An improved hybrid of SVM and SCAD for pathway analysis

Muhammad Faiz Misman*, Mohd Saberi Mohamad, Safaai Deris, Afnizanfaizal Abdullah, Siti Zaiton Mohd Hashim

Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, 81310, Skudai, Johor Darul Takzim, Malaysia; Muhammad Faiz Misman – Email: faizmisman@gmail.com; *Corresponding author

Received September 29, 2011; Accepted October 02, 2011; Published October 14, 2011

Abstract:
Pathway analysis has lead to a new era in genomic research by providing further biological process information compared to traditional single gene analysis. Beside the advantage, pathway analysis provides some challenges to the researchers, one of which is the quality of pathway data itself. The pathway data usually defined from biological context free, when it comes to a specific biological context (e.g. lung cancer disease), typically only several genes within pathways are responsible for the corresponding cellular process. It also can be that some pathways may be included with uninformative genes or perhaps informative genes were excluded. Moreover, many algorithms in pathway analysis neglect these limitations by treating all the genes within pathways as significant. In previous study, a hybrid of support vector machines and smoothly clipped absolute deviation with groups-specific tuning parameters (gSVM-SCAD) was proposed in order to identify and select the informative genes before the pathway evaluation process. However, gSVM-SCAD had showed a limitation in terms of the performance of classification accuracy. In order to deal with this limitation, we made an enhancement to the tuning parameter method for gSVM-SCAD by applying the B-Type generalized approximate cross validation (BGACV). Experimental analyses using one simulated data and two gene expression data have shown that the proposed method obtains significant results in identifying biologically significant genes and pathways, and in classification accuracy.

Keywords: pathway analysis, smoothly clipped absolute deviation, support vector machines.

Background:
Incorporation of prior pathway data into microarray analysis has become a popular research area in bioinformatics due to the advantages in providing the further biological interpretation compared to the single gene microarray analysis. These advantages have triggered various experiments and approaches in order to identify informative genes and pathways that contribute to certain cellular processes. Two most common approaches to identify the informative genes and pathways are enrichment analysis (EA) and machine learning (ML) approaches [1]. In EA approaches, genes are viewed as functional pathways where the pathways with large number of differentially expressed genes between two conditions are considered as significant [1]. While in ML approaches, the genes in the pathways are discriminated between two or more phenotypes of interests and the pathway with genes that improves the prediction of the phenotype considered as significant pathways.

However, there are some challenges in pathway analysis, such as, some pathways may be included with uninformative genes or perhaps informative genes were excluded [1]. Another concern is that usually pathway data are curated from the biological context free; it can be that only several genes take part in certain cellular process when it goes to the phenotype specific analysis (e.g. lung cancer research). In order to deal with the challenges in the quality of pathway data, several researchers in EA approaches attempt to refining the pathway...
data to specific conditions by removing the unaltered genes in the pathways, and including the additional information of the functional interpretation of the pathways or gene groups [1]. In ML approaches, several researchers have included the gene selection methods in order to select the informative genes within pathways before the pathway evaluation process [2, 3]. This is because, gene selection methods provide several advantages such as: (1) improves the classification accuracy, (2) removes uninformative genes, and (3) reduces computational time [13].

Misman et al. [4] had proposed the gSVM-SCAD in order to identify the significant genes and pathways, simultaneously dealing with the challenges in pathway data quality. gSVM-SCAD is a hybrid method of support vector machines (SVM) and smoothly clipped absolute deviation (SCAD) with group-specific tuning parameters that select the informative genes within pathways before the pathway evaluation process. gSVM-SCAD had shown its superior performance in identifying significant genes and pathways compared to the hybrid of SVM and L1 penalty function (L1 SVM), and almost with all other 5 classifiers without gene selection methods [4]. However, in this work, we detected that gSVM-SCAD had some limitations.

