A Hybrid of SVM and SCAD with Group-Specific Tuning Parameter for Pathway-Based Microarray Analysis

Muhammad Faiz Misman, Mohd Saberi Mohamad, Safaai Deris,
Raja Nurul Mardhiah Raja Mohamad, Siti Zaiton Mohd Hashim, and Sigeru Omatu

Abstract. The incorporation of pathway data into the microarray analysis had lead to a new era in advance understanding of biological processes. However, this advancement is limited by the two issues in quality of pathway data. First, the pathway data are usually made from the biological context free, when it comes to a specific cellular process (e.g. lung cancer development), it can be that only several genes within pathways are responsible for the corresponding cellular process. Second, pathway data commonly curated from the literatures, it can be that some pathway may be included with the uninformative genes while the informative genes may be excluded. In this paper, we proposed 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. Our experiments on lung cancer and gender data sets show that gSVM-SCAD obtains significant results in classification accuracy and in selecting the informative genes and pathways.

Muhammad Faiz Misman · Mohd Saberi Mohamad · Safaai Deris ·
Raja Nurul Mardhiah Raja Mohamad
Artificial Intelligence & Bioinformatics Research Group, Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Johor, Malaysia
e-mail: faizmisman@gmail.com, {saberi,safaai}@utm.my,
mardhiah.mohamad@gmail.com

Siti Zaiton Mohd Hashim
Soft Computing Research Group, Faculty of Computer Science and Information Systems,
Universiti Teknologi Malaysia, 81310, Skudai, Johor Darul Takzim, Malaysia
e-mail: sitizaiton@utm.my

Sigeru Omatu
Department of Electronics, Information and Communication Engineering,
Osaka Institute of Technology, Osaka 535-8585, Japan
e-mail: omatu@rsh.oit.ac.jp

S. Omatu et al. (Eds.): Distributed Computing and Artificial Intelligence, AISC 151, pp. 387–394.
springerlink.com

1 Introduction

In order to obtain further biological information, researchers in recent years have begun to incorporate the microarray data with biological prior knowledge such as pathway data. Currently there are two approaches used in pathway-based microarray analysis, enrichment analysis approaches (EA) and supervised machine learning approaches (ML) [1, 2].

Beside the advantages, this pathway-based microarray analysis also provides some challenges to researchers. One of the challenges is the quality of the pathway data. When the pathway data is curated from the literature or other resources, the informative genes may be excluded while uninformative genes may be included [1]. Chen *et al.* [3] stated that since the pathway data are defined from the biological context free, when dealing in the specific biological context (e.g. cancer development), typically only a subset of genes within pathway are responsible for the corresponding cellular process. In order to deal with these challenges, we used the ML approaches since it have an advantage compared to EA, where ML can select informative genes within pathways by including the gene selection method while EA tends to consider all the genes within pathways are equally important [1]. This is because, gene selection methods provide several advantages such as improves the classification accuracy, remove uninformative genes, and it can reduce computational time [4]. Therefore, we proposed a hybrid of support vector machines and smoothly clipped absolute deviation with group-specific tuning parameter method (gSVM-SCAD) with aim to effectively select the informative genes and pathways that related to a specific biological context.

2 The Proposed Method and Experimental Data

Given a data set $\{(x_i,y_i)\}$, $y_i \in \{-1,1\}$ is the sample tissue with possible two classes $y_i = -1$ and $y_i = 1$ for each data set used in this paper, while $x_i = (x_{i1,\dots}\ x_{id}) \in R^d$ 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 [5]. This has been widely used especially in microarray classification area [6]. SVM distinguish input variables into its classes by a margin of

$$\min_{\beta,c}\Sigma[1-y_i f(x_i)]_+ + \text{pen}_\lambda(\beta) \qquad (1)$$

where $[1\text{-}y_i f(x_i)]_+$ is the SVM convex hinge loss function, while $\text{pen}_\lambda(\beta)$ is the penalty function with parameters λ, where $\beta = (\beta_1,\dots, \beta_i)$ are the coefficients of the hyperplane, while c is the intercept of the hyperplane. Even though SVM has proven its superior ability in classifying high dimensional data, the standard SVM can suffer from irrelevant data, since all the variables are used for constructing the classifier [5]. This is due to the usage of the L_2 penalty in a soft-thresholding function for the common SVM. The detailed applications of L_2 penalty in a soft-thresholding function and its drawbacks in identifying noises can be obtained from [5].

2.1 SVM-SCAD

A penalty function is usually used as a variable selection in the statistics, in bioinformatics it is called as gene selection. SCAD is different from other popular penalty functions such as LASSO, also called as the L_1 penalty [7]. 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 [8]. SCAD penalty has the form of

$$\mathrm{pen}_{\lambda}(\beta) = \Sigma_{j=1}^{d} P_{\lambda}(\beta_j) \tag{2}$$

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, the continuous differentiable penalty function is defined as

$$\mathrm{pen}_{\lambda}(\beta_j) = \begin{cases} \lambda|\beta| & \text{if } |\beta| \leq \lambda \\ -(|\beta|^2 - 2a\,\lambda|\beta| + \lambda^2)/(2(a-1)) & \text{if } \lambda < |\beta| \leq a\lambda \\ ((a+1)\,\lambda)/2 & \text{if } |\beta| > a\lambda \end{cases}$$

where a and λ are tuning parameters with $a > 2$ and $\lambda > 0$ [8]. For a tuning parameter a, Fan and Li [8] suggested the parameter $a = 3.7$ due to the minimal achievement in a bayes risk while λ is a tuning parameter obtained using general approximate cross validation (GACV) method (as discussed latter).

In order to surmount the limitations of the SVM due to its inability to distinguish between noise and informative data, Zhang *et al.* [5] proposed the SVM-SCAD by replacing the L_2 penalty in Equation (1) with Equation (2), which takes the form

$$\min_{\beta,c} \frac{1}{n} \Sigma[1-y_i f(x_i)]_+ + \Sigma^{d}_{j=1} P_{\lambda}(\beta_j) \tag{3}$$

and thus the SVM-SCAD can simultaneously provide gene selection and classification. In order to select the informative genes, SVM-SCAD have to minimize the Equation (3) using the successive quadratic algorithm (SQA) and repeated for kth times until convergence. 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 $\epsilon = y_i - f(x_i)$.

2.2 Tuning Parameter Selection Method

In SCAD there are two tuning parameters namely a and λ that plays an important role in determining an effective predictive model. The tuning parameter selector method in SVM-SCAD is only used to estimate the nearly optimal λ in order to identify the effective predictive model for SCAD. In this paper, a GACV by Wahba *et al.* [9] is used in order to select the nearly optimal λ. The formula on calculating the GACV as given below:

$$\mathrm{GACV}_\lambda = \frac{1}{n}\sum_{i=1}^{n}[1 - y_i f(x_i)_\lambda]_+ + \mathrm{DF}_\lambda \tag{4}$$

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

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

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 (refer [10] for further explanations on RKHS). 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 $\mathrm{DF}_\lambda = K(0,0)/n\gamma^2$ where γ is the hard margin of an SVM [9]. The nearly optimal tuning parameter λ is obtained by minimizing the error rate from the GACV.

2.3 The Proposed Method (gSVM-SCAD)

Since parameter *a* in SVM-SCAD has been setup as 3.7 [8], there is only parameter λ that play an important role. In order to incorporate pathway data, the gSVM-SCAD used group-specific parameters λ_j estimation, using the framework proposed by Tai and Pan [11]. In this paper, there are k groups of genes where k = 1...n, each gene is 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 Equation (2) to the form of

$$\mathrm{pen}_{\lambda k}(\beta_j) = \sum_{j=1}^{d} P\lambda_k\left(\beta_j\right) \tag{5}$$

by allowing each pathway to have it own parameter λ_k as in (5) instead of general λ in Equation (2), the genes within pathways can be selected and classified more accurately. Figure 1 illustrates the procedure of gSVM-SCAD.

There are several main differences between gSVM-SCAD and other current methods in ML. First, it provides the genes selection method to select the informative genes within a pathway that related to the phenotype of interest. Second, the penalty function SCAD is more robust when dealing with a high number of genes, and it selects important genes more consistently than popular L_1 penalty function [5]. And lastly, with group-specific tuning parameters, the gSVM-SCAD provides more flexibility in choosing the best λ for each pathway. Therefore, by selecting the informative genes within a pathway, the gSVM-SCAD can be seen as the best method in dealing with pathway data quality problems in pathway-based microarray analysis.

```
Algorithm: gSVM-SCAD
Input:    GE: Microarray data
          PD: Pathway data
          $\lambda$ : Tuning parameter
Output:   SP: Informative 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 and assign genes from GE that have same pathway info as 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 TP = 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 $\lambda$ 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 $\frac{1}{n}\Sigma[1\text{-}y_i f(x_i)]_+ + \Sigma^d_{j=1} P_\lambda(\beta_j)$
                 $k = k + 1$
            If $\beta_j^k \le \epsilon$ 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 cross validation
     End-for
End
```

Fig. 1 The gSVM-SCAD procedure

2.4 Experimental Data

The performance of the gSVM-SCAD is tested using two types of data, microarray and pathway data. The role of pathway data is as a metadata or prior biological knowledge. For the pathway data, there are a total of 480 pathways with 168 taken from KEGG and the other 312 pathways from BioCarta. The information of the microarray data sets is shown in Table 1. Both data can be downloaded at http://bioinformatics.med.yale.edu/pathway-analysis/datasets.htm.

Table 1. Microarray data sets

Name	Total samples	Total genes	Class	Reference
Lung	86	7129	2 (normal and tumor)	[13]
Gender	32	22283	2 (male and female cells)	unpublished

3 Results and Discussion

3.1 Performance Evaluation

In order to evaluate the performance of gSVM-SCAD, we used a 10-fold cross validation (10-fold CV) classification accuracy. The selected gene and pathways are validated with the biological literatures and databases. The biological validation results can be obtained in our supplementary page (http://www.utm.my/aibig/people/mohd-saberi-mohamad/research/supplementary-information.html).

For the performance evaluation of SCAD penalty function, SCAD was compared with L_1 penalty function by hybridizing it with an SVM classifier (L_1 SVM), obtained from R package penalizedSVM [14]. The L_1 SVM also applied with group-specific tuning parameters to determine λ. Then, the gSVM-SCAD was compared with the current SVM-SCAD with respect to one general tuning parameter for all pathways, the tuning parameter $\lambda = 0.4$ as used by Zhang *et al.* [5]. For comparison with other classification methods without any gene selection process, the gSVM-SCAD was compared with four classifiers that are without gene selection method. The classifiers are PathwayRF [12], multi layer perceptron neural networks with 3 layers (MLP), k-nearest neighbor with one neighbours (kNN), and linear discriminant analysis (LDA). The results of the experiment were shown in Table 2.

In comparing gSVM-SCAD with L_1-SVM and SVM-SCAD, it is interesting to note that gSVM-SCAD outperforms the other two penalized classifiers in both data sets with gSVM-SCAD is 18.63% higher than L_1-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 L_1 penalty function [5]. In contrast to L_1 penalty, SCAD produces sparse solution by thresholding small estimated β to zer (Please refer [5] and [8] for further information of the robustness of non convex penalty in microarray data). Therefore, the proposed method with SCAD penalty function selected more informatively genes within a pathway than the LASSO penalty. Table 2 further shows that the gSVM-SCAD had better results than the SVM-SCAD, with 20.27%, 9.37% higher in lung cancer and gender data sets respectively. It is demonstrated that group tuning parameters in the gSVM-SCAD provided flexibility in determining the λ for each pathway compared to the use of a general λ for whole pathways. This is because usually the genes within pathway have a different prior distribution.

Table 2 further shows that result in lung cancer data set outperformed compared to gender data set. This is because one feature selection method may find many different subsets of features (in this research, features are referred as gene and pathway) that can achieve similar or different classification accuracy [15, 16]. It is believed that, this is related to the instability of the SVM-SCAD as a gene selection method in selecting the informative genes within pathway, since this research focuses only on accuracy-based strategy in analyzing the performance of the gSVM-SCAD. By using the accuracy-based strategy the stability in feature selection method may not be fully reliable in selecting the true informative genes [15].

Table 2 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	53.50	77.96
MLP	70.39	81.54
kNN	61.73	82.44
LDA	63.24	75.81
PathwayRF [13]	71.00	81.75

Note:

The texts in **Bold** are the highest 10-fold CV accuracy.

The texts in *italic* are the methods from the self-running experiment.

