{mxn}$ where $e_{i,j}$ represents the expression level of the gene j in the tissue sample i. In this paper, two gene expression data sets are used: lung cancer, and gender. The information of the data sets is shown in Table 2.

Table 2. Gene expression data sets

Name	No. of samples	No. of genes	Class	Reference
Lung	86	7129	2(normal and tumor)	[15]
Gender	32	22283	2(male and female cells)	unpublished

3.2 Biological pathway data

Total 480 pathways are used in this research. 168 pathways were taken from KEGG and 312 pathways were taken Biocarta pathway database. In a pathway data set, the first column represents the pathway name while the second column represents the gene name.

4 Experimental results and discussion

In this paper, to evaluate the performance of the gSVM-SCAD, we used a 10-fold cross validation (10-fold CV) classification accuracy. The results obtained from the gSVM-SCAD are validated with the biological literatures and databases. Since the limited pages for this paper, we only chose the top five pathways with highest 10-fold CV accuracy from both data sets for biological validation (commonly applied with several authors such as Pang et al. (2006) and Wang et al. (2008)).

4.1 Performance evaluation

For the performance evaluation of the SCAD penalty function, the SCAD was compared with the popular L_1 penalty function by hybridizing it with an SVM classifier (L_1 SVM), obtained from R package named penalizedSVM [16]. The L_1 SVM also applied with group-specific tuning parameters to determine λ. This experiment was done intentionally to test the robustness of the SCAD penalty in identifying informative genes when dealing with large coefficients compare to the L1 method. Then the gSVM-SCAD was compared with the current SVM-SCAD with respect to one general parameter tuning for all pathways, the tuning parameters $\lambda = 0.4$ as used in previous research [7]. For comparison with other classification methods without any gene selection process, the gSVM-SCAD was compared with four classifiers that are without any penalty function or gene selection method. The classifiers are PathwayRF [14], neural networks, k-nearest neighbour with one neighbours (kNN), and linear discriminant analysis (LDA). The purpose of this comparisons is to show that not all genes in a pathway contribute to a certain cellular process. The results of the experiment are shown in Table 3.

Table 3. A comparison of averages of 10-fold CV accuracy from the top ten pathways with other methods

Method	Lung Cancer (%)	Gender (%)
gSVM-SCAD	**73.77**	**87.33**
L_1-SVM	55.14	80.76
SVM-SCAD	53.5	77.96
Neural Networks	70.39	81.54
kNN	61.73	82.44
LDA	63.24	75.81
PathwayRF	71.00	81.75

As shown in Table 3, in comparing the gSVM-SCAD with L1-SVM and SVM-SCAD, it is interesting to note that the gSVM-SCAD outperforms the other two penalized classifiers in both datasets with gSVM-SCAD is 18.63% higher than L1-SVM for lung cancer data set, and 6.57% higher in gender data set. This is due to the SCAD as a non-convex penalty function is more robust to biasness when dealing with a large number of coefficients β in selecting informative genes compared to the L1 penalty function [6]. Therefore the proposed method with SCAD penalty function is more efficient in selecting informative genes within a pathway compare to LASSO penalty. Table 3 further shows that the gSVM-SCAD had better results than SVM-SCAD, with 20.27% and 9.37% higher in lung cancer and gender data sets respectively. It is demonstrated that group specific tuning parameters in gSVM-SCAD provides flexibility in determining the λ for each pathway compared to the use of general λ for every pathway. This is because genes within pathway usually have a different prior distribution.

In order to show that not all genes in a pathway contributed to the development of specific cellular processes, the gSVM-SCAD is compared with four classifiers. The results are also shown in Table 3. For lung cancer data, gSVM-SCAD outperformed all the classifiers, with 2.77% higher than PathwayRF, 3.8% higher than neural networks, 10.53% higher than LDA, and lastly 12.04% higher than kNN. For gender data, result obtained by gSVM-SCAD is 5.58% higher than PathwayRF, 5.79% higher than neural networks, 4.89% higher than kNN one neighbor and 11.52% higher than LDA.

From the results in Table 2, the gSVM-SCAD shows a better performance when compared to almost four classifiers for both data sets. This is because the standard classifiers built a classification model using all genes within the pathways. If there are uninformative genes inside the pathways, it reduced the classification performance. In contrast, the gSVM-SCAD does not include all genes in the pathways into development of a classification model, as not all genes in a pathway contribute to cellular processes, due to the quality of pathway data.

4.2 Biological validation

The gSVM-SCAD has been tested using the lung cancer data set that has two possible output classes: tumor and normal. The selected genes and the top five pathways presented in Table 4. For lung cancer data set, we used HLungDB [17] and genecards version 3.06 (www.genecards.org) to validate the selected genes within pathways.

For the WNT signaling pathway, it is reported that the pathway plays a significant role in the development of lung and other colorectal cancers [18]. From 24 genes inside the pathways, 16 genes selected by the gSVM-SCAD where twelve genes were validated as related to the lung cancer development while other remaining genes are not contributing to the lung cancer. With respect to the second pathway, it also contributes to the development of lung cancer, since the AKAP95 protein plays an important role in cell mitosis [19], the gSVM-SCAD identifies seven out of 10 genes included in the pathway with four genes such as DDX5, PRKACB, CDK1, and CCNB1 playing an important role in lung cancer development, while others have no evidence in lung cancer development.

The gSVM-SCAD identifies that induction of apoptosis pathway as one of the lung cancer related pathway, where this pathway has been reported by Lee et al. [20] as one of the contributor to the lung cancer development. Thirteen out of eighteen genes in the induction of apoptosis have been selected by the gSVM-SCAD as the significant genes, with thirteen genes being related to the lung cancer. For the Tyrosine metabolism pathway, there are no references showing that this pathway is related to the lung cancer development. However, the gSVM-SCAD has selected several genes within this pathway that play an important role in the development of lung cancer, such as AOC3, DDC, GHR, TPO, NAT6, ALDH3A1, ADH7, MAOA, MAOB and ADH1C. This makes it possible that this pathway may relate to the development of lung cancer and thus prompting biologists to conduct further research on this pathway. While for the Activation of Csk pathway, Masaki et al. [21] have reported that the activation of this pathway plays an important role in the development of lung cancer, with three genes marked as lung cancer genes.

Table 4. Selected genes from the top five pathways in the lung cancer data set.

Pathways	No. of genes	Selected gene(s)
WNT Signaling Pathway	24	**APC** , **MYC, AXIN1, GSK3B, CTNNB1** [18], **HNF1A** [22], **CREBBP** [23], **HDAC1** [24], **WNT1** [25], **CSNK1A1** [26], **CSNK2A1** [15], **TLE1** [27] *PPARD, PPP2CA, TAB1, DVL1*
AKAP95 role in mitosis and chromosome dynamics	10	**DDX5** [28]**, PRKACB** [15]**, CDK1** [29]**, CCNB1** [30], *PPP2CA, PRKAR2B, PRKAR2A*
Induction of apoptosis	36	**FADD** [31], **TNFSF10** [32], **CASP3, CASP6, CASP7, CASP8, CASP9, CASP10** [20], **BCL2** [33], **BIRC3** [34], **TRAF** [35], **BIRC** [36], **TNFRSF25** [37] , **RARA** [38], *TRADD, RELA, DFFA, RIPK1*
Tyrosine metabolism	45	**AOC3, NAT6, ADH7, MAOA, MAOB** [15], **DDC** [39], **ADH1C** [40], **GHR** [41], **TPO** [42], **ALDH3A1** [43], *ADH5, PNMT, TAT, ARD1A, DBT*, *AOC2*, *ALDH1A3*, *AOX1*, *PRMT2*, *FAH*, *ALDH3B2*, *KAT2A*, *ADH6*, *ADH4, GOT2*
Activation of Csk	30	**PRKACB, HLA-DQB1** [15]**, CREBBP** [21] *CD247, IL23A, PRKAR1B*, *GNGT1*, *CD3D*, *CD3E*

For the gender dataset, we used 480 pathways. For this data set, we were looking the genes within pathways that existed in the lymphoblastoid cell lines for both male and female [14]. The top 5 pathways with highest 10-fold CV accuracy are shown in Table 5. Our gSVM-SCAD had selected 11 genes out of total 726 genes within top 5 pathways. From the 11 genes, 8 genes are proved to be related in lymphoblastoid cell lines for both male and female gender.

Table 5. Selected genes from the top five pathways in gender dataset

Pathways	No. of genes	Selected gene(s)
Testis genes from xhx and netaffx	111	**RPS4Y1** [44]
GNF female genes	116	**XIST**
RAP down	434	**DDX3X, HDHD1A** [44] **NDUFS3** [45]
XINACT	34	*RSP4X*, **DDX3X, PRKX** [44]
Willard inact	31	*RPS4X*, **STS** [44], *RPS4P17*

5 Conclusion

This paper focuses to identify the significant genes and pathways that relate to phenotypes of interest by proposing the gSVM-SCAD. From the experiments and analyses, the gSVM-SCAD is shown to outperform the other ML methods in both data sets. In comparison of penalty function, gSVM-SCAD has shown its superiority in selecting the informative genes within pathways compare to L_1 SVM. By providing group-specific tuning parameters, gSVM-SCAD had shown a better performance compare to an SVM-SCAD that provides a general penalty term for all pathways. Furthermore, majority of the genes selected by gSVM-SCAD from both data lung cancer and gender data sets are proved as biologically relevance.

Despite the good performance based on the comparisons done in this paper, gSVM-SCAD still possesses a limitation because SCAD penalty is a parametric method that relies on the parameter λ to balance the trade-off between data fitting and model parsimony [7] and the results will be affected if improper selection by the GACV. This can be seen in Table 4 where there are still a lot false positives. When λ is too small, it can lead to the overfitting of the training model and give too little sparse to the produced classifier; and if λ is too big, it can lead to the underfitting to the training model, which again can give very sparse to the classifier [7]. Therefore, further research regard this matter shall be done to surmount the limitation in gSVM-SCAD.

Acknowledgement. We would like to thank Malaysian Ministry of Science, Technology and Innovation for supporting this research by an e-science research grant (Grant number: 06-01-06-SF1029). This research is also supported by a GUP research grant (Grant number: Q.J130000.2507.04H16) that was sponsored by Universiti Teknologi Malaysia.

References

1. Wang, X., Dalkic, E., Wu, M., Chan, C.: Gene Module Level Analysis: Identification to Networks and Dynamics. Curr Opin Biotechnol. 19: 482-91 (2008)
2. Chen, X., Wang, L., Smith, J.D., Zhang, B.: Supervised principle component analysis for gene set enrichment of microarray data with continuous or survival outcome. Bioinformatics. 24: 2474-81 (2008)
3. Hummel, M., Meister, R., Mansmann, U.: GlobalANCOVA: Exploration and Assessment of Gene Group Effects. Bioinformatics. 24: 78-85 (2008)
4. Tai, F., Pan, W.: Incorporating Prior Knowledge of Predictors into Penalized Classifiers with Multiple Penalty Terms. Bioinformatics. 23: 1775-82 (2007)
5. Tai, F., Pan, W.: Incorporating Prior Knowledge of Gene Functional Groups into Regularized Discriminant Analysis of Microarray Data. Bioinformatics. 23: 3170-7 (2007)
6. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. JASA. 96: 1348-60 (2001)
7. Zhang, H.H., Ahn, J., Lin, X., Park, C.: Gene selection using support vector machines with non-convex penalty. Bioinformatics. 22: 88-95 (2006)

8. Tibshirani, R.: Regression shrinkage and selection via the lasso. J R Stat Soc Series B (Statistical Methodology). 58: 267-288 (1996)
9. Wang, J.T., Wu, X.: Kernel design for RNA classification using Support Vector Machines. Int J Data Min Bioinform. 1: 57-76 (2006)
10. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.: Gene selection for cancer classification using support vector machines. Machine Learning. 46: 389-422 (2002)
11. Wu, Y., Liu, Y.: Robust Truncated Hinge Loss Support Vector Machines. JASA. 102: 974-83 (2007)
12. Wahba, G., Lin, Y., Zhang, H.: GACV for support vector machines, or , another way to look at margin-like quantities. In: Smola, A.J., Bartlett, P., Schoelkopf, B., Schurmans, D. (eds.) Advances in Large Margin Classifiers. pp. 297-309. MIT Press (2000)
13. Wang, H., Li, R., Tsai, C.L.: Tuning parameters selectors for the smoothly clipped absolute deviation method. Biometrika. 94: 553-68 (2007)
14. Pang, H., Lin, A., Holford, M., Enerson, B.E., Lu, B., Lawton, M.P., Floyd, E., Zhao, H.: Pathway analysis using random forest classification and regression. Bioinformatics. 16: 2028-36 (2006)
15. Battacharjee, A., Richards, W.G., Satunton, J., *et al.*: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. PNAS. 98: 13790-5 (2001)
16. Becker, N., Werft, W., Toedt, G., Lichter, P., Benner, A.: PenalizedSVM: A R-package for feature selection SVM classification. Bioinformatics. 25: 1711-2 (2009)
17. Wang, L., Xiong, Y., Sun, Y., Fang, Z., Li, L., Ji, H., Shi, T.: HLungDB: an integrated database of human lung cancer research. Nucleic Acids Res. 38: D665-9 (2010)
18. Mazieres, J., He, B., You, L., Xu, Z., Jablons, D.M.: Wnt signaling in lung cancer. Cancer Letters. 222: 1-10 (2005)
19. Collas, P., Le Guellec, K., Taskén, K.: The A-kinase-anchoring protein AKAP95 is a multivalent protein with a key role in chromatin condensation at mitosis. J Cell Biol. 147: 1167-79 (1999)
20. Lee, S.Y., Choi, Y.Y., Choi, J.E., *et al.*: Polymorphisms in the caspase genes and the risk of lung cancer. J Thorac Oncol. 5: 1152-8 (2010)
21. Masaki, T., Igarashi, K., Tokuda, M., *et al.*: pp60$^{\text{c-src}}$ activation in lung adenocarcinoma. Eur J Cancer. 39: 1447-55 (2003)
22. Lanzafame, S., Caltabiano, R., Puzzo, L., Immè, A.: Expression of thyroid transcription factor 1 (TTF-1) in extra thyroidal sites: papillary thyroid carcinoma of branchial cleft cysts and thyroglossal duct cysts and struma ovarii. Pathologica. 98: 640-4 (2006)
23. Tillinghast, G.W., Partee, J., Albert, P., Kelly, J.M., Burtow, K.H., Kelly, K.: Analysis of genetic stability at the EP300 and CREBBP loci in a panel of cancer cell lines. Genes Chromosomes and Cancer. 37: 121-31 (2003)
24. Sasaki, H., Moriyama, S., Nakashima, Y., Kobayashi, Y., Kiriyama, M., Fukai, I., Yamakawa, Y., Fujii, Y.: Histone deacetylase 1 mRNA expression in lung cancer. Lung Cancer. 46: 171-8 (2004)
25. Huang, C.L., Liu, D., Ishikawa, S., Nakashima, T., Nakashima, N., Yokomise, H., Kadota, K., Ueno, M.: Wnt1 overexpression promotes tumour progression in non-small cell lung cancer. Eur J Cancer. 44: 2680-88 (2008)
26. Wrage, M., Ruosaari, S., Eijk, P.P., *et al.*: Genomic profiles associated with early micrometastasis in lung cancer: relevance of 4q deletion. Clin Cancer Res. 15: 1566-74 (2009)

27. Jagdis, A., Rubin, B.P., Tubbs, R.R., Pacheco, M., Nielsen, T.O.: Prospective evaluation of TLE1 as a diagnostic immunohistochemical marker in synovial sarcoma. Am J Surg Pathol. 33: 1743-51 (2009)
28. Su, L.J., Chang, C.W., Wu, Y.C., *et al*.: Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. BMC Genomics. 8: 140 (2007)
29. Hsu, T.S., Chen, C., Lee, P.T., *et al*.: 7-Chloro-6-piperidin-1-yl-quinoline-5,8-dione (PT-262), a novel synthetic compound induces lung carcinoma cell death associated with inhibiting ERK and CDC2 phosphorylation via a p53-independent pathway. Cancer Chemother Pharmacol. 62: 799-808 (2008)
30. Kosacka, M., Korzeniewska, A., Jankowska, R.: The evaluation of prognostic value of cyclin B1 expression in patients with resected non-small-cell lung cancer stage I-IIIA—preliminary report. Pol Merkur Lekarski. 28: 117-21 (2010)
31. Bhojani, M.S., Chen, G., Ross, B.D., Beer, D.G., Rehemtulla, A.: Nuclear localized phosphorylated FADD induces cell proliferation and is associated with aggressive lung cancer. Cell Cycle. 4: 1478-81 (2005)
32. Sun, W., Zhang, K., Zhang, X., *et al*.: Identification of differentially expressed genes in human lung squamous cell carcinoma using suppression substractive hybridization. Cancer Lett. 212: 83-93 (2004)
33. Nhung, N.V., Mirejovsky, T., Mirejovsky, P., Melinova, L.: Expression of p53, p21 and bcl-2 in prognosis of lung carcinomas. Cesk Patol. 35: 117-21 (1999)
34. Ekedahl, J., Joseph, B., Grigoriev, M.Y., *et al*.: Expression of inhibitor of apoptosis proteins in small- and non-small-cell lung carcinoma cells. Exp Cell Res. 279: 277-90 (2002)
35. Li, X., Yang, Y., Ashwell, J.D.: TNF-RII and c-IAPI mediate ubiquitination and degradation of TRAF2. Nature. 416: 345-7 (2002)
36. Kang, H.G., Lee, S.J., Chae, M.H., *et al*.: Identification of polymorphisms in the XIAP gene and analysis of association with lung cancer risk in a Korean population. Cancer Genet Cytogenet. 180: 6-13 (2008)
37. Anglim, P.P., Galler, J.S., Koss, M.N., *et al*.: Identification of a panel of sensitive and specific DNA methylation markers for squamous cell lung cancer. Mol Cancer. 7: 62 (2008)
38. Wan, H., Hong, W.K., Lotan, R.: Increased retinoic acid responsiveness in lung carcinoma cells that are nonresponsive despite the presence of endogenous retinoic acid receptor (RAR) beta by expression of exogenous retinoid receptors retinoid X receptor alpha, RAR alpha, and RAR gamma. Cancer Res. 61: 556-64 (2001)
39. Vos, M.D., Scott, F.M., Iwai, N., Treston, A.M.: Expression in human lung cancer cell lines of genes of prohormone processing and the neuroendocrine phenotype. J Cell Biochem Suppl. 24: 257-68 (1996)
40. Freudenheim, J.L., Ram, M., Nie, J., *et al*.: Lung cancer in humans is not associated with lifetime total alcohol consumption or with genetic variation in alcohol dehydrogenase 3 (ADH3). J Nutr. 133: 3619-24 (2003)
41. Cao, G., Lu, H., Feng, J., Shu, J., Zheng, D., Hou, Y.: Lung cancer risk associated with Thr495Pro polymorphism of GHR in Chinese population. Jpn J Clin Oncol. 38: 308-16 (2008)
42. Werynska, B., Ramlau, R., Podolak-Dawidziak, M., *et al*.: Serum thrombopoietin levels in patients with reactive thrombocytosis due to lung cancer and in patients with essential thrombocythemia. Neoplasma. 50: 447-51 (2003)
43. Muzio, G., Trombetta, A., Maggiora, M., *et al*.: Arachidonic acid suppresses growth of human lung tumor A549 cells through down-regulation of ALDH3A1 expression. Free Radic Biol Med. 40: 1929-38 (2006)

44. Johnston, C.M., Lovell, F.L., Leongamornlert, D.A., Stranger, B.E., Dermitzakis, E.T., Ross, M.T.: Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. PLoS Genetics. 4: e9 (2008)

45. Zhu, X., Peng, X., Guan, M.X., Yan, Q.: Pathogenic mutations of nuclear genes associated with mitochondrial disorders. Acta Biochim Biophys Sinica. 41: 179-87 (2009)