Protein-protein interaction network construction for cancer using a new $L_{1/2}$-penalized Net-SVM model

H. Chai, H.H. Huang, H.K. Jiang, Y. Liang and L.Y. Xia

State Key Laboratory of Quality Research in Chinese Medicines & Faculty of Information Technology, Macau University of Science and Technology, Taipa, Macau, China

Corresponding author: Y. Liang
E-mail: yliang@must.edu.mo

Genet. Mol. Res. 15 (3): gmr.15038794
Received May 16, 2016
Accepted June 3, 2016
Published July 25, 2016
DOI http://dx.doi.org/10.4238/gmr.15038794

ABSTRACT. Identifying biomarker genes and characterizing interaction pathways with high-dimensional and low-sample size microarray data is a major challenge in computational biology. In this field, the construction of protein-protein interaction (PPI) networks using disease-related selected genes has garnered much attention. Support vector machines (SVMs) are commonly used to classify patients, and a number of useful tools such as lasso, elastic net, SCAD, or other regularization methods can be combined with SVM models to select genes that are related to a disease. In the current study, we propose a new Net-SVM model that is different from other SVM models as it is combined with $L_{1/2}$-norm regularization, which has good performance with high-dimensional and low-sample size microarray data for cancer classification, gene selection, and PPI network construction. Both simulation studies and real data experiments demonstrated that our

proposed method outperformed other regularization methods such as lasso, SCAD, and elastic net. In conclusion, our model may help to select fewer but more relevant genes, and can be used to construct simple and informative PPI networks that are highly relevant to cancer.

Key words: SVM model; L1/2 regularization; Gene selection; Protein-protein interaction networks; Machine learning

INTRODUCTION

Classifying cancer patients and identifying cancer-related genes using high-dimensional, low-sample size microarray data is an important problem in cancer treatment and drug design. To date, many methods have been used to try to solve this problem including logistic regression, Cox models, and gene-pair methods (Hosmer and Lemeshow, 2004; Ma et al., 2004; Li et al., 2008; Goeman, 2010). The support vector machine (SVM) model (Suykens and Vandewalle, 1999) with different regularization methods is one of the most widely used supervised learning methods, and can be applied for disease classification and feature selection. Owing to the high dimensionality and low sample size of microarray gene data, the SVM model is usually regularized with penalties such as those of L_1-norm (lasso) or L_2-norm (ridge) regularization (Hoerl and Kennard, 1970; Zhu et al., 2004). The goals of the regularization methods are to minimize regression errors and to select relevant variables simultaneously through generation of sparse solutions. The SVM model has shown great success in outcome prediction for different kinds of cancers. However, the weakness of the SVM model is that its results are based purely on computational or algorithmic points, which may not be biologically meaningful in cancer treatment (Bair and Tibshirani, 2004). In order to overcome this shortcoming, Li and Li (2008) proposed a simple and fast network-constrained regularization procedure, which could identify related genes and build networks that were relevant to the disease or disease outcome. Recently, many similar methods have been proposed using gene expression data to construct protein-protein interaction (PPI) networks based on other supervised learning methods such as logistic regression or Cox models combined with different regularization methods (Chuang et al., 2007; Brouard et al., 2010; Zhang et al., 2013a,b). Generally, the widely used L_1-norm and L_2-norm regularization methods may identify a large number of irrelevant disease genes, which greatly increases research costs and makes the constructed networks more complex. Xu et al. (2012) proposed that the $L_{1/2}$ regularization method could find more sparse solutions. Moreover, $L_{1/2}$ regularization has some good statistical properties such as sparsity, unbiasedness, and oracle properties, and has already been successfully applied to real data analyses (Liang et al., 2013; Liu et al., 2014; Chai et al., 2015). However, according to the literature, $L_{1/2}$ regularization has only been used for gene identification and has not been combined with the SVM model (Bair and Tibshirani, 2004), and the results of $L_{1/2}$ regularization are only based on algorithms. In order to obtain more simple, accurate, and biologically meaningful results, we combined a network-constrained procedure and $L_{1/2}$ regularization in a newly proposed Net-SVM model. Our method can be used for disease classification, disease-related gene identification, and can also be used to construct relevant PPI networks. The genes identified by our method can provide molecular interaction information about disease-related biological processes, which can be combined with protein network information collected from biological databases such

as BioGRID that contains biological interaction data from more than 43,000 publications (Stark et al., 2006). The constructed networks, which combine this protein network data with molecular interaction information extracted from gene expression and biological process analyses, have been demonstrated to be biologically meaningful and can effectively remove background noise (Li and Li, 2008; Zhang et al., 2013a). In the current study, a new version of the Net-SVM model with $L_{1/2}$ regularization is introduced; a coordinate descent algorithm is presented that completes the Net-SVM model with $L_{1/2}$ regularization; the performances of different methods on simulated and real datasets are demonstrated; we provide a biological analysis of the selected genes and constructed PPI networks in real cancer datasets; and conclusions about our proposed method are given.

MATERIAL AND METHODS

The $L_{1/2}$ penalized Net-SVM model

In this study, we defined a network $G = (V, E, W)$, where V is the set of genes in the dataset and $e = (u \sim v)$ represents the set of edges that gene u and v are linked in the PPI network. W is the weight of the edges where $w(u, v)$ indicates the weight of the edge $e = (u \sim v)$. We set d_v as the degree of gene v, such that it is the number of edges that are linked with gene u. The normalized Laplacian matrix L for W with $u \sim v$ can be defined as previously described as follows (Chung, 1997):

$$L(u,v) = \begin{cases} 1 - \dfrac{w(u,v)}{d_u} & \text{if } u = v \text{ and } d_u \neq 0 \\ -\dfrac{w(u,v)}{\sqrt{d_u d_v}} & \text{if u and v are linked otherwise} \\ 0 & \end{cases} \qquad \text{(Equation 1)}$$

This matrix L is always a non-negative definite matrix, and we can obtain many useful properties from the graph of the corresponding set of eigenvalues or spectrum.

Consider that the dataset contains n samples and p genes, with $y = (y_1, y_2, ...y_n)^T$ where $y \in (0,1)$, $X = (x_{i1}, x_{i2}, ..., x_{ip})$ indicate the p-dimension covariates. The support vector machine (SVM) model solves the following problem:

$$\min \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \sum_{j=i}^{p} \beta_j h_j(x_i) \right) \right] \qquad \text{(Equation 2)}$$

where $\{h_1(x_1) \ldots h_p(x_p)\}$ are the dictionary of basic functions.

Add the regularization part to the SVM model, it can be written as

$$\min \sum_{i=1}^{n} \left[1 - y_i \left(\beta_0 + \sum_{j=i}^{p} \beta_j h_j(x_i) \right) \right] + \lambda \|\beta\|^q \qquad \text{(Equation 3)}$$

where the λ is the tuning parameter.

Following previous study (Suykens and Vandewalle, 1999), our proposed Net-SVM model with network constraints can be defined as:

$$f(\lambda_1,\lambda_2,\beta)=\left\{\min\sum_{i=1}^{n}\left[1-y_i\left(\beta_0+\sum_{j=i}^{p}\beta_j h_j(x_i)\right)\right]+\lambda_1\|\beta\|^q\right\}+\lambda_2\beta^T L\beta \quad \text{(Equation 4)}$$

where λ_1 and λ_2 are the tuning parameters. The first term is the log-likelihood function of the SVM model and the regularization part is used to induce a sparse solution. The second part is a network constraint based on the Laplacian matrix, which can be used to induce a smooth solution of the network.

According to previous methods (Zou and Hastie, 2005), we produced a new set of <X*, Y*>

$$X^*_{(n+p)*p}=(1+\lambda_2)^{-1/2}\begin{pmatrix}X\\ \sqrt{\lambda_2}S^T\end{pmatrix}, Y^*_{(n+p)}=\begin{pmatrix}Y\\ 0\end{pmatrix} \quad \text{(Equation 5)}$$

where $L=U\Gamma U^T$ and $S=U\Gamma^{1/2}$. Let * = $\sqrt{1+\lambda_2}\beta$ and $\gamma=\frac{\lambda_1}{\sqrt{1+\lambda_2}}$. The Equation 4 can also be written as:

$$f(\lambda_1,\lambda_2,\beta)=f(\gamma,\beta^*)=\min\sum_{i=1}^{n+p}\left[1-y_i^*\left(\beta_0^*+\sum_{j=i}^{p}\beta_j^* h_j(x_i^*)\right)\right]+\gamma\sum_{j=1}^{p}\left|\beta_j^*\right|^q \quad \text{(Equation 6)}$$

Equation 6 allows us to solve the Net-SVM as an equivalent optimization problem with regularization. The L_1-type problem can be written as:

$$f(\gamma,\beta^*)=\min\sum_{i=1}^{n+p}\left[1-y_i^*\left(\beta_0^*+\sum_{j=i}^{p}\beta_j^* h_j(x_i^*)\right)\right]+\gamma\sum_{j=1}^{p}\left|\beta_j^*\right|^1 \quad \text{(Equation 7)}$$

In general, the L_1-type regularization method can efficiently solve the optimization problem. However, due to the high-dimensionality and low-sample size of microarray data from biological samples, the L_1-type regularization may produce many inconsistent gene selections and some results are subject to extra bias. In order to solve this problem, Xu et al. (2010) proposed the $L_{1/2}$ regularization method to obtain a more sparse solution. The sparsity, unbiasedness, and oracle properties of the $L_{1/2}$ regularization led us to predict that it would be more suitable for biological datasets. The Net-SVM model with the $L_{1/2}$ regularization can be written as:

$$f(\gamma,\beta^*)=\min\sum_{i=1}^{n+p}\left[1-y_i^*\left(\beta_0^*+\sum_{j=i}^{p}\beta_j^* h_j(x_i^*)\right)\right]+\gamma\sum_{j=1}^{p}\left|\beta_j^*\right|^{\frac{1}{2}}=(Y^*-X^*\beta^*)^T(Y^*-X^*\beta^*)+\sum_{j=1}^{p}\left|\beta_j^*\right|^{\frac{1}{2}} \quad \text{(Equation 8)}$$

A coordinate descent algorithm for the $L_{1/2}$-penalized Net-SVM model

Next, a coordinate descent algorithm was designed to implement the $L_{1/2}$-penalized Net-SVM model. The main idea of the coordinate descent algorithm for the $L_{1/2}$-penalized Net-SVM model is very simple and efficient. The target function Equation 8 is optimized with respect to the value of the coefficient β_j, and the coordinated descent algorithm is repeated for many cycles from $j = 1$ to p iteratively until all coefficients converge. The coordinate descent algorithm applied for L_1-type regularization by the soft thresholding operator was defined as follows:

$$\beta(j) = Soft(\omega_j, \lambda) = \begin{cases} \omega_j + \lambda & \text{if } \omega_j < \lambda \\ \omega_j - \lambda & \text{if } \omega_j > \lambda \\ 0 & \text{if } |\omega_j| < \lambda \end{cases} \quad \text{(Equation 9)}$$

Following methods described by Xu et al. (2012), the following new half threshold function was used instead of Equation 9:

$$\beta(j) = Half(\omega_j, \lambda) = \begin{cases} \frac{2}{3}\omega_j \left(1 + \cos\left(\frac{2(\pi - \varphi_\lambda(\omega_j))}{3}\right)\right) & \text{if } |\omega_j| > \frac{\sqrt[3]{54}}{4}(\lambda)^{2/3} \\ 0 \quad \text{otherwise} \end{cases} \quad \text{(Equation 10)}$$

where $\varphi_\lambda(\omega_j) = \arccos\left(\frac{\lambda}{8}\left(\left(\frac{\omega_j}{3}\right)^{-3/2}\right)\right)$.

Based on this new half threshold function, the coordinate descent algorithm was designed with the Newton-Raphson iterative method for the $L_{1/}$-penalized Net-SVM model, and proceeds as follows:

Coordinate descent algorithm for the L1/2 penalized Net-SVM model

Step 1: Initial all $\beta = 0$ ($j = 1,2 \ldots p$) and λ; set $m = 0$; Step 2: Construct the Laplacian matrix L, the X^* and Y^*; Step 3: Solve $(Y^* - X^*\beta^*)^T(Y^* - X^*\beta^*) + \sum_{j=1}^{p} |\beta_j^*|^{\frac{1}{2}}$, subject to the constraints of the Net-SVM model with penalties; Step 4: Make $m = m + 1$, update $\beta = Half(\omega_j, \lambda)$; Step 5: Repeat Steps 3, 4 until all $\beta(m)$ are converged.

RESULTS

Simulation experiments

To test the performance of our proposed $L_{1/2}$-penalized Net-SVM model, we compared the results of Net-SVM models with the following four regularization approaches: elastic net, lasso, SCAD, and $L_{1/2}$. We generated the test network datasets according to the following algorithm (Li and Li, 2008):

Step 1: Suppose that there are 200 independent transcription factors x_n which each

transcription factor regulates 10 different genes x_m, so that the constructed network contain about (200*10+200 = 2200) variables, set N = 100. That means the dimension p = 2200, and the size of the dataset N = 100. The transcription factors x_n, x_m are generated by the normal distribution $N(0,1)$.

Step 2: Consider the correlation between the transcription factors and their respective regulated genes, set the correlation coefficient r=0.75, the regulated genes x_m will be rewritten as: $x_m = (1 - 0.75) * x_m + 0.75 * x_n$. Combine the x_m and x_n, we get the total variables X_i.

Step 3: Generate $w = \left(5, \underbrace{\frac{5}{\sqrt{5}} \ldots \frac{5}{\sqrt{5}}}_{10}, -5, \underbrace{\frac{-5}{\sqrt{5}}, \ldots \frac{-5}{\sqrt{5}}}_{10}, 3, \underbrace{\frac{3}{\sqrt{5}}, \ldots \frac{3}{\sqrt{5}}}_{10}, -3, \underbrace{\frac{-3}{\sqrt{5}}, \ldots \frac{-3}{\sqrt{5}}}_{10}, 0 \ldots 0 \right)$ and the noise control parameter $\varepsilon \sim N(0, \sigma_e^2)$.

Step 4: The corresponding y_i is defined as: $if \ \frac{\exp(X_i w + s)}{1 + \exp(X_i w + s)} \geq 0.5, \ y_i = 1; \ else \ y_i = -1$.

In the current study, a 10-fold cross validation approach was used to tune the regularization parameters for the different penalized Net-SVM models. In order to obtain more accurate results, all tuned methods were applied to the different datasets 100 times.

Three parameters were used to evaluate the accuracy of the different methods in the test experiments, including the percent correct, sensitivity, and specificity. We defined the true positive (TP) as the number of correctly selected genes, false positive (FP) as the number of irrelevant selected genes, false negative (FN) as the number of genes related to the disease that were not selected, and true negative (TN) as the number of the irrelevant genes that were not selected by different methods.

$$\text{percent correct} = \frac{\text{selected correct genes}}{\text{total selected genes}}$$

$$Sensitivity = \frac{TP}{TP + FN}, Specificity = \frac{TN}{TN + FP}$$

Table 1 shows the performances of the Net-SVM models with different regularizations. We found that when comparing the numbers of correctly selected genes, the Net-SVM with elastic net regularization selected the most correct genes (43.82), whereas the Net-SVM with $L_{1/2}$ regularization selected the least (42.65); nevertheless, the differences of the results obtained by the different methods were very small. For the number of total selected genes, the Net-SVM with lasso or elastic net regularizations selected the largest numbers of genes. On the contrary, the Net-SVM model with $L_{1/2}$ regularization only selected approximately 56.43 genes. Thus, the accuracy of gene selection with the $L_{1/2}$ regularization was higher (75.58%) than those obtained with SCAD (61.15%), lasso (13.23%) and elastic net (9.83%) regularizations.

When comparing the sensitivities, we found that the values obtained by the different methods were very close. In regards to specificity, the performance of Net-SVM with $L_{1/2}$ regularization was the best. This indicates that too many irrelevant genes were selected by the other three methods. The misclassification errors are shown in the last column of Table 1. The Net-SVM model with elastic net regularization resulted in the largest misclassification error value (8.12), while the Net-SVM with $L_{1/2}$ regularization achieved the lowest value (4.59). As mentioned above, the Net-SVM model with $L_{1/2}$ regularization selected the lowest

number of genes in the datasets, and thus obtained the highest accuracy in gene selection. Furthermore, this method exhibited the best performance in the classification compared to the other methods. Therefore, we concluded that the Net-SVM model with $L_{1/2}$ regularization may be an accurate and efficient method for high-dimensional and low-sample size biological datasets in cancer research.

Table 1. Gene selection performance of different Net-SVM models with different regularization methods.

Methods	Selected correct genes	Total selected genes	Percent correct	Sensitivity	Specificity	Misclassification error
Net-SVM+$L_{1/2}$	42.65	56.43	75.58%	96.93%	99.36%	4.59
Net-SVM+SCAD	42.83	70.04	61.15%	97.34%	98.74%	5.03
Net-SVM+Lasso	43.31	327.28	13.23%	98.43%	86.83%	7.74
Net-SVM+Elastic net	43.82	445.53	9.83%	99.59%	81.37%	8.12

Figures 1-4 show the coefficient paths and misclassification errors obtained by the different methods for one run in the simulation experiments. The vertical dotted lines indicate the optimal solutions as determined by the minimal misclassification values computed in the 10-fold cross validation. We found that the solution path obtained by the Net-SVM model with $L_{1/2}$ regularization was sparser than those obtained by the other three methods.

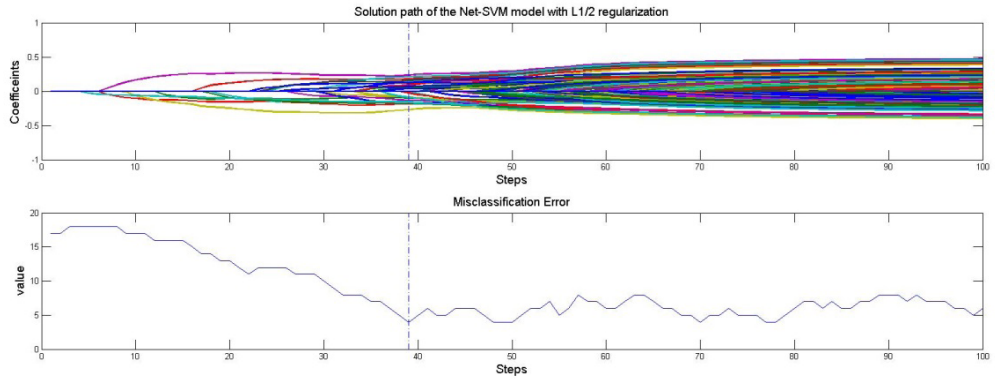

Figure 1. Performance obtained by the Net-SVM model with $L_{1/2}$ regularization.

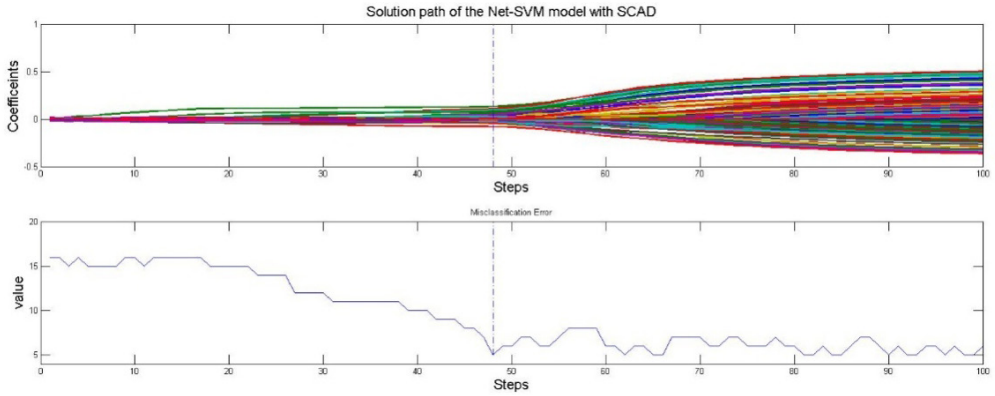

Figure 2. Performance obtained by the Net-SVM model with SCAD regularization.

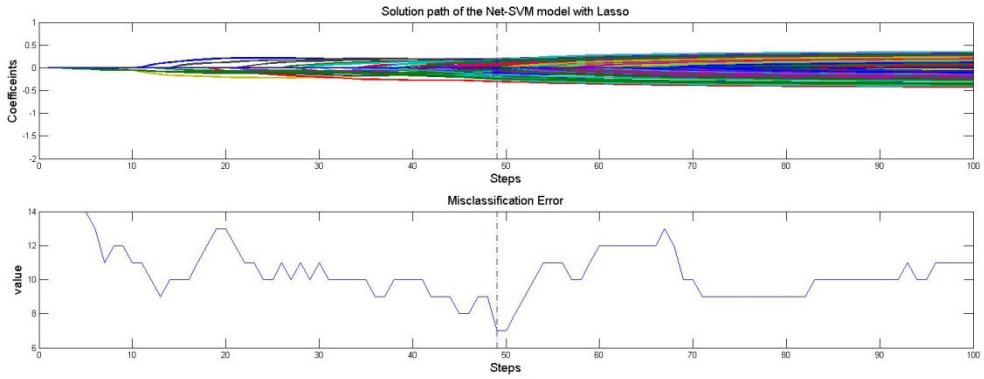

Figure 3. Performance obtained by the Net-SVM model with Lasso regularization.

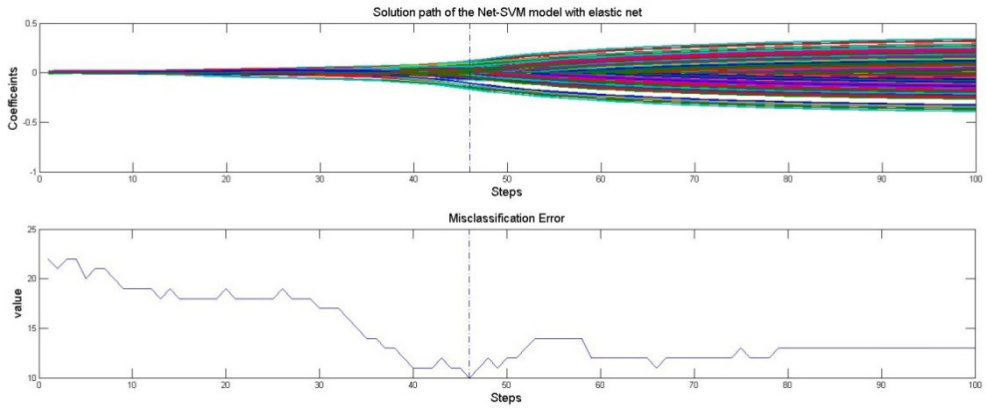

Figure 4. Performance obtained by the Net-SVM model with elastic net regularization.

Real biological data experiments

In order to further evaluate the performances of the four Net-SVM models with different regularizations, two real datasets were used.

Prostate cancer dataset

The prostate cancer dataset was produced by Dinesh et al. (Singh et al., 2002), and contains information on approximately 12,600 genes from 102 total samples, including those from 52 prostate cancer patients and 50 healthy controls. We evaluated the prediction performance of the four different Net-SVM models using the following random partition: 3/4 of the samples (N = 77) were used as the training set, and the other 25 samples were used for testing prediction capability.

Lung cancer dataset

The lung cancer dataset GDS3527 (Landi et al., 2008) was downloaded from the NCBI GEO Database (http://www.ncbi.nlm.nih.gov/sites/GDSbrowser). This lung cancer gene expression dataset contains information on 22284 genes from 58 lung cancer patients and 49 healthy controls. A total of 80 samples were used for the training, and the other 27 samples were used for testing.

Tables 2 and 3 show the average results of the Net-SVM models with different regularizations applied to the two real datasets for 100 runs. Our results revealed that the $L_{1/2}$-penalized Net-SVM model selected the lowest number of genes, whereas the elastic net regularization selected the most genes. The numbers of wrongly classified patients by the four methods were very similar, and the performance of the $L_{1/2}$-penalized Net-SVM model was the best. Performance is a very important factor in clinical research, where the goal is to obtain an accurate result using the lowest number of genes to reduce research costs.

Table 2. The results of the four Net-SVM models with different regularization methods in prostate tumor dataset.

	Selected genes	Connected genes	Connected edges	Cross validation error	Test error
Net-SVM+$L_{1/2}$	68.74	54.96		4.01/77	2.95/25
Net-SVM+SCAD	76.93	61.48		4.07/77	2.97/25
Net-SVM+Lasso	120.52	93.41		4.15/77	3.06/25
Net-SVM+Elastic net	215.17	182.62		4.21/77	3.07/25

Table 3. The results of the four Net-SVM models with different regularization methods in lung cancer dataset.

	Selected genes	Connected genes	Connected edges	Cross validation error	Test error
Net-SVM+$L_{1/2}$	180.32	76.57	80.15	6.56/80	3.88/27
Net-SVM+SCAD	214.56	102.11	111.48	6.76/80	3.92/27
Net-SVM+Lasso	306.19	178.26	239.34	7.02/80	4.16/27
Net-SVM+Elastic net	421.73	243.44	333.06	6.95/80	4.09/27

Biological analysis of the selected genes and constructed PPI networks in lung cancer

Here, we provide a brief biological analysis of the results for the lung cancer dataset GDS3527. In Figures 5-8, the PPI networks related to lung cancer obtained by the four Net-SVM models with different regularizations are presented. The results clearly demonstrate that the PPI network obtained by the $L_{1/2}$-penalized Net-SVM model is more concise than the other three networks. At the same time, the classification errors obtained by the $L_{1/2}$-penalized Net-SVM model were the lowest compared to those obtained by the other methods (Table 3). Hence, we expect that our proposed method may help researchers construct disease-related PPI networks quickly and accurately.

In evaluating the four PPI networks constructed by the different methods, we found that some important genes were present in all four networks, including RPA3, TAL1, MIF, SPP1, NME1, TTN, HSPB2, CRYAB, CAV1, and ENO1, which were mostly found at center nodes in the PPI networks and had many network branches. Interestingly, we found that although these genes were at important nodes in the constructed the PPI networks, they may not be the decisive nodes that determine whether the person has the disease. Table 4 lists the 15 top-ranked disease-related genes that were selected by the four different regularization methods. The genes in bold are the genes that were selected by all four methods. As seen in Table 4,

we found that only three center nodes were present in the 15 top-ranked informative genes, including SPP1, TAL1, and CAV1. Moreover, these three genes have all been implicated in cancer. Specifically, SPP1, the protein encoded by the *SPP1* gene, is involved in the attachment of osteoclasts and has been previously implicated in cancer (Wu et al., 2014; Lin et al., 2015). Additionally, TAL1 and CAV1 have both been suggested to play a role in cancer (Patel et al., 2014; Loosveld et al., 2014; Sayhan et al., 2015; Zhao et al., 2015).

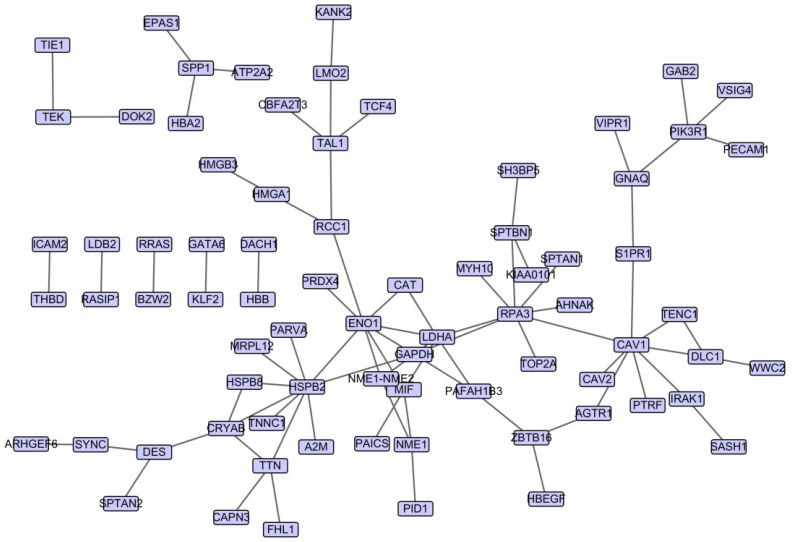

Figure 5. PPI network for lung cancer obtained by $L_{1/2}$ penalized Net-SVM.

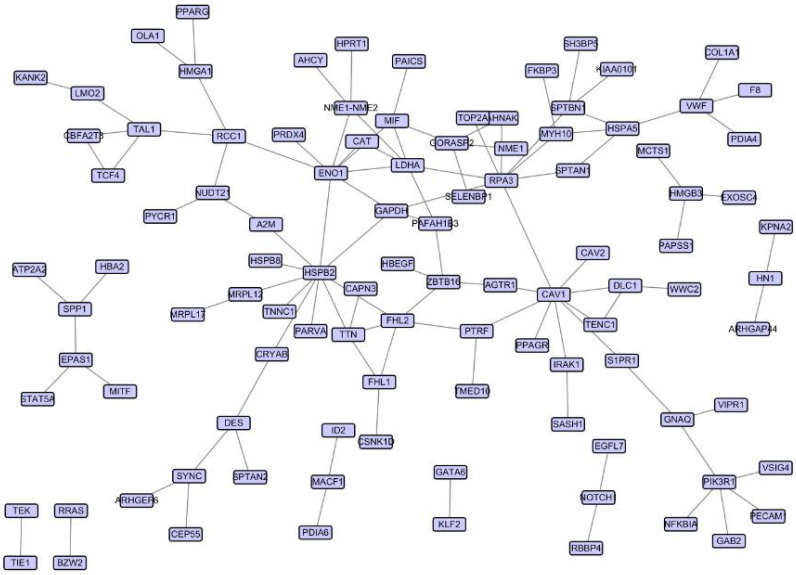

Figure 6. PPI network for lung cancer obtained by SCAD penalized Net-SVM.

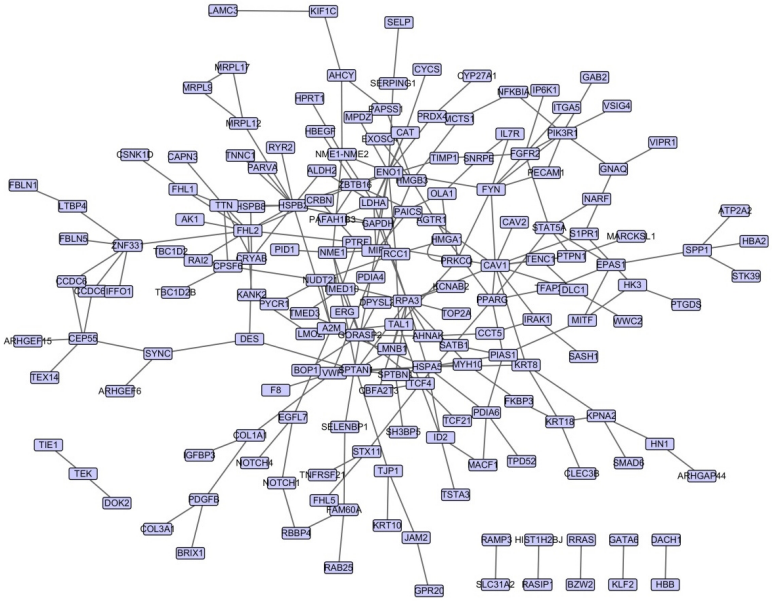

Figure 7. PPI network for lung cancer obtained by Lasso penalized Net-SVM.

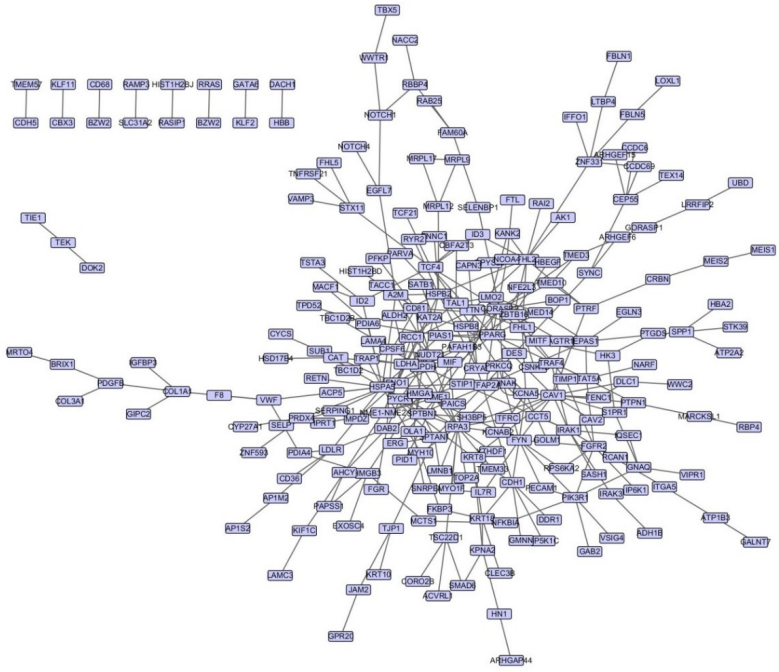

Figure 8. PPI network for lung cancer obtained by elastic net penalized Net-SVM.