A SCAD performance relies on the tuning parameter. If the tuning parameter is too small, it can bring little sparsity and over fit to the classifier model. While if it is too large, it can make very sparse to the classifier model but produced poor over fit to the classifier model. Therefore, it is important to choose an appropriate tuning parameter selector method for the gSVM-SCAD, since the SCAD plays an important role in determining the true classification model for the SVM. The generalized cross validation (GCV) proposed by Fan and Li. [5] is widely used in the past literatures as a tuning parameter selector method for SVM-SCAD. Unfortunately, GCV had some limitations where it poorly performs when dealing with the small number of variables (genes) and large sample sizes [6]. This is due to the reason that usually some pathways contained not more than 100 genes and even some pathways contained less than 10 genes. We believed that this scenario can lead to the poor performance of SCAD in selecting the informative genes and simultaneously identifying significant genes. In order to surmount the limitation of the gSVM-SCAD, we propose the B-type generalized approximate cross validation (BGACV) as a new tuning parameter selector method for gSVM-SCAD. Our proposed method is denoted as gSVM-SCADBGACV.

Methodology:

Experimental Data:

There are two types of data used: gene expression and pathway data. For gene expression data, the canine and lung cancer datasets were used to evaluate the performance of the gSVM-SCADBGACV (the information of the datasets is shown in Table 1, see supplementary material). For the pathway data, there are total of 441 pathways with 129 taken from KEGG and 312 pathways curated from BioCarta database. Both gene expression and pathway data can be downloaded at http://bioinformatics.med.yale.edu/pathway-analysis/datasets.htm. For the simulated data, total 1060 simulated genes are generated, where 848 genes are informative while other 212 genes are uninformative. The sample size for this experiment has been setup as 80. All 1060 genes are distributed into the Gi groups of genes where i = 1,...,15 and each group have a different size of genes. For the distribution of genes, each Gi contains i x 10 of genes. Moreover, 80% of genes in each group are informative. This creates 15 groups of genes with sizes varying from 10 to 150. The simulation data were created using the sim.data function in R package penalized SVM [14].

gSVM-SCAD with BGACV Tuning Parameter Method (The Proposed Method):

Our proposed method (gSVM-SCADBGACV) consists of three steps as shown in Figure 1. Step 1: group genes based on the pathway information provided by the pathway data. Step 2 is the most important step where the genes within pathways are evaluated using the SVM-SCAD, the uninformative genes are discarded from their pathway. In step 2.1, the proposed BGACV is used in order to select an appropriate tuning parameter for SVM-SCAD. In step 3, the classification error from the selected genes for each pathway is calculated. Step 1 and 3 is similar to the gSVM-SCAD but in step 2.1, there are some changes where BGACV is used instead of GACV.

![Figure 1: The gSVM-SCADBGACV procedure.](image-url)
good results when dealing with the diverging number of parameters, Wang et al. [11] admitted that their method still showed some limitation when dealing with the data with high number of samples and low number of variables. Therefore, we used the GACV formula [12] and transformed the GACV formula into BGACV. The GACV proved it robustness when dealing with data with high number of samples and low number of variables [12]. The formula of GACV is given in supplementary material.

For the canine dataset in Table 2 (see supplementary material), it is interesting to note that gSVM-SCADBGACV outperforms gSVM-SCAD and the other machine learning methods with 85.95% accuracy. The gSVM-SCADBGACV achieved 4.87% higher than gSVM-SCAD, 3.62% higher than neural networks, slightly 0.04% higher than PathwayRF, and 4.96% higher than KNN1 and KNN3 respectively, and 9.34% higher than LDA. For the lung cancer dataset, gSVM-SCADBGACV also showed a significant improvement with 76.35% accuracy and 2.78% higher than gSVM-SCAD. While with other methods, gSVM-SCADBGACV obtained 4.55% higher than PathwayRF, 6.16% higher than neural networks, 14.82% higher than KNN1, 11.75% higher than KNN3, 13.31% higher than LDA, 21.41% higher than L1 SVM, 23.05% higher than SVM-SCAD.

