Technology and Health Care 29 (2021) S287–S295
DOI 10.3233/THC-218026
IOS Press

Cancer classification and biomarker selection via a penalized logsum network-based logistic regression model

Zhiming Zhou[a], Haihui Huang[a,b] and Yong Liang[c,*]

[a]*Faculty of Information Technology, Macau University of Science and Technology, Macau, China*
[b]*Shaoguan University, Shaoguan, Guangdong, China*
[c]*Macau Institute of Systems Engineering and Collaborative Laboratory of Intelligent Science and Systems, Macau University of Science and Technology, Macau, China*

Abstract.
BACKGROUND: In genome research, it is particularly important to identify molecular biomarkers or signaling pathways related to phenotypes. Logistic regression model is a powerful discrimination method that can offer a clear statistical explanation and obtain the classification probability of classification label information. However, it is unable to fulfill biomarker selection.
OBJECTIVE: The aim of this paper is to give the model efficient gene selection capability.
METHODS: In this paper, we propose a new penalized logsum network-based regularization logistic regression model for gene selection and cancer classification.
RESULTS: Experimental results on simulated data sets show that our method is effective in the analysis of high-dimensional data. For a large data set, the proposed method has achieved 89.66% (training) and 90.02% (testing) AUC performances, which are, on average, 5.17% (training) and 4.49% (testing) better than mainstream methods.
CONCLUSIONS: The proposed method can be considered a promising tool for gene selection and cancer classification of high-dimensional biological data.

Keywords: Regularization, gene selection, log-sum penalty, network-based knowledge

1. Introduction

Microarray technology is one of the most recent advances in cancer research, and using this method, the expression levels of thousands of genes can be recorded simultaneously. In genomic analysis, the identification of molecular biomarkers or signal pathways associated with phenotypes is a particularly important issue. Logistic regression is a powerful discrimination method, enables clear statistical interpretation, and derives classification probability of classification label information.

From a biological point of view, only a few genes are related to the target disease, and most genes are not involved in cancer classification. Unrelated genes may cause noise and reduce the accuracy of prediction. In addition, from a machine learning perspective, too many features may cause overfitting and negatively affect classification performance.

[*] Corresponding author: Yong Liang, Macau Institute of Systems Engineering and Collaborative Laboratory of Intelligent Science and Systems, Macau University of Science and Technology, Macau, China. E-mail: yliang@must.edu.mo.

The regularization method has been widely used when dealing with high-dimensional problems. A popular regularization method is the absolute shrinkage and selection operator (Lasso) or L_1 penalty [1], which can simultaneously perform feature selection and model construction. L_1 penalty extensions, such as the SCAD penalty [2], which is symmetrical, non-concave, have also been common for many years. The adaptive Lasso [3] uses dynamic weights to penalize different coefficients. However, in some cases, L_1 type regularization may result in inconsistent feature selection and often introduce extra bias in parameter estimation into the statistical model [4]. Xu et al. [4] suggested $L_{1/2}$ regularization, which and can produce a more accurate solution than the L_1 penalty [5–7]. But in the analysis of gene expression data, $L_{1/2}$ regularization may not be sparse enough. In theory, L_0 regularization produces more sparse and better solutions [8], but this is an NP problem. Therefore, Candes et al. suggested the logsum penalty [9], a method that nicely approximates the L_0 penalty. The logsum penalty has been successfully applied in much research in recent years, such as impact force recognition [10], drug discovery [11], etc. However, the logsum penalty lacks a mechanism to incorporate prior knowledge of genetic biology. Combining gene expression with the analysis of network knowledge can reduce noise and detect complicating factors in genomic data analysis of many regression and classification models [12–15].

In short, existing methods show good results in terms of feature selection and model construction, but they either cannot produce sufficient sparsity or do not use any network interaction knowledge.

In this paper, we investigate the sparse logistic regression model with a logsum network-based (Logsum-Net) penalty, in particular for gene selection in cancer classification.

The major contributions of this paper are as follows:

1. A new method of gene selection is put forward. The logsum method is an efficient tool for feature selection; however, this method is recommended from a strictly computational view, and there is no built-in design that can use a priori biological structure information. Unlike previous research, in this research, a Logsum-Net regularization that aims to integrate prior biological graph information is proposed.
2. Beyond several mainstream methods, we have carried out a simulation experiment and experimented on a breast gene expression data set, and the experimental results show that the method is feasible. For the breast cancer data set, the AUC of our method is, on average, 5.17% (training) and 4.49% (testing) better than mainstream methods.

2. The penalized logsum network-based logistic regression model

Suppose 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 the ith sample with p genes and y_i is the corresponding dependent variable consisting of a binary value of either zero or 1. Define a classifier $f(x) = e^x/(1 + e^x)$, and the logistic regression is defined as:

$$P(y_i = 1|X_i) = f(X_i'\beta) = \frac{\exp(X_i'\beta)}{1 + \exp(X_i'\beta)} \tag{1}$$

where $\beta = (\beta_1, \ldots, \beta_p)$ are the estimated variables. We transform Eq. (1) using simple algebra:

$$l(\beta) = -\sum_{i=1}^{n}\{y_i \log[f(X_i'\beta)] + (1 - y_i)\log[1 - f(X_i'\beta)]\} \tag{2}$$

Equation (2) is easily to overfitted when applied to a problem with high dimensions and low sample size. Regularization is a commonly used technique to solve high-dimensional problems, which can be expressed as:

$$L(\lambda, \beta) = l(\beta) + P(\beta), \tag{3}$$

where $P(\beta)$ is the regularization or penalty term. A typical penalty is the Lasso (L_1) method [1] which has the form $P_{\lambda,Lasso}(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|^1$, where $\lambda > 0$. With the singularity characteristic of L_1 regularization, an L_1 penalized regression model can force small coefficients to zero. There are various versions of the L_1 penalty, such as elastic net [16], SCAD [2]. However, L_1 type regularization has two drawbacks: it is biased and may not be sparse enough. Xu et al. [4] proposed an $L_{1/2}$ method of the form $P_{\lambda,L1/2}(\beta) = \lambda \sum_{j=1}^{p} |\beta_j|^{1/2}$, which is able to generate more a sparser solution than L_1 methods. However, in the analysis of genomics data, $L_{1/2}$ regularization may not be sparse enough. Theoretically, L_0 regularization produces better solutions with more sparsity, but this is an NP problem. Therefore, Candes et al. [9] proposed the logsum penalty, which approximates L_0 regularization much better. The logsum regularization is shown as follows:

$$P_{\lambda,Logsum}(\beta) = \lambda \sum_{j=1}^{p} \log(|\beta_j| + \varepsilon), \tag{4}$$

where $\varepsilon > 0$ should be set arbitrarily small to make the logsum penalty closely resemble the L_0-norm. However, the logsum penalty is unable to use any prior biological knowledge such as the gene-regulatory network, as this method was proposed from a purely computational point of view.

There has been much research into network-based penalties. Li and Li [17], Chen et al. [18] and Wang et al. [19], for example, suggested a Lasso network-based method for the study of gene expression. However, the result achieved by the Lasso method is not sparse enough for genome data. In this paper, we propose a logsum network-based (Logsum-Net) method as follows:

$$P_{\lambda_1,\lambda_2,Logsum\text{-}net}(\beta) = P_{\lambda_1,Logsum}(\beta) + \lambda_2 \beta L \beta, \tag{5}$$

where L is a symmetric Laplace matrix that combines the knowledge of biological network, and the $\beta L \beta$ term forces a smooth result on the network.

$$P_{\lambda_1,\lambda_2,Logsum\text{-}net}(\beta) = P_{\lambda_1,Logsum}(\beta) + \lambda_2 \sum_{1 \leqslant i < k \leqslant p;} w_{ik} \left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_k}{\sqrt{d_k}} \right)^2, \tag{6}$$

where $w_{ik} \in [0, 1]$ depends on whether features i and k are linked or not; d_i and d_k denote the degree (includes out or in degree) of features i and k; λ_1 and λ_2 are penalty tuning parameters. The logsum network-based logistic regression model (Logsum-NL) can be formed as follows:

$$\hat{\beta} = \text{argmin}_{\beta} \{ l(\beta) + P_{\lambda_1,\lambda_2,Logsum\text{-}Net}(\beta) \}, \tag{7}$$

This equation not only ensures sparsity in the solutions and, making them more appropriate for biological interpretation, but also smooths the regression coefficient of the genes that are connected in the network.

3. Algorithm

We first proposed a new threshold-based solver of the Logsum-Net penalty. Then, we applied an efficient coordinate descent algorithm (CDA) [20] to solve the Logsum-NL.

3.1. Threshold solver of the Logsum-Net penalty

Assuming a linear model with p predictors:

$$y = x_1\beta_1 + x_2\beta_2 + \ldots + x_p\beta_p,$$

For simplicity, the predictors and responses are all standardized. A Logsum-Net linear model can be expressed as:

$$\mathcal{L}(\beta) = \frac{1}{2}||\hat{\omega} - \beta||^2 + P_{\lambda_1,\lambda_2,\textit{Logsum-Net}}(\beta), \tag{8}$$

where $\hat{y} = X\hat{\omega}$ and $\hat{\omega} = X^T y$.

Recall that $P_{\lambda_1,\lambda_2,\textit{Logsum-Net}}(\beta) = P_{\lambda_1,\textit{Logsum}}(\beta) + \lambda_2 \sum_{1\leqslant i<k\leqslant p;} w_{ik}\left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_k}{\sqrt{d_k}}\right)^2$, which can be rewritten as:

$$\begin{aligned} P_{\lambda_1,\lambda_2,\textit{Logsum-Net}}(\beta) &= P_{\lambda_1,\textit{Logsum}}(\beta) + \lambda_2 \sum_{i=1}^{p} w_{ij}\left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_j}{\sqrt{d_j}}\right)^2 \\ &+ \lambda_2 \sum_{1\leqslant i<k\leqslant p; i,k\neq j} w_{ik}\left(\frac{\beta_i}{\sqrt{d_i}} - \frac{\beta_k}{\sqrt{d_k}}\right)^2. \end{aligned} \tag{9}$$

The first partial derivative concerning β_j of Eq. (8) is given by:

$$\frac{\partial}{\partial\beta_j}\mathcal{L}(\beta) = \beta_j - \hat{\omega} + \lambda_1\frac{1}{\beta_j + \varepsilon} + 2\lambda_2\beta_j - t, \tag{10}$$

where $t = \lambda_2 \sum_{i=1}^{p} \frac{w_{ij}\beta_i}{\sqrt{d_i d_j}}$.

By setting Eq. (9) $= 0$, a threshold-based solver for the jth item in Logsum-Net penalized linear regression model can be shown as follows:

$$\beta_j = \begin{cases} \textit{sign}(\hat{\omega}_j)\frac{c_1+\sqrt{c_2}}{2} & \text{if } c_2 > 0 \\ 0 & \text{if } c_2 \leqslant 0 \end{cases}, \tag{11}$$

where $c_1 = \frac{\hat{\omega}+t-\varepsilon-2\lambda_2\varepsilon}{1+2\lambda_2}$, $c_2 = c_1^2 - 4\left(\frac{\lambda_1-\hat{\omega}_j\varepsilon-t\zeta}{1+2\lambda_2}\right)$.

3.2. Algorithm for the Logsum-Net penalized logistic regression model

By the Taylor series method, we can rewrite Eq. (2) as follows:

$$l(\beta) \approx \frac{1}{2n}\sum_{i=1}^{n}(Z_i - X_i\beta)'W_i(Z_i - X_i\beta) \tag{12}$$

where $Z_i = X_i\tilde{\beta} + \frac{y_i - f(X_i\tilde{\beta})}{f(X_i\tilde{\beta})(1-f(X_i\tilde{\beta})}$ is the estimated response, and $W_i = f(X_i\tilde{\beta})(1-f(X_i\tilde{\beta})$ is the weight for the estimated response. $f(X_i\tilde{\beta}) = \exp(X_i\tilde{\beta})/(1+\exp(X_i\tilde{\beta}))$ is the evaluated value under the current parameters. Thus, we can redefine the partial residual for fitting the current $\tilde{\beta}$ as $\check{Z}_i^{(j)} = \sum_{k\neq j} x_{ik}\tilde{\beta}_k$ and $\omega_j = \sum_{i=1}^{n} W_i x_{ij}(Z_i - \check{Z}_i^{(j)})$. The whole algorithm for the Logsum-NL as follows:

Step 1: Set all $\beta_j \leftarrow 0$ ($j = 1, 2, \ldots, p$) and X, y, ε.
$m \leftarrow 0$, λ_1 and λ_2 are chosen by gird search;
Step 2: Calculate $Z(m)$ and $W(m)$ based on the current $\beta(m)$;

```
Step 3: Update each $\beta_j(m)$ and cycle over $j = 1, \ldots, p$
Step 3.1: Calculate $\tilde{Z}_i^{(j)}(m) \leftarrow \sum_{k \neq j} x_{ik}\beta_k(m)$, $t = \lambda_2 \sum_{i=1}^{p} \frac{w_{ij}\beta_i}{\sqrt{d_i d_j}}$.
$\omega_j(m) \leftarrow \sum_{i=1}^{n} W_i(m) x_{ij}(Z_i(m) - \tilde{Z}_i^{(j)}(m))$,
$c_1 = \frac{\omega_j(m)+t-\varepsilon-2\lambda_2\varepsilon}{1+2\lambda_2}$ and $c_2 = c_1^2 - 4\left(\frac{\lambda_1-\omega_j(m)\varepsilon-t\varepsilon}{1+2\lambda_2}\right)$.
Step 3.2: Update $\beta_j(m)$ by Eq. (11);
Step 4: Let $m \leftarrow m+1$, $\beta(m+1) \leftarrow \beta(m)$;
If $\beta(m)$ dose not convergence, then repeat Steps 2, 3.
```

4. Results

4.1. Simulation

Here, we have conducted simulation research to measure the gene selection and classify the ability of the proposed method. Several penalized logistic model technologies are compared in the experiment: the Lasso method, the $L_{1/2}$ method, the SCAD method, the elastic net method, and the L1-Net method. We follow Li's work [17] to perform a simulation experiment; that is, a graph with 200 different transcription factors (TFs) is simulated. Ten 10 genes are regulated by a TF, which means the simulated network consists of 2,200. The dependent variable y is designated as a binary value of zero or 1, and is related to the first 4 TFs and their target genes.

Two models are suggested in the simulation. In every model, there were 200 instances: 100 training and 100 for testing.

In the first model, we assumed TF and its target played an activator or repressor role in the outcome variable:

$$\beta = \left(3, \underbrace{\frac{3}{\sqrt{10}}, \ldots, \frac{3}{\sqrt{10}}}_{10}, -3, \underbrace{\frac{-3}{\sqrt{10}}, \ldots, \frac{-3}{\sqrt{10}}}_{10}, 5, \underbrace{\frac{5}{\sqrt{10}}, \ldots, \frac{5}{\sqrt{10}}}_{10}, -5, \underbrace{\frac{-5}{\sqrt{10}}, \ldots, \frac{-5}{\sqrt{10}}}_{10}, 0, \ldots, 0\right)$$

In the second model, a TF could be both an activator and repressor at the same time, and the rest setting is similar to the first model:

$$\beta = \left(3, \frac{-3}{\sqrt{10}}, \frac{-3}{\sqrt{10}}, \frac{-3}{\sqrt{10}}, \underbrace{\frac{3}{\sqrt{10}}, \ldots, \frac{3}{\sqrt{10}}}_{7}, -3, \frac{3}{\sqrt{10}}, \frac{3}{\sqrt{10}}, \frac{3}{\sqrt{10}}, \underbrace{\frac{-3}{\sqrt{10}}, \ldots, \frac{-3}{\sqrt{10}}}_{7}, \right.$$

$$\left. 5, \frac{-5}{\sqrt{10}}, \frac{-5}{\sqrt{10}}, \frac{-5}{\sqrt{10}}, \underbrace{\frac{5}{\sqrt{10}}, \ldots, \frac{5}{\sqrt{10}}}_{7}, -5, \frac{5}{\sqrt{10}}, \frac{5}{\sqrt{10}}, \frac{5}{\sqrt{10}}, \underbrace{\frac{-5}{\sqrt{10}}, \ldots, \frac{-5}{\sqrt{10}}}_{7}, 0, \ldots, 0\right)$$

The cross-validation (CV) technique has been widely used in parameter tuning. Here, we use a 10-CV method [21,22] to identify the optimal tuning parameters for the training set. Genes with zero coefficients in the predicated model will be considered irrelevant to the predictor variables [23].

To consider the impact of variable correlation on the method more fully, a variable ρ was used to control the correlation between TF and its target.

Table 1
Simulation study – gene selection performance

		Lasso		$L_{1/2}$		SCAD		ElasticNet		L_1-Net		Logsum-NL	
	ρ	P	TP	P	TP	P	TP	P	TP	P	TP	P	TP
Model 1	0.2	56.3	6.9	54.9	9.3	387.6	25	348.7	18.6	134.7	24.1	104.3	32.2
	0.5	76.7	12.4	58.7	17.7	449.7	33.7	377.4	31.2	130.2	24.8	110.9	34.5
	0.7	69.3	14.3	71.9	22.3	499.1	41.5	404.9	35.5	165.6	30.8	128.1	40.9
Model 2	0.2	53.7	7.9	47	7.9	366.4	15.2	312.2	17.8	109.4	32.8	94.1	34.7
	0.5	60.1	10.2	50.6	12	293.7	27.2	339.9	23.9	136.1	27	121.1	39.1
	0.7	61.8	10.5	61.3	10.9	380.7	32.1	372.2	30.4	215.2	36.6	128.2	38.3

Table 2
Simulation study – classification performance

		Lasso	$L_{1/2}$	SCAD	ElasticNet	L_1-Net	Logsum-NL
	ρ			Accuracy			
Model 1	0.2	85.47%	82.43%	82.10%	84.64%	85.85%	92.93%
	0.5	82.43%	80.08%	79.00%	77.38%	80.71%	92.16%
	0.7	77.96%	81.33%	78.93%	79.79%	81.44%	91.54%
Model 2	0.2	87.67%	89.60%	88.41%	86.00%	89.21%	95.18%
	0.5	84.30%	89.19%	88.33%	87.55%	87.65%	94.36%
	0.7	81.34%	85.52%	84.82%	84.09%	86.50%	92.63%

The simulation process was repeated 500 times, and we use P and TP to report the feature selection ability of this method. P refers to the number of non-zero coefficient genes in the prediction model, and TP refers to the number of true non-zero coefficient genes in the model. The classification accuracy for the test set was also calculated, and Tables 1 and 2 summarize the results of each model.

As shown in Table 1, compared to other algorithms, our method is more accurate at identifying real genes. For example, in Model 1, when $\rho = 0.7$, the mean TP identified by the Logsum-NL method is 40.9, while the number of true non-zero genes is 44. Our method selects almost entirely true genes. In addition, our method also performs well in the classification task. As dedicated in Table 2, our method has higher accuracy than sparse logistic regression models such as Lasso, $L_{1/2}$, SCAD, elastic net, and L_1-Net.

These results show that this method is a useful tool for classification and feature selection.

4.2. *Real data*

To further prove the performance of the proposed method, we compared our approach with the other five regularization methods in an analysis of TCGA breast cancer. This data describes 20,501 genes in 806 different breast cancer samples. We retained only samples with complete information. After that, 85 TNBC and 460 non-TNBC were further divided into two groups: training ($n = 327$; 51 TNBC, 276 non-TNBC) and testing ($n = 218$; 34 TNBC, 184 non- TNBC) sets.

An extensive biological interactive network was obtained from BioGrid, which consists of 15,211 nodes (gene or other entities) and 336,119 interactions. A prepared network L with 11,320 genes and 224,458 edges was gained when we were mapping the downloaded network into the gene expression data.

We also added two new methods for performance comparison: SPL-Logsum [11] and HLR [7]. Table 3 shows that the Logsum-NL method gained higher predicting AUC performance than other mainstream regularization methods.

It can be seen from Table 4 that the genes identified by our Logsum-NL method include the SplA/Ryanodine Receptor Domain and SOCS Box intron 1 (SPSB 1), which has recently been identified

Table 3
The results for breast cancer

Method	# selected genes	Training AUC (10-CV)	Testing AUC
Lasso [1]	66	81.56%	86.83%
$L_{1/2}$ [6]	50	80.82%	85.28%
SCAD [2]	54	87.67%	79.94%
ElasticNet [16]	82	80.04%	82.61%
L_1-Net [24]	133	86.97%	88.45%
HLR [7]	86	87.13%	86.90%
SPL-Logsum [11]	79	87.26%	88.71%
Logsum-NL	91	89.66%	90.02%

Table 4
The top ten ranked genes

Rank	Lasso	$L_{1/2}$	SCAD	ElasticNet	L_1-Net	HLR	SPL-Logsum	Logsum-NL
1	MYOM2	CALCA	TMEM63B	AKAP14	ATP1A2	CA12	ATP1A2	ESR1
2	HOXB9	FETUB	WFDC1	APCS	GGA1	FHOD1	ABCA8	SPSB1
3	DCBLD1	GABRB1	DAPK3	C15orf41	KCNA2	ETV4	MYOM2	DAPK3
4	PGBD5	PNMAL2	SLC25A1	GABRB1	GABRD	DAPK3	FNBP1	RTEL1
5	MYOM2	TM4SF4	OVOL1	KCNA2	GBE1	LHFP	GATA3	TM4SF4
6	GABRB1	PPP2R2D	SLC27A4	ATP1A2	UPK3A	GATA1	TM4SF4	EPHA8
7	APCS	ADAMTS6	MIF	GABRD	HEYL	DAPK3	CANT1	GBE1
8	USP9Y	KCNA2	OPN4	DCBLD1	CPNE6	UPK3A	FOXO1	PCDH10
9	HSP90AA1	ATP1A2	GADD45B	KIF15	GUCY1A2	ADAMTS6	GBE1	ATP1A2
10	KCNA2	FHOD1	OLIG1	HOXB9	TAF9B	SDC1	ABI3BP	LY6H

as spontaneously regulated during breast tumor recurrence, and necessary and sufficient for promoting tumor recurrence [25]. The estrogen receptor (ESR 1) is one of the important markers for the classification of breast cancer subtypes in clinics, which can be used to not only guide prognosis but also decide treatment [26]. In breast tumors, protoprotein 10 (PCDH10) is down-regulated and methylated excessively [27]. The lymphocyte antigen 6 family member H (LY6H) is a cancer biomarker and therapeutic target that induces invasion and metastasis. LY6H is involved in the development of breast cancer by affecting the cellular pathway Ras/ERK. This gene may be a new marker for diagnosis and gene therapy in breast cancer patients [28].

5. Discussion and conclusions

The logsum method is a powerful method for feature selection. However, it is unable to use any previous biological structure information. To overcome this drawback, in this paper we first propose Logsum-Net regularization to integrate biological network knowledge. Then, we suggest the penalized logsum-net regularization logistic regression model (logsum-NL) for gene selection and cancer classification. For a real large dataset, the proposed method has achieved 89.66% (training) and 90.02% (testing) AUC performance which are, on average, 5.17% (training) and 4.49% (testing) better than mainstream methods. Therefore, the proposed logsum-NL method is a promising tool for gene selection and cancer classification of high-dimensional biological data. The limitation of this article is that it does not include an in-depth analysis of the selected genes. Future directions for research include further analysis of the potential clinical application of the selected genes, and investigation of the method with other high-dimensional data/models.

Acknowledgments

This research was supported by Macau Science and Technology Development Funds (0158/2019/A3, 0002/2019/APD) of Macau SAR of China, Special Innovation Projects of Universities in Guangdong Province (2018KTSCX205), the Natural Science Foundation of Guangdong Province (2018A030307033) and the National Natural Science Foundation of China (6201101081, 62006155).

Conflict of interest

None to report.

References

[1] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc B. 1996; 267–88.

[2] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001; 96: 1348–60. doi: 10.1198/016214501753382273.

[3] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006; 101: 1418–29.

[4] Xu Z, Zhang H, Wang Y, Chang X, Liang Y. L1/2 regularization. Sci China Inf Sci. 2010; 53: 1159–69.

[5] Huang H-H, Liang Y. Hybrid L1/2+2 method for gene selection in the Cox proportional hazards model. Comput Methods Programs Biomed. 2018; 164: 65–73. doi: 10.1016/j.cmpb.2018.06.004.

[6] Liang Y, Liu C, Luan XZ, Leung KS, Chan TM, Xu ZB, et al. Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification. BMC Bioinformatics. 2013; 14: 198.

[7] Huang H-H, Liu X-Y, Liang Y. Feature selection and cancer classification via sparse logistic regression with the hybrid L1/2+2 regularization. PLoS One. 2016; 11: e0149675. doi: 10.1371/journal.pone.0149675.

[8] Chu G-J, Liang Y, Wang J-X. Novel harmonic regularization approach for variable selection in Cox's proportional hazards model. Comput Math Methods Med. 2014; 2014: 857398. doi: 10.1155/2014/857398.

[9] Candès EJ, Wakin MB, Boyd SP. Enhancing sparsity by reweighted l 1 minimization. J Fourier Anal Appl. 2008; 14: 877–905. doi: 10.1007/s00041-008-9045-x.

[10] Qiao B, Liu J, Liu J, Yang Z, Chen X. An enhanced sparse regularization method for impact force identification. Mech Syst Signal Process. 2019; 126: 341–67. doi: 10.1016/J.YMSSP.2019.02.039.

[11] Xia L-Y, Wang Q-Y, Cao Z, Liang Y. Descriptor selection improvements for quantitative structure-activity relationships. Int J Neural Syst. 2019; 29: 1950016. doi: 10.1142/S0129065719500163.

[12] Huang H, Liang Y. A novel Cox proportional hazards model for high-dimensional genomic data in cancer prognosis. IEEE/ACM Trans Comput Biol Bioinforma. 2019: 1–1. doi: 10.1109/TCBB.2019.2961667.

[13] Huang HH, Liu XY, Li HM, Liang Y. Molecular pathway identification using a new L1/2 solver and biological network-constrained mode. Int J Data Min Bioinform. 2017; 17: 189. doi: 10.1504/IJDMB.2017.085277.

[14] Huang H-H, Liang Y. An integrative analysis system of gene expression using self-paced learning and SCAD-Net. Expert Syst Appl. 2019; 135: 102–12. doi: 10.1016/J.ESWA.2019.06.016.

[15] Huang H-H, Liang Y, Liu X-Y. Network-based logistic classification with an enhanced L1/2 solver reveals biomarker and subnetwork signatures for diagnosing lung cancer. Biomed Res Int. 2015; 2015: 713953. doi: 10.1155/2015/713953.

[16] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B. 2005; 67: 301–20.

[17] Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics. 2008; 24: 1175–82. doi: 10.1093/bioinformatics/btn081.

[18] Chen J, Zhang S, C. S. Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. Bioinformatics. 2016; 32: 1724–32. doi: 10.1093/bioinformatics/btw059.

[19] Wang R, Su C, Wang X, Fu Q, Gao X, Zhang C, et al. Global gene expression analysis combined with a genomics approach for the identification of signal transduction networks involved in postnatal mouse myocardial proliferation and development. Int J Mol Med. 2018; 41: 311–21. doi: 10.3892/ijmm.2017.3234.

[20] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010; 33: 1–22.

[21] Chui K, Lytras M. A novel MOGA-SVM multinomial classification for organ inflammation detection. Appl Sci. 2019; 9: 2284. doi: 10.3390/app9112284.

[22] Imai S, Takekuma Y, Kashiwagi H, Miyai T, Kobayashi M, Iseki K, et al. Validation of the usefulness of artificial neural networks for risk prediction of adverse drug reactions used for individual patients in clinical practice. PLoS One. 2020; 15: e0236789. doi: 10.1371/journal.pone.0236789.

[23] Peng X, Yang Y. Algorithms for interval-valued fuzzy soft sets in stochastic multi-criteria decision making based on regret theory and prospect theory with combined weight. Appl Soft Comput. 2017; 54: 415–30. doi: 10.1016/J.ASOC.2016.06.036.

[24] Ren J, Du Y, Li S, Ma S, Jiang Y, Wu C. Robust network-based regularization and variable selection for high-dimensional genomic data in cancer prognosis. Genet Epidemiol. 2019; 43: 276–91. doi: 10.1002/gepi.22194.

[25] Feng Y, Pan T-C, Pant DK, Chakrabarti KR, Alvarez JV, Ruth JR, et al. SPSB1 promotes breast cancer recurrence by potentiating c-MET signaling. Cancer Discov. 2014; 4: 790–803. doi: 10.1158/2159-8290.CD-13-0548.

[26] Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490: 61–70. doi: 10.1038/nature11412.

[27] Li Z, Chim JCS, Yang M, Ye J, Wong BCY, Qiao L. Role of PCDH10 and its hypermethylation in human gastric cancer. Biochim Biophys Acta – Mol Cell Res. 2012; 1823: 298–305. doi: 10.1016/J.BBAMCR.2011.11.011.

[28] Kong HK, Yoon S, Park JH. The regulatory mechanism of the LY6K gene expression in human breast cancer cells. J Biol Chem. 2012; 287: 38889–900. doi: 10.1074/jbc.M112.394270.