Table 4. The 15 top-ranked informative genes selected by Net-SVM models with different regularization methods.

Rank	$L_{1/2}$	SCAD	Lasso	Elastic net
1	**SPP1**	**SPP1**	**SPP1**	**SPP1**
2	TEK	**TAL1**	**AGTR1**	**AGTR1**
3	PECAM1	**AGTR1**	HK3	**CAT**
4	**TAL1**	HSPB2	RASIP1	HK3
5	HIST1H2BJ	TEK	CD34	**TAL1**
6	**AGTR1**	MIF	**CAT**	RASIP1
7	RASIP1	**SASH1**	**TAL1**	CD34
8	**CAV1**	CAV2	FHL5	TTN
9	EPAS1	**CAV1**	LDHA	FHL5
10	**SASH1**	NME1-NME2	ARHGEF15	LDHA
11	S1PR1	**CAT**	TTN	ARHGEF15
12	NME1	ENO1	**CAV1**	VSIG4
13	FHL1	NUDT21	**SASH1**	GOLM1
14	**CAT**	TTN	MIF	**SASH1**
15	CRYAB	EPAS1	NME1	**CAV1**

In addition to these genes, three other genes were selected by all four models, including SASH1, AGTR1, and CAT. SASH1 has been shown to play an important role in tumor formation (Martini et al., 2011). AGTR1 has been shown to be important in controlling blood pressure and volume in the cardiovascular system, and can be found in a KEGG cancer pathway. CAT encodes for a catalase that is an important antioxidant enzyme in the human body to defend against oxidative stress. Oxidative stress plays an important role in the development of many chronic or late-onset diseases such as cancer, asthma, and diabetes. Thus, this gene is also associated with cancer (Shen et al., 2015). There were some other genes that were selected by the Net-SVM model with $L_{1/2}$ regularization that were not selected by the other models, which were also related to cancer. For example, previous studies have demonstrated that high expression of CRYAB was correlated with poor survival in non-small cell lung cancer patients (Qin et al., 2014). Another gene, NME1, has been shown to have a great effect on cancer inhibition, and thus is very important in cancer treatment (Banerjee et al., 2015; Niitsu, 2014). Therefore, the genes only selected by the $L_{1/2}$-penalized Net-SVM model were all related to the cancer. Above all, we believe that our Net-SVM model with $L_{1/2}$ regularization can identify cancer-related genes accurately and efficiently.

DISCUSSION

In the current study, we proposed a new Net-SVM model with $L_{1/2}$ regularization, which can be used to identify highly relevant biomarkers in high-dimensional and low-sample size biological datasets. The method can also be used to construct corresponding PPI disease networks. This model was completed by a coordinate descent algorithm with the Newton-Raphson iterative method. Our method was shown to select fewer genes, but this did not reduce the accuracy of its predictions. The results of the simulation and real data experiments both showed that the performance of the Net-SVM model with $L_{1/2}$ regularization was better than those of the other methods in regards to gene selection and classification. Our method can also be used to construct simple and accurate PPI networks for cancer diagnosis, and hence, the Net-SVM model with $L_{1/2}$ regularization may be a competitive method for gene selection and PPI network construction.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research supported by the Macau Science and Technology Development Funds (Grant #099/2013/A3) from the Macau Special Administrative Region of the People's Republic of China.

REFERENCES

Bair E and Tibshirani R (2004). Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biol.* 2: E108. http://dx.doi.org/10.1371/journal.pbio.0020108

Banerjee S, Jha HC and Robertson ES (2015). Regulation of the metastasis suppressor Nm23-H1 by tumor viruses. *Naunyn Schmiedebergs Arch. Pharmacol.* 388: 207-224. http://dx.doi.org/10.1007/s00210-014-1043-8

Brouard C, Szafranski M and d'Alché-Buc F (2010). Regularized output kernel regression applied to protein-protein interaction network inference. In NIPS MLCB Workshop.

Chai H, Liang Y and Liu XY (2015). The L(1/2) regularization approach for survival analysis in the accelerated failure time model. *Comput. Biol. Med.* 64: 283-290. http://dx.doi.org/10.1016/j.compbiomed.2014.09.002

Chuang HY, Lee E, Liu YT, Lee D, et al. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3: 140. http://dx.doi.org/10.1038/msb4100180

Chung FR (1997). Spectral graph theory. Vol. 92. American Mathematical Soc.

Goeman JJ (2010). L1 penalized estimation in the Cox proportional hazards model. *Biom. J.* 52: 70-84.

Hoerl AE and Kennard RW (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12: 55-67. http://dx.doi.org/10.1080/00401706.1970.10488634

Hosmer J and Lemeshow S (2004). Applied logistic regression. John Wiley & Sons, New York.

Landi MT, Dracheva T, Rotunno M, Figueroa JD et al. (2008). Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PloS One* 3: e1651.

Li C and Li H (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* 24: 1175-1182. http://dx.doi.org/10.1093/bioinformatics/btn081

Li J, Tang X, Liu J, Huang J, et al. (2008). A novel approach to feature extraction from classification models based on information gene pairs. *Pattern Recognit.* 41: 1975-1984. http://dx.doi.org/10.1016/j.patcog.2007.11.019

Liang Y, Liu C, Luan XZ, Leung KS, et al. (2013). Sparse logistic regression with a L 1/2 penalty for gene selection in cancer classification. *BMC Bioinformatics* 14: 1. http://dx.doi.org/10.1186/1471-2105-14-198

Lin Y, McKinnon KE, Ha SW and Beck GR, Jr. (2015). Inorganic phosphate induces cancer cell mediated angiogenesis dependent on forkhead box protein C2 (FOXC2) regulated osteopontin expression. *Mol. Carcinog.* 54: 926-934. http://dx.doi.org/10.1002/mc.22153

Liu C, Liang Y, Luan XZ, Leung KS, et al. (2014). The L 1/2 regularization method for variable selection in the Cox model. *Appl. Soft Comput.* 14: 498-503. http://dx.doi.org/10.1016/j.asoc.2013.09.006

Loosveld M, Bonnet M, Gon S, Montpellier B, et al. (2014). MYC fails to efficiently shape malignant transformation in T-cell acute lymphoblastic leukemia. *Genes Chromosomes Cancer* 53: 52-66. http://dx.doi.org/10.1002/gcc.22117

Ma XJ, Wang Z, Ryan PD, Isakoff SJ, et al. (2004). A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. *Cancer Cell* 5: 607-616. http://dx.doi.org/10.1016/j.ccr.2004.05.015

Martini M, Gnann A, Scheikl D, Holzmann B, et al. (2011). The candidate tumor suppressor SASH1 interacts with the actin cytoskeleton and stimulates cell-matrix adhesion. *Int. J. Biochem. Cell Biol.* 43: 1630-1640. http://dx.doi.org/10.1016/j.biocel.2011.07.012

Niitsu N (2014). The association of nm23-H1 expression with a poor prognosis in patients with peripheral T-cell lymphoma, not otherwise specified. *J. Clin. Exp. Hematop.* 54: 171-177. http://dx.doi.org/10.3960/jslrt.54.171

Patel B, Kang Y, Cui K, Litt M, et al. (2014). Aberrant TAL1 activation is mediated by an interchromosomal interaction in human T-cell acute lymphoblastic leukemia. *Leukemia* 28: 349-361. http://dx.doi.org/10.1038/leu.2013.158

Qin H, Ni Y, Tong J, Zhao J, et al. (2014). Elevated expression of CRYAB predicts unfavorable prognosis in non-small cell lung cancer. *Med. Oncol.* 31: 142. http://dx.doi.org/10.1007/s12032-014-0142-1

Sayhan S, Diniz G, Karadeniz T, Ayaz D, et al. (2015). Expression of caveolin-1 in peritumoral stroma is associated with histological grade in ovarian serous tumors. *Ginekol. Pol.* 86: 424-428. http://dx.doi.org/10.17772/gp/2398

Shen Y, Li D, Tian P, Shen K, et al. (2015). The catalase C-262T gene polymorphism and cancer risk: a systematic review and meta-analysis. *Medicine (Baltimore)* 94: e679. http://dx.doi.org/10.1097/MD.0000000000000679

Singh D, Febbo PG, Ross K, Jackson DG, et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1: 203-209. http://dx.doi.org/10.1016/S1535-6108(02)00030-2

Stark C, Breitkreutz BJ, Reguly T, Boucher L, et al. (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 34 (suppl 1): D535-D539. http://dx.doi.org/10.1093/nar/gkj109

Suykens JA and Vandewalle J (1999). Least squares support vector machine classifiers. *Neural Process. Lett.* 9: 293-300. http://dx.doi.org/10.1023/A:1018628609742

Wu XL, Lin KJ, Bai AP, Wang WX, et al. (2014). Osteopontin knockdown suppresses the growth and angiogenesis of colon cancer cells. *World J. Gastroenterol.* 20: 10440-10448. http://dx.doi.org/10.3748/wjg.v20.i30.10440

Xu ZB, Zhang H, Wang Y, Chang X, et al. (2010). L1/2 regularization. *Sci. China Information Sci.* 53: 1159-1169. http://dx.doi.org/10.1007/s11432-010-0090-0

Xu ZB, Chang X, Xu F and Zhang H (2012). Regularization: A thresholding representation theory and a fast solver. Neural Networks and Learning Systems. *IEEE Transactions on* 23: 1013-1027.

Zhang W, Ota T, Shridhar V, Chien J, et al. (2013a). Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLOS Comput. Biol.* 9: e1002975. http://dx.doi.org/10.1371/journal.pcbi.1002975

Zhang W, Wan YW, Allen GI, Pang K, et al. (2013b). Molecular pathway identification using biological network-regularized logistic models. *BMC Genomics* 14 (Suppl 8): S7. http://dx.doi.org/10.1186/1471-2164-14-S8-S7

Zhao Z, Han FH, Yang SB, Hua LX, et al. (2015). Loss of stromal caveolin-1 expression in colorectal cancer predicts poor survival. *World J. Gastroenterol.* 21: 1140-1147. http://dx.doi.org/10.3748/wjg.v21.i4.1140

Zhu J, Rosset S, Hastie T and Tibshirani R (2004). 1-norm support vector machines. *Adv. Neural Inf. Process. Syst.* 16: 49-56.

Zou H and Hastie T (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.* 67: 301-320. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x