In order to show that not all genes in pathways are contributed to the development of specific cellular processes, the gSVM-SCAD is compared with four classifiers. The results are also shown in Table 2. For the lung cancer data set, it shows that the gSVM-SCAD outperformed all the classifiers, with 2.77% higher than PathwayRF, 3.38% higher than MLP, 10.53% higher than LDA, and lastly 12.04% higher than kNN. While for the gender data set, the gSVM-SCAD obtained 5.58% higher than PathwayRF, 5.79% higher than MLP, 4.89% higher than kNN one neighbour 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 all two data sets. This is because the standard classifiers built a classification model using all genes within 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 the development of a classification model, as not all genes in a pathway contribute to cellular processes, due to the quality of pathway data.

4 Summary

This paper focuses on to identify the informative genes and pathways that relate to phenotypes of interests by proposing the gSVM-SCAD. From the experiments and analyses, the gSVM-SCAD was shown to outperform the other supervised machine learning methods in almost all three data sets. In comparison of penalty functions, 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. The proposed method also had shown its ability in identifying the informative genes and pathways.

Acknowledgments. This work is financed by Institutional Scholarship MyPhD provided by the Ministry of Higher Education of Malaysia. We also would like to thank Universiti Teknologi Malaysia for supporting this research by UTM GUP research grants (vot number: Q.J130000.7123.00H67 and Q.J130000.7107.01H29).

References

[1] Wang, X., Dalkic, E., Wu, M., et al.: Gene module level analysis: identification to networks and dynamics. Curr. Opin. Biotechnol. 19, 482–491 (2008), doi:10.1016/j.copbio.2008.07.011

[2] Misman, M.F., Deris, S., Hashim, S.Z.M., et al.: Pathway-based microarray analysis for defining statistical significant phenotype-related pathways: a review of common approaches. In: Int. Conf. Inf. Manag. Eng. (2009), doi:10.1109/ICIME.2009.103

[3] Chen, X., Wang, L., Smith, J.D., et al.: Supervised principle component analysis for gene set enrichment of microarray data with continuous or survival outcome. Bioinformatics 24, 2474–2481 (2008), doi:10.1093/bioinformatics/btn458

[4] Mohamad, M.S., Omatu, S., Deris, S., et al.: A modified binary particle swarm optimization for selecting the small subset of informative genes from gene expression data. IEEE Trans. Inf. Technol. Biomed (2011), doi:10.1109/TITB.2011.2167756

[5] Zhang, H.H., Ahn, J., Lin, X., et al.: Gene selection using support vector machines with non-convex penalty. Bioinformatics 22, 88–95 (2006), doi:10.1093/bioinformatics/bti736

[6] Guyon, I., Weston, J., Barnhill, S., et al.: Gene selection for cancer classification using support vector machines. Mach Learn 46, 389–422 (2002), doi:10.1093/bioinformatics/btl386

[7] Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B (Methodol.) 58, 267–288 (1996), doi:10.1.1.35.7574

[8] Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96(456), 1348–1360 (2001), doi:10.2307/3085904

[9] 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. MIT Press, Cambridge (2000)

[10] Wahba, G.: Support vector machines, reproducing kernel Hilbert spaces, and randomized GACV. In: Schoelkopf, A.B., Burges, C.J.C., Smola, A.J. (eds.) Advances in Kernel Methods – Support Vector Learning. MIT Press, Cambridge (1999)

[11] Tai, F., Pan, W.: Incorporating Prior Knowledge of Predictors into Penalized Classifiers with Multiple Penalty Terms. Bioinformatics 23, 1775–1782 (2007), doi:10.1093/bioinformatics/btm234

[12] Pang, H., Lin, A., Holford, M., et al.: Pathway analysis using random forest classification and regression. Bioinformatics 16, 2028–2036 (2006), doi:10.1093/bioinformatics/btl344

[13] Battacharjee, A., Richards, W.G., Satunton, J., et al.: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. 98, 13790–13795 (2001), doi:10.1073/pnas.191502998

[14] Becker, N., Werft, W., Toedt, G., et al.: PenalizedSVM: A R-package for feature selection SVM classification. Bioinformatics 25, 1711–1712 (2009), doi:10.1093/bioinformatics/btp286

[15] He, Z., Yu, W.: Stable feature selection for biomarker discovery. Computational Biology and Chemistry 34, 215–225 (2010), doi:10.1016/j.compbiolchem.2010.07.002

[16] Zucknick, M., Richardson, S., Stronach, E.A.: Comparing the Characteristics of Gene Expression Profiles Derived by Univariate and Multivariate Classification Methods. Statistical Application in Genetics and Molecular Biology 7, 1–28 (2008), doi:10.2202/1544-6115.1307