International Journal of Automation and Computing 13(2), April 2016, 191-198
DOI: 10.1007/s11633-015-0919-5

Biomarker Identification of Rat Liver Regeneration via Adaptive Logistic Regression

Liu-Yuan Chen[1,2] Jie Yang[1] Guo-Guo Xu[3] Yun-Qing Liu[3] Jun-Tao Li[2] Cun-Shuan Xu[3]

[1]School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China
[2]College of Mathematics and Information Sciences, Henan Normal University, Xinxiang 453007, China
[3]State Key Laboratory Cultivation Base for Cell Differentiation Regulation, Henan Normal University, Xinxiang 453007, China

Abstract: This paper is devoted to identifying the biomarkers of rat liver regeneration via the adaptive logistic regression. By combining the adaptive elastic net penalty with the logistic regression loss, the adaptive logistic regression is proposed to adaptively identify the important genes in groups. Furthermore, by improving the pathwise coordinate descent algorithm, a fast solving algorithm is developed for computing the regularized paths of the adaptive logistic regression. The results from the experiments performed on the microarray data of rat liver regeneration are provided to illustrate the effectiveness of the proposed method and verify the biological rationality of the selected biomarkers.

Keywords: Adaptive logistic regression, gene selection, microarray classification, grouping effect, rat liver regeneration.

1 Introduction

Microarray classification is performed on "high dimension, small samples" data, where the number of genes is much larger than the number of samples[1, 2]. Hence, one of the important problems is to identify a small number of discriminatory biomarker genes[1−4]. Support vector machine (SVM) is an important statistical learning method, and has been widely applied to artificial intelligence area, such as ontology matching[5], classifiction of spectra of emission line stars[6], multivariate calibration[7], degree prediction of malignancy in brain glioma[8] and so on. Following the same idea, the machine-learning-based methods for gene selection and microarray have attracted much attention in bioinformatics[9−18].

The standard L_2-norm support vector machine[1−3] is the most typical learning machine for selecting genes before classification. Since it can make some of the fitted coefficients be exactly zero, the L_1-norm penalty has the advantage of automatically selecting relevant variables[10−12]. Lasso is the typically L_1-norm penalized learning machine[10]. Combining the L_1-norm penalty with the hinge loss, 1-norm support vector machine was proposed[11]. Combining the L_1-norm penalty with the logistic loss, the sparse logistic regression was proposed[12]. Note that non-convex L_P-norm penalty has the similar feature for the L_1-norm penalty. A new sparse logistic regression[13] for automatic gene selection was proposed by combining the logistic loss with non-convex L_p-norm penalty. To select genes in groups, the elastic net was proposed in [14]. Following the same idea, many elastic net penalized methods, such as the doubly regularized support vector machine (DRSVM)[15], the huberized support vector machine (HSVM)[16], were proposed. In order to adaptively control the size of the selected groups, the adaptive elastic net penalized methods were proposed[17, 18]. Especially, the partly adaptive elastic net (PAEN)[18] was proposed by introducing the proper data-driven weights to the penalty terms. Since the same weight is imposed on both L_1-norm penalized coefficient and L_2-norm penalized coefficient, PAEN can automatically identify the significant genes within each group and thus encourage an adaptive grouping effect. Motivated by the properties of gene selection proposed by Li et al.[18] and the idea of the data driven proposed by Meng et al.[19], this paper proposes an adaptive logistic regression for identifying the biomarkers of rat liver regeneration by combining the logistic loss and the adaptive elastic net penalty. The paper is organized as follows. Section 2 presents the preliminary of the problem. Section 3 gives the statistical model and the property of the adaptive logistic regression. Experimental results obtained on the microarray data of rat liver regeneration are presented in Section 4. Finally, Section 5 is a summary about the work.

2 Problem formulation and preliminary

Given a training data set for binary microarray classification problem, $D = \{(x_i, y_i)\}_{i=1}^n$, where $x_i (x_i = x_{i1}, x_{i2}, \cdots, x_{ip})$ is a multidimensional input vector with dimension

Research Article
Manuscript received April, 2014; accepted July 8, 2014; published online January 11, 2016
This work was supported by National Nature Science Foundation of China (No. 61203293), Key Scientific and Technological Project of Henan Province (No. 122102210131), Program for Science and Technology Innovation Talents in Universities of Henan Province (No. 13HASTIT040), Foundation of Henan Educational Committee (No. 13A120524), Henan Normