Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 713953, 7 pages
http://dx.doi.org/10.1155/2015/713953

Research Article

Network-Based Logistic Classification with an Enhanced $L_{1/2}$ Solver Reveals Biomarker and Subnetwork Signatures for Diagnosing Lung Cancer

Hai-Hui Huang, Yong Liang, and Xiao-Ying Liu

Faculty of Information Technology & State Key Laboratory of Quality Research in Chinese Medicines, Macau University of Science and Technology, Avenida Wai Long, Taipa 999078, Macau

Correspondence should be addressed to Yong Liang; yliang@must.edu.mo

Received 24 October 2014; Revised 5 April 2015; Accepted 30 April 2015

Academic Editor: Jennifer Wu

Identifying biomarker and signaling pathway is a critical step in genomic studies, in which the regularization method is a widely used feature extraction approach. However, most of the regularizers are based on L_1-norm and their results are not good enough for sparsity and interpretation and are asymptotically biased, especially in genomic research. Recently, we gained a large amount of molecular interaction information about the disease-related biological processes and gathered them through various databases, which focused on many aspects of biological systems. In this paper, we use an enhanced $L_{1/2}$ penalized solver to penalize network-constrained logistic regression model called an enhanced $L_{1/2}$ net, where the predictors are based on gene-expression data with biologic network knowledge. Extensive simulation studies showed that our proposed approach outperforms L_1 regularization, the old $L_{1/2}$ penalized solver, and the Elastic net approaches in terms of classification accuracy and stability. Furthermore, we applied our method for lung cancer data analysis and found that our method achieves higher predictive accuracy than L_1 regularization, the old $L_{1/2}$ penalized solver, and the Elastic net approaches, while fewer but informative biomarkers and pathways are selected.

1. Introduction

Identifying molecular biomarker or signaling pathway involved in a phenotype is a particularly important problem in genomic studies. Logistic regression is a powerful discriminating method and has an explicit statistical interpretation which can obtain probabilities of classification regarding the class label information.

A key challenge in identifying diagnosis or prognosis biomarkers using the logistic regression model is that the number of observations is much smaller than the size of measured biomarkers in most of the genomic studies. Such limitation causes instability in the algorithms used to select gene marker. Regularization methods have been widely used in order to deal with this problem of high dimensionality. For example, Shevade and Keerthi proposed the sparse logistic regression based on the Lasso regularization [1, 2]. Meier et al. investigated logistic regression with group Lasso [3]. Usually, the Lasso type procedures are often called L_1-norm type regularization methods. However, L_1 regularization may yield inconsistent selections when applied to variable selection in some situations [4] and often introduces the extra bias in the estimation [5]. In many genomic studies, we need a sparser solution for interpretation and accurate outcomes, but L_1 regularization has a gap to meet these requirements. Thus, a further improvement of regularization is urgently required. L_q $(0 < q < 1)$ regularization can assuredly generate more sparse and precise solutions than L_1 regularization. Moreover, $L_{1/2}$ penalty can be taken as a representative of L_q $(0 < q < 1)$ penalty and has demonstrated many attractive properties which do not appear in some L_1 regularization approaches, such as unbiasedness, sparsity, and oracle properties [6–8].

So far, we observed dense molecular interaction information about the disease-related biological processes and gathered it through databases focused on many aspects of biological systems. For example, BioGRID records collected various biological interactions from more than 43,468

publications [9]. These regulatory relationships are usually represented by a network. Combining these pieces of graphic information extracted from the biological process with an analysis of the gene-expression data had provided useful prior information to detective noise and removes confounding factors from biological data for several classification and regression models [10–14].

Inspired by the aforementioned methods and ideas, here, we define a network-constrained logistic regression model with $L_{1/2}$ penalty following the framework established by [11], where the predictors are based on the gene-expression data with biologic network knowledge. The proposed model is aimed at identifying some biomarkers and subnetworks regarding diseases. In order to achieve a better prediction, we use an enhanced half thresholding algorithm for $L_{1/2}$ regularization, which is more efficient than the old half thresholding approach in the literature [6, 15, 16].

The rest of the paper is organized as follows. In Section 2, we proposed a new version of the network-constrained logistic regression model with $L_{1/2}$ regularization. In Section 3, we presented an enhanced half thresholding method for $L_{1/2}$ regularization and the corresponding coordinate descent algorithm. In Section 4, we evaluated the performance of our proposed approach on the simulated data and presented the applications of the proposed methods to an analysis of lung cancer data. We concluded the paper with Section 5.

2. $L_{1/2}$ Penalized Network-Constrained Logistic Regression Model

Generally, assuming that dataset D has n samples, $D = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$, where $X_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is ith sample with p genes and y_i is the corresponding variable that takes a value of 0 or 1. Define a classifier $f(x) = e^x/(1 + e^x)$ and the logistic regression is defined as

$$P\left(y_i = 1 \mid X_i\right) = f\left(X'_i\beta\right) = \frac{\exp\left(X'_i\beta\right)}{1 + \exp\left(X'_i\beta\right)}, \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)$ are the coefficients to be estimated. We can obtain β by minimizing the log-likelihood function of the logistic regression. Following [11], to combine biological network with an analysis of the gene microarray data, we used a Laplacian constraint approach here. Consider a graph $G = (V, E)$, where V is the set of genes that meet p explanatory variables and E is the set of edges. If gene u and gene v are connected, then there is an edge between gene u and gene v, which is denoted by $E_{uv} = 1$; else $E_{uv} = 0$. w_{uv} denotes the weight of edge E_{uv}. The normalized Laplacian matrix L for G is defined by

$$L_{uv} = \begin{cases} 1 - \dfrac{w_{uv}}{d_u} & \text{if } u = v,\ d_u \neq 0, \\ -\dfrac{w_{uv}}{\sqrt{d_u d_v}} & \text{if } u, v \text{ are adjacent} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where d_u and d_v are the degrees of genes u and v, respectively. The degrees of gene u (or v) describe the number of the edges that connected with u (or v). For $\lambda \geq 0$, the network-constrained logistic regression model is presented as

$$L(\lambda, \beta) = -\sum_{i=1}^{n} \left\{ y_i \log\left[f\left(X'_i\beta\right)\right] + (1 - y_i) \log\left[1 - f\left(X'_i\beta\right)\right] \right\} + \lambda\beta^T L\beta, \tag{3}$$

where the first term in (3) is the log-likelihood function of the logistic model and the second term is a network constraint based on the Laplacian matrix, which induces a smooth solution of β on the graph.

Directly computing (3) performs poorly for both prediction and biomarker selection purposes when the gene number $p \gg$ the sample size n. Therefore, the regularization approach is vitally needed. When adding a regularization term to (3), the sparse network-constrained logistic regression can be written as

$$L(\lambda_1, \lambda_2, \beta) = -\sum_{i=1}^{n} \left\{ y_i \log\left[f\left(X'_i\beta\right)\right] + (1 - y_i) \log\left[1 - \mathrm{f}\left(X'_i\beta\right)\right] \right\} + \lambda_1 \sum_{j=1}^{p} P\left(\beta_j\right) + \lambda_2\beta^T L\beta, \tag{4}$$

where $\lambda_1 > 0$ is a regularization parameter. In Zhang et al. [13], the authors used Lasso (L_1) which has the regularization term $P(\beta) = \sum_{j=1}^{p} |\beta_j|$ to penalize (4). However, the result of the Lasso type (L_1) regularization is not good enough for interpretation, especially in genomic research. Besides this, L_1 regularization is asymptotically biased [17, 18]. To improve the solution's sparsity and its predictive accuracy, we need to think beyond L_1 regularization to L_q penalties. In mathematics, L_q $(0 < q < 1)$ type regularization $|\beta|_q = \sum |\beta|^q$ with the lower value of q would lead to better solutions with more sparsity and gives asymptotically unbiased estimates [17]. Moreover, $L_{1/2}$ penalty can be taken as a representative of L_q $(0 < q < 1)$ penalty and has permitted an analytically expressive thresholding representation [6, 7]. Therefore, we proposed a novel $L_{1/2}$ net approach based on $L_{1/2}$ regularization to penalize the network-constrained logistic regression model, as shown in

$$L(\lambda_1, \lambda_2, \beta) = -\sum_{i=1}^{n} \left\{ y_i \log\left[f\left(X'_i\beta\right)\right] + (1 - y_i) \log\left[1 - f\left(X'_i\beta\right)\right] \right\} + \lambda_1 |\beta|_{1/2} + \lambda_2\beta^T L\beta, \tag{5}$$

where $|\beta|_{1/2} = \sum_{j=1}^{p} |\beta_j|^{1/2}$.

3. A Coordinate Descent Algorithm for the Network-Constrained Logistic Model with the Enhanced $L_{1/2}$ Thresholding Operator

$L_{1/2}$ penalty function is nonconvex, which raises numerical challenges in fitting the models. Recently, the coordinate

descent algorithms [19] for solving nonconvex regularization models (SCAD [20], MCP [21]) have shown significant efficiency and convergence [22]. Since the computational burden increases only linearly with the feature number p, the coordinate descent algorithm can be a powerful tool for solving high-dimensional problems. Its standard procedure can be demonstrated as follows: for every coefficient β_j ($j = 1, 2, \ldots, p$), to partially optimize the target function with respect to β_j, and fix the remaining elements β_k ($k = 1, 2, \ldots, p$ and $k \neq j$) at their most recently updated values. The specific form of updating β depends on the thresholding operator of the penalty.

In this paper, we present an enhanced $L_{1/2}$ thresholding operator for the coordinate descent algorithm:

$$\begin{aligned}&\beta_j\\&= \text{Enhanced\_Half}\left(\omega_j, \lambda\right)\\&= \begin{cases}\dfrac{2}{3}\omega_j\left(1+\cos\left(\dfrac{2\left(\pi-\varphi_\lambda\left(\omega_j\right)\right)}{3}\right)\right) & \text{if } \left|\omega_j\right| > \dfrac{\sqrt[3]{54}}{4}(\lambda)^{2/3}\\ 0 & \text{otherwise,}\end{cases}\end{aligned} \tag{6}$$

where $\varphi_\lambda(\omega) = \arccos((\lambda/8)(|\omega|/3)^{-3/2})$, $\pi = 3.14$, $\omega_j = \sum_{i=1}^{n} x_{ij}(y_i - \widetilde{y}_i^{(j)})$, and $\widetilde{y}_i^{(j)} = \sum_{k\neq j} x_{ik}\beta_k$ as the partial residual for fitting β_j.

Remark. This enhanced $L_{1/2}$ thresholding operator $(\sqrt[3]{54}/4)(\lambda)^{2/3}$ outperforms the old $L_{1/2}$ thresholding $(3/4)(\lambda)^{2/3}$ introduced in [6, 15, 16]. We know that the quantity of the regularization solutions depends seriously on the value of the regularization parameter λ. Based on this enhanced $L_{1/2}$ thresholding operator, when λ is chosen by some efficient strategies for the parameter tuning, such as cross validation, the convergence of algorithm (6) is proved [7].

The Laplacian matrix L is nonnegative definite; thus, it can be written as $L = SS^T$ by Cholesky decomposition. Following C. Li and H. Li [11] approach, (4) can be expressed as

$$\begin{aligned}L\left(\lambda_1, \lambda_2, \beta\right) = L\left(\gamma, \beta^*\right) = -\sum_{i=1}^{n}\left\{y_i^* \log\left[f\left(X^*{}_i\beta^*\right)\right]\right.\\ \left.+\left(1-y_i^*\right)\log\left[1-f\left(X^*{}_i\beta^*\right)\right]\right\} + \sum_{j=1}^{p}\gamma\left|\beta_j^*\right|^{1/2},\end{aligned} \tag{7}$$

where $X^*{}_{(n+p)\times p} = (1+\lambda_2)^{-1/2}\left(\frac{X}{\sqrt{\lambda_2}S^T}\right)$, $Y^*_{(n+p)} = \binom{Y}{0}$, $\beta^* = \sqrt{1+\lambda_2}\beta$, and γ is the regularization parameter and can be expressed as $\gamma = \lambda_1/\sqrt{1+\lambda_2}$.

One-term Taylor series expansion for (7) can be written as

$$\begin{aligned}L\left(\gamma, \beta^*\right) \approx \frac{1}{2n}\sum_{i=1}^{n}\left(Z_i - X^*{}_i\beta^*\right)' W_i\left(Z_i - X^*{}_i\beta^*\right)\\ +\sum_{j=1}^{p}P\left(\beta^*{}_j\right),\end{aligned} \tag{8}$$

where $Z_i = X^*{}_i\widetilde{\beta^*} + (y_i^* - f(X^*{}_i\widetilde{\beta^*}))/f(X^*{}_i\widetilde{\beta^*})(1-f(X^*{}_i\widetilde{\beta^*}))$ is the estimated response and $W_i = f(X^*{}_i\widetilde{\beta^*})(1-f(X^*{}_i\widetilde{\beta^*}))$ is the weight for the estimated response. $f(X^*{}_i\widetilde{\beta^*}) = \exp(X^*{}_i\widetilde{\beta^*})/(1+\exp(X^*{}_i\widetilde{\beta^*}))$ is the evaluated value under the current parameters. Thus, we can redefine the partial residual for fitting current $\widetilde{\beta^*}$ as $\check{Z}_i^{(j)} = \sum_i^n W_i(\widetilde{Z}_i - \sum_{k\neq j} x_{ik}^*\widetilde{\beta^*}_k)$ and $\omega_j = \sum_{i=1}^n x_{ij}^*(Z_i - \check{Z}_i^{(j)})$. The procedure of the coordinate descent algorithm for $L_{1/2}$ penalized network-constrained logistic model is described as follows.

Algorithm 1 (the coordinate descent algorithm for $L_{1/2}$ penalized network-constrained logistic model). We consider the following.

Step 1. Initialize all $\beta_j(m) \leftarrow 0$ ($j = 1, 2, \ldots, p$) and y^*, X^*, and set $m \leftarrow 0$, γ chosen by cross validation.

Step 2. Calculate $Z(m)$ and $W(m)$ and approximate the loss function (8) based on the current $\beta(m)$.

Step 3. Update each $\beta_j(m)$ and cycle over $j = 1, \ldots, p$, until $\beta_j(m)$ does not change.

Step 3.1. Compute $Z_i{}^{(j)}(m) \leftarrow \sum_{i=1}^n W_i(m)(Z_i(m) - \sum_{k\neq j} x_{ik}^*\beta_k(m))$ and $\omega_j(m) \leftarrow \sum_{i=1}^n x_{ij}(Z_i(m) - \check{Z}_i{}^{(j)}(m))$.

Step 3.2. Update $\beta_j(m) \leftarrow \text{Enhanced\_Half}(\omega_j(m), \gamma)$.

Step 4. Let $m \leftarrow m+1$, $\beta(m+1) \leftarrow \beta(m)$.

If $\beta(m)$ dose not converge, then repeat Steps 2 and 3.

4. Simulation and Application

4.1. Analyses of Simulated Data. We evaluate the performance of four methods: the network-constrained logistic regression models with L_1 regularization (L_1 net), $L_{1/2}$ regularization with old thresholding value $(3/4)(\lambda)^{2/3}$ ($L_{1/2}$ net) and with the enhanced thresholding value $(\sqrt[3]{54}/4)(\lambda)^{2/3}$ (enhanced $L_{1/2}$ net), and the Elastic net regularization approach (Elastic net). We first simulated the graph structure to mimic gene regulatory network: assuming that the graph consists of 200 independent transcription factors (TFs) and each TF regulates 10 unlike genes, so there are a total of 2200 variables, $X = (x_1, x_2, \ldots, x_p)$, $p = 2200$. The training and the independent test data sets include the sample sizes of 100, respectively. Each TF x_n and its regulated genes x_m were generated by the normal distribution $N(0, 1)$. We set the correlation rate between x_n and its regulated gene x_m as 0.75, $x_m = (1-0.75)\times x_m + (0.75)\times x_n$. The binary responder y_i ($1 \le i \le 100$), which is associated with the matrix X of TFs and their regulated genes, is calculated based on the following formula and rule:

$$\begin{aligned}&y_i = 1 \text{ (Label 1)},\\&\text{if } P\left(y_i = 1 \mid X_i\right) = \frac{\exp\left(X_i\beta + \epsilon\right)}{1+\exp\left(X_i\beta+\epsilon\right)} \geqslant 0.5;\ \text{else } y_i = 0 \text{ (Label 0)},\end{aligned} \tag{9}$$

Table 1: Simulation results of the enhanced $L_{1/2}$ net, $L_{1/2}$ net, L_1 net, and Elastic net, respectively.

Model	Misclassification errors (%)				Sensitivity (%)				Specificity (%)			
	Eh_$L_{1/2}$	$L_{1/2}$	L_1	Elastic	Eh_$L_{1/2}$	$L_{1/2}$	L_1	Elastic	Eh_$L_{1/2}$	$L_{1/2}$	L_1	Elastic
1	**9.22**	9.85	11.81	13.12	**0.985**	0.971	0.968	0.873	0.969	0.970	0.962	0.981
	(0.36)	(0.31)	(0.41)	(0.12)	(0.00)	(0.00)	(0.02)	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)
2	**10.76**	10.83	13.21	14.14	0.939	0.939	0.943	0.835	**0.987**	0.981	0.987	0.980
	(0.33)	(0.36)	(0.24)	(0.23)	(0.00)	(0.00)	(0.01)	(0.00)	(0.02)	(0.01)	(0.01)	(0.00)

Simulation results (averaged over 100 runs) for comparison of misclassification errors, sensitivity, and specificity used the enhanced $L_{1/2}$ net, $L_{1/2}$ net, L_1 net, and the Elastic net, respectively. The standard errors are given in parentheses.

where $\beta = (2, \underbrace{2/\sqrt{5}, \ldots, 2/\sqrt{5}}_{10}, -2, \underbrace{-2/\sqrt{5}, \ldots, -2/\sqrt{5}}_{10}, 4, \underbrace{4/\sqrt{5}, \ldots, 4/\sqrt{5}}_{10}, -4, \underbrace{-4/\sqrt{5}, \ldots, -4/\sqrt{5}}_{10}, 0, \ldots, 0)$ for Model 1, and $\varepsilon \sim N(0, \sigma_e^2)$.

Model 2 was defined similar to Model 1, except that we considered the case when the TF can have positive and negative effects on its regulated genes at the same time:

$$\beta = \Bigg(2, \frac{-2}{\sqrt{5}}, \frac{-2}{\sqrt{5}}, \frac{-2}{\sqrt{5}}, \underbrace{\frac{2}{\sqrt{5}}, \ldots, \frac{2}{\sqrt{5}}}_{7}, -2, \frac{2}{\sqrt{5}}, \frac{2}{\sqrt{5}}, \frac{2}{\sqrt{5}}, \underbrace{\frac{-2}{\sqrt{5}}, \ldots, \frac{-2}{\sqrt{5}}}_{7}, 4, \frac{-4}{\sqrt{5}}, \frac{-4}{\sqrt{5}}, \frac{-4}{\sqrt{5}}, \underbrace{\frac{4}{\sqrt{5}}, \ldots, \frac{4}{\sqrt{5}}}_{7}, 4, \frac{4}{\sqrt{5}}, \frac{4}{\sqrt{5}}, \frac{4}{\sqrt{5}}, \underbrace{\frac{-4}{\sqrt{5}}, \ldots, \frac{-4}{\sqrt{5}}}_{7}, 0, \ldots, 0 \Bigg). \quad (10)$$

In these two models, the 10-fold cross validation approach was conducted on the training datasets to tune the regularization parameters of the enhanced $L_{1/2}$ net, $L_{1/2}$ net, and L_1 net. Both penalized parameters for L_1 and ridge regularization in the Elastic net were tuned by the 10-fold cross validation on the two-dimensional parameter surfaces. We repeated the simulations over 100 times and then computed the misclassification error, the sensitivity, and the specificity averagely for each net model on the test datasets.

Table 1 summarizes the simulation results from each regularization net model. In general, our proposed enhanced $L_{1/2}$ net model achieved the smallest misclassification errors in Models 1 (9.22%) and 2 (10.76%) compared with the other regularization methods including the old $L_{1/2}$ thresholding method (9.85% for Model 1 and 10.83% for Model 2), L_1 net (11.81% for Model 1 and 13.21% for Model 2), and the Elastic net (13.12% for Model 1 and 14.14% for Model 2). Meanwhile, the enhanced $L_{1/2}$ net resulted in the highest sensitivity in Model 1 (98.5%) compared with the other methods. Moreover, the enhanced $L_{1/2}$ net obtained the best specificity in Model 2 (98.7%) amongst the other approaches. To sum up, the enhanced $L_{1/2}$ net outperforms the other three algorithms in terms of prediction accuracy, sensitivity, and specificity.

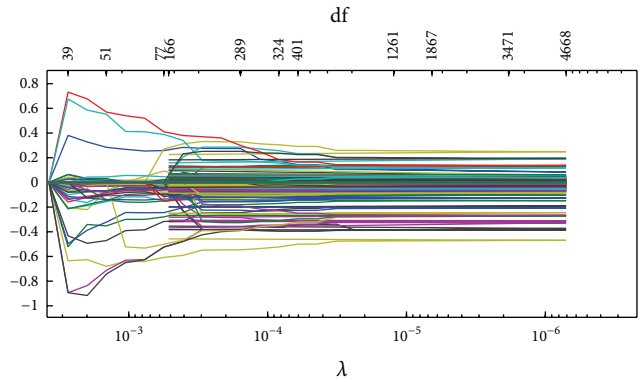

Figure 1: The solution paths of the enhanced $L_{1/2}$ net for the lung cancer dataset in one sample run.

4.2. Analysis of Lung Cancer. In this section, we merged the protein-protein interaction (PPI) network (see http://www.thebiogrid.org/) with a lung cancer (LC) gene-expression dataset [23] to demonstrate the performance of our proposed enhanced $L_{1/2}$ net method. The gene-expression dataset contains the expression profiles of 22284 genes for 107 patients, in which 58 had lung cancer. To test the generalization ability of the proposed method, we divided the dataset into the training set (sample size $n = 70$; 38 LC, 32 non-LC) which covered 2/3 samples of the dataset and the test set (sample size $n = 37$; 20 LC, 17 non-LC) which covered the other 1/3 specimens of the dataset. The 10-fold cross validation approach was conducted on the training dataset to tune the regularization parameters. By combining the gene-expression data with the PPI network, the final PPI network includes 8619 genes and 28293 edges.

Figures 1–4 display the solution paths of the four regularization net methods for the LC dataset in one sample run. Here, x-axis displays the values of the running lambda (the running lambda of L_1 penalty in the Elastic net approach), and x-axis at the top (degrees of freedom) means the number of nonzero coefficients of beta. y-axis is the values of the coefficients beta which measure the gene importance. The predictive model builds from the training set and then tests its predictive performance on the test set. The detailed results were represented in Table 2.

As shown in Table 2, the enhanced $L_{1/2}$ net selected the fewest number of genes and edges compared to $L_{1/2}$

Table 2: The results of the enhanced $L_{1/2}$ net, $L_{1/2}$ net, L_1 net, and Elastic net on LC dataset, respectively.

	Selected genes	Connected genes	Connected edges	Cross validation error	Test error
Eh_$L_{1/2}$ net	171	54	41	6/70	5/37
$L_{1/2}$ net	193	61	47	6/70	6/37
L_1 net	500	150	121	7/70	6/37
Elastic	636	337	510	6/70	6/37

Results of analysis of LC gene expression dataset by four procedures, including the number of genes selected, the number of linked PPI network genes, the number of linked PPI network edges, the CV error, and test errors.

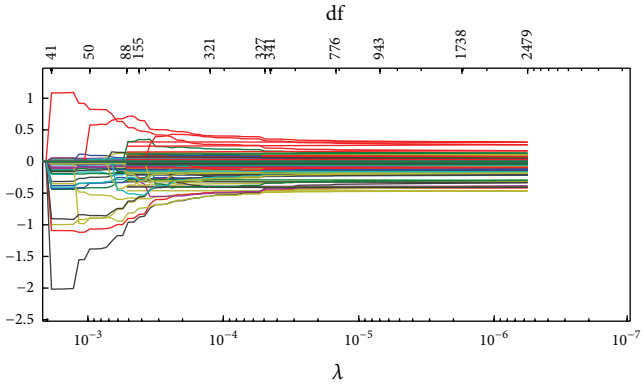

Figure 2: The solution paths of $L_{1/2}$ net for the lung cancer dataset in one sample run.

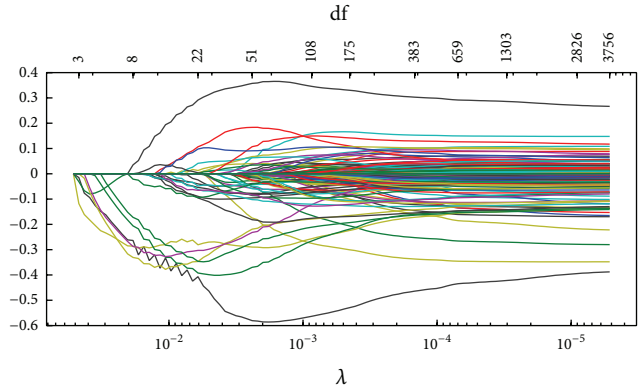

Figure 3: The solution paths of L_1 net for the lung cancer dataset in one sample run.

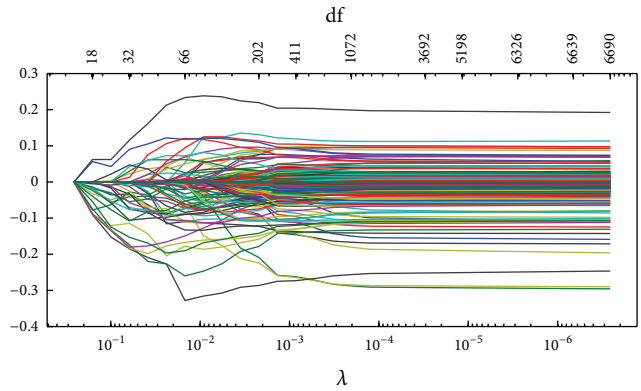

Figure 4: The solution paths of the Elastic net for the lung cancer dataset in one sample run.

net, L_1 net, and the Elastic net. Meanwhile, the predictive performance of the enhanced $L_{1/2}$ net outperforms the other three regularization net algorithms.

To further evaluate the performance of the enhanced $L_{1/2}$ net procedure, we report its capacity of identifying the biomarkers related to lung cancer. NK2 homeobox 1 (Nkx2-1) protein regulates transcription of genes specific for lung. It is used as a biomarker to determine lung cancer in anatomic pathology. It also has a critical role in maintaining lung tumor cells [24, 25]. Epidermal growth factor receptor (EGFR) is known to play a key role in cell proliferation and apoptosis. EGFR overexpression and activity could result in tumor growth and progression [26] and somatic mutations within the tyrosine kinase domain of EGFR, which have been identified in a subset of lung adenocarcinoma [27, 28]. The enhanced $L_{1/2}$ net (Figure 5) and $L_{1/2}$ net successfully identified these two important biomarkers for LC. However, neither L_1 net nor the Elastic net selected them both.

Except to identify these two significant biomarkers (EGFR and Nkx2-1), the enhanced $L_{1/2}$ net also selected several pathways that were associated with lung cancer. For example, one of the subnetworks includes genes involving molecular proliferation (e.g., genes ARF4, EGFR, DCN, BRCA1, and ITIH5). As these gene express significantly and continuously, it promotes lung cancer progression. On the other hand, this group is linked to ENO1. We are unable to get a clear testimony to sustain this relationship by looking at PPI database. However, a recent report [29] has demonstrated that ENO1 is the promising biomarker that may provide more diagnostic efficacy for lung cancer. This link implies a functional relationship and suggests the important role of ENO1 in lung cancer.

All these results reveal that the enhanced $L_{1/2}$ net is more reliable than $L_{1/2}$ net, L_1 net, and the Elastic net approaches for selecting key markers from high-dimensional genomic data. Another advantage of our proposed method is that it has the ability to recognize novel and potential relationships with biologic significance. It is mentionable that our proposed method is inclined to identify fewer but more informative genes (or edges) than L_1 net and the Elastic net approaches in genomic data and that means the proposed method has allowed the researcher to more easily concentrate on the key targets for functional studies or downstream applications.

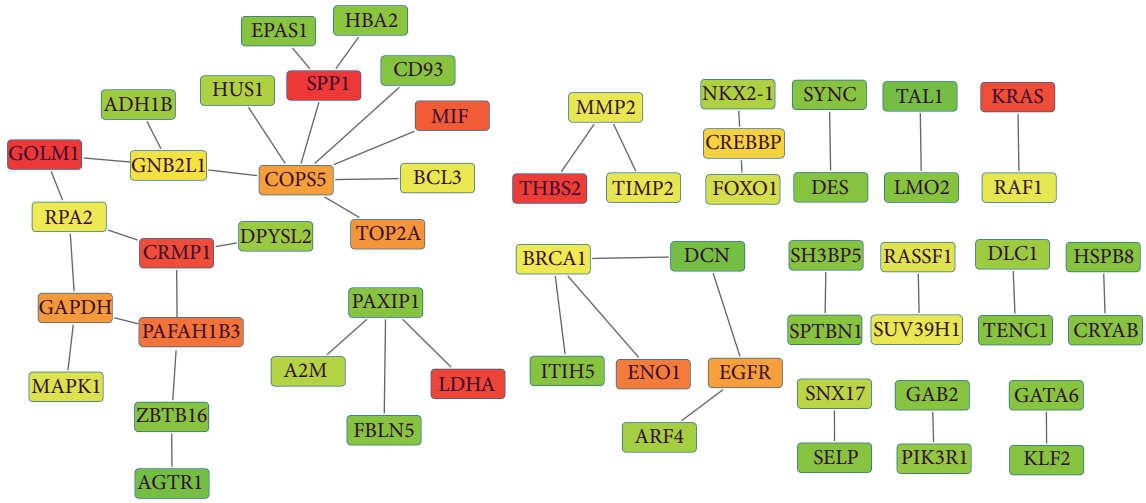

FIGURE 5: Subnetworks identified by the enhanced $L_{1/2}$ net for lung cancer datasets (only those genes that are linked on the PPI network are plotted).

5. Conclusions

In biological molecular research, especially for cancer, the analysis of combining biological pathway information with gene-expression data may play an important role to search for new targets for drug design. In this paper, we use the enhanced $L_{1/2}$ solver to penalized network-constrained logistic regression model to integrate lung cancer gene-expression with protein-to-protein interaction network. We develop the corresponding coordinate descent algorithm as a novel biomarker selection approach. This algorithm is extremely fast and easy to implement. Both simulation and real genomic data studies showed that the enhanced $L_{1/2}$ net is a ranking procedure compared with $L_{1/2}$ net (using the old thresholding operator), L_1 net, and the Elastic net in the selection of biomarker and subnetwork.

We successfully identified several important clinical biomarkers and subnetwork that are driving lung cancer. The proposed method has provided new information to investigators in biological studies and can be the efficient tool for identifying cancer related biomarker and subnetwork.

Conflict of Interests

The authors declare no conflict of interests.

Acknowledgment

This work was supported by the Macau Science and Technology Development Funds (Grant no. 099/2013/A3) of Macau SAR of China.

References

[1] S. K. Shevade and S. S. Keerthi, "A simple and efficient algorithm for gene selection using sparse logistic regression," *Bioinformatics*, vol. 19, no. 17, pp. 2246–2253, 2003.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 58, no. 1, pp. 267–288, 1996.

[3] L. Meier, S. van de Geer, and P. Bühlmann, "The group Lasso for logistic regression," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 70, no. 1, pp. 53–71, 2008.

[4] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[5] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *The Annals of Statistics*, vol. 37, no. 1, pp. 246–270, 2009.

[6] Z. B. Xu, H. Zhang, Y. Wang, X. Y. Chang, and Y. Liang, "L1/2 regularization," *Science in China Series F*, vol. 40, no. 3, pp. 1–11, 2010.

[7] Z. Xu, X. Chang, F. Xu, and H. Zhang, "$L_{1/2}$ regularization: a thresholding representation theory and a fast solver," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 7, pp. 1013–1027, 2012.

[8] J. Zeng, S. Lin, Y. Wang et al., "$L_{1/2}$ regularization: convergence of iterative half thresholding algorithm," *IEEE Transactions on Signal Processing*, vol. 62, no. 9, pp. 2317–2329, 2014.

[9] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic acids research.*, vol. 34, supplement 1, pp. D535–D539, 2006.

[10] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, article 140, 2007.

[11] C. Li and H. Li, "Network-constrained regularization and variable selection for analysis of genomic data," *Bioinformatics*, vol. 24, no. 9, pp. 1175–1182, 2008.

[12] Z. Tian, T. Hwang, and R. Kuang, "A hypergraph-based learning algorithm for classifying gene expression and arrayCGH data with prior knowledge," *Bioinformatics*, vol. 25, no. 21, pp. 2831–2838, 2009.

[13] W. Zhang, Y.-W. Wan, G. I. Allen, K. Pang, M. L. Anderson, and Z. Liu, "Molecular pathway identification using biological

network-regularized logistic models," *BMC Genomics*, vol. 14, no. 8, article S7, 2013.

[14] H. Sun, W. Lin, R. Feng, and H. Li, "Network-regularized high dimensional Cox regression for Analysis of Genomic data," *Statistica Sinica*, vol. 24, pp. 1433–1459, 2014.

[15] Y. Liang, C. Liu, X.-Z. Luan et al., "Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification," *BMC Bioinformatics*, vol. 14, no. 1, article 198, 2013.

[16] C. Liu, Y. Liang, X.-Z. Luan et al., "The L1/2 regularization method for variable selection in the Cox model," *Applied Soft Computing Journal*, vol. 14, pp. 498–503, 2014.

[17] K. Knight and W. Fu, "Asymptotics for Lasso-type estimators," *Annals of Statistics*, vol. 28, no. 5, pp. 1356–1378, 2000.

[18] D. Malioutov, M. Çetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 3010–3022, 2005.

[19] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[20] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[21] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[22] P. Breheny and J. Huang, "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *Annals of Applied Statistics*, vol. 5, no. 1, pp. 232–253, 2011.

[23] M. T. Landi, T. Dracheva, M. Rotunno et al., "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival," *PLoS ONE*, vol. 3, no. 2, Article ID e1651, 2008.

[24] Y. Li, K. Eggermont, V. Vanslembrouck, and C. M. Verfaillie, "NKX2-1 activation by SMAD2 signaling after definitive endoderm differentiation in human embryonic stem cell," *Stem Cells and Development*, vol. 22, no. 9, pp. 1433–1442, 2013.

[25] B. A. Weir, M. S. Woo, G. Getz et al., "Characterizing the cancer genome in lung adenocarcinoma," *Nature*, vol. 450, no. 7171, pp. 893–898, 2007.

[26] H. Zhang, A. Berezov, Q. Wang et al., "ErbB receptors: from oncogenes to targeted cancer therapies," *The Journal of Clinical Investigation*, vol. 117, no. 8, pp. 2051–2058, 2007.

[27] J. G. Paez, P. A. Jänne, J. C. Lee et al., "EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy," *Science*, vol. 304, no. 5676, pp. 1497–1500, 2004.

[28] S. R. Chowdhuri, L. Xi, T. H.-T. Pham et al., "EGFR and KRAS mutation analysis in cytologic samples of lung adenocarcinoma enabled by laser capture microdissection," *Modern Pathology*, vol. 25, no. 4, pp. 548–555, 2012.

[29] L. Yu, J. Shen, K. Mannoor, M. Guarnera, and F. Jiang, "Identification of ENO1 as a potential sputum biomarker for early stage lung cancer by shotgun proteomics," *Clinical Lung Cancer*, vol. 15, no. 5, pp. 372–378, 2014.