From these results, it is proved that BGACV plays an important role in increasing the performance of SCAD in gSVM-SCAD. For the canine dataset, gSVM-SCAD showed quite poor performance, ranked third behind neural networks and PathwayRF. However, when BGACV is applied to the gSVM-SCAD, it showed a significant improvement even it obtained slightly higher than PathwayRF result. This is because, BGACV is consistently identified nearly optimal for SCAD and simultaneously provides nearly unbiased coefficient estimation when dealing with large coefficients. The results from the gSVM-SCADBGACV show significant achievement, since the pathway analysis usually dealing with the lower size of genes and larger size of samples.

Conclusion:
In this work, the proposed gSVM-CADBGACV is shown to outperform the gSVM-SCAD and other supervised machine learning methods. It is also shown that, a tuning parameter selector method plays an important role in gSVM-SCAD in dealing with the small genes size and large sample size that usually happened in pathway-based microarray analysis. The gSVM-SCADBGACV consistently showed better performance from the simulation study, the 10-fold CV accuracy, and in biological validation compared to gSVM-SCAD (see supplementary material for biological validation results). We have only focused on static gene expression data; however, our approach can be implemented with modification to time-course or survival gene expression data.

Acknowledgement:
We would like to thank Universiti Teknologi Malaysia for supporting this research by the UTM GUP research grant (Vot number: Q130000.7107.01H29).

References:

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.
Supplementary material:

The formula of GACV is shown below:

$$GACV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (1 - y_i \beta_{\lambda,i}) + DF_{\lambda}$$

Where $n$, $\beta$, and $\lambda$ refers to number of samples, coefficients of the hyperplane, and tuning parameter, respectively while $DF_{\lambda}$ is the degree of freedom of parameter $\lambda$, where

$$DF_{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_{k=1}^{n} \beta_{\lambda,k}^2}{\lambda} + \sum_{i=1}^{n} \frac{\log n}{n}$$

Where $\beta_{\lambda,k}$ and $\|k(x_i,x_i)\|_2$ is the reproducing kernel hilbert space (RKHS) with SVM reproducing kernel $K$ [12]. According to Shi et al. [6], in the transformation akaike information criterion (AIC) to BIC, BIC replaces degree of freedom, $DF_{\lambda}$ with $(\frac{\log n}{n})$. Therefore, the BGACV can be defined as,

$$BGACV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (1 - y_i \beta_{\lambda,i}) + \log n \cdot DF_{\lambda}$$

The nearly optimal tuning parameter $\lambda$ is obtained by minimizing the error rate from the BGACV.

### Table 1: Gene expression datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>No. of samples</th>
<th>No. of genes</th>
<th>Class</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung</td>
<td>86</td>
<td>7129</td>
<td>2(normal and tumor)</td>
<td>[7]</td>
</tr>
<tr>
<td>Canine</td>
<td>31</td>
<td>12473</td>
<td>2(control and drug induced</td>
<td>[8]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Canine</th>
<th>Lung cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>gSVM-SCAD\text{BGACV}</td>
<td>85.96</td>
<td>76.55</td>
</tr>
<tr>
<td>gSVM-SCAD (Misman et al., 2010)</td>
<td>81.08</td>
<td>73.77</td>
</tr>
<tr>
<td>PathwayRF (Pang et al., 2006)</td>
<td>85.92</td>
<td>71.00</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>82.33</td>
<td>70.39</td>
</tr>
<tr>
<td>KNN1</td>
<td>81.00</td>
<td>61.73</td>
</tr>
<tr>
<td>KNN3</td>
<td>81.00</td>
<td>64.80</td>
</tr>
<tr>
<td>LDA</td>
<td>76.62</td>
<td>63.24</td>
</tr>
<tr>
<td>L_{1} SVM</td>
<td>72.33</td>
<td>55.14</td>
</tr>
<tr>
<td>SVM-SCAD</td>
<td>73.67</td>
<td>53.5</td>
</tr>
</tbody>
</table>

Note: The texts in **Bold** are the highest 10-fold CV accuracy. The texts in *italic* are the methods from the self-running experiment.

**Biological Validation:**

The five most significant pathways in the lung cancer dataset are (1) WNT signaling pathway, (2) AKAP95 role in mitosis and chromosome dynamics pathway, (3) Induction of apoptosis pathway, (4) Antisense pathway, and (5) Cycling of Ran in nucleocytoplasmic transport pathway as shown in **Table 3**. For the WNT signaling pathway, Mazieres et al. [1] reported that the pathway plays a significant role in the development of lung and other colorectal cancers. 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 [2]. 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. [3] as one of the contributor to the lung cancer development. For the Keratinocyte Differentiation pathway, Massion et al. [4] had shown that this pathway has contributed to the development of lung cancer. While for the Activation of Csk pathway, Masaki et al. [5] 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. The selected genes for each pathway are also presented in **Table 3**.
For the canine dataset as shown in Table 4, the gSVM-CADBGACV has been detected. Four pathways are similar to the pathways that are detected by Pang et al. [33]. The pathways are Aminoacyl-tRNA biosynthesis pathway, Granzyme A mediated Apoptosis Pathway, TGF-beta signaling pathway, and Antisense Pathway. Hence, all these four pathways are considered significant based on the literature. For the Starch and Metabolism pathway, it may not seem directly related to vascular pathology, but it contained two genes that related to the metastasis for gene RUVBL2 [34] and inflammation for UGDH [35]. This makes it possible that this pathway may relate to the vascular pathology and thus prompting biologists to conduct further research on this pathway. The selected genes for each pathway are presented in Table 4. Fifteen genes were selected, with thirteen genes detected as informative.

### Table 3: Selected genes from the top five pathways in the lung cancer dataset

<table>
<thead>
<tr>
<th>Pathways</th>
<th>No. of genes</th>
<th>Selected gene(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WNT Signaling Pathway</td>
<td>24</td>
<td>APC [1], HNF1A [6], CREBBP [7], HDAC1 [8], MYC [1], WNT1 [9], CSNK1A1 [10], AXIN1 [1], CSNK2A1 [10], CTNNB1 [1], GSK3B [1], TLE1 [11], PPARD, PPP2CA, TAB1, DVL1</td>
</tr>
<tr>
<td>AKAP95 role in mitosis and chromosome dynamics Induction of apoptosis</td>
<td>10</td>
<td>DDX5 [12], PRKACB [10], CDK1 [13], CCNB1 [14], PPP2CA, PRKAR2B, PRKAR2A</td>
</tr>
<tr>
<td>Keratinocyte Differentiation</td>
<td>45</td>
<td>IRS1 [2], AKAP95, CSNK1A1, CTNNB1, GSK3B, PPP2CA, PRKAR2B, PRKAR2A</td>
</tr>
<tr>
<td>Activation of Csk</td>
<td>30</td>
<td>SET, APEX1 [33], DFFB [33]</td>
</tr>
</tbody>
</table>

Note: The genes in italic text are uninformative. The genes in bold text are genes that directly related to the lung cancer.

### Table 4: Selected genes from the top five pathways in the canine dataset

<table>
<thead>
<tr>
<th>Pathways</th>
<th>No. of genes</th>
<th>Selected gene(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGF-beta signaling pathway</td>
<td>40</td>
<td>RHOA [33], BMP4 [33], SPI [33]</td>
</tr>
<tr>
<td>Starch and sucrose metabolism</td>
<td>59</td>
<td>TESK1, UGDH, RUVBL2 [34], UGDH [35]</td>
</tr>
<tr>
<td>Aminoacyl-tRNA biosynthesis</td>
<td>19</td>
<td>NARX [33], FARS [33], FARS [33]</td>
</tr>
<tr>
<td>Granzyme A mediated Apoptosis Pathway</td>
<td>5</td>
<td>SET, APEX1 [33], DFFB [33]</td>
</tr>
<tr>
<td>Antisense Pathway</td>
<td>34</td>
<td>SFPQ [33], MATR3 [33], ADAR [33]</td>
</tr>
</tbody>
</table>

Note: The genes in italic text are uninformative. The genes in bold text are genes that directly related to vascular injury in coronary arteries related.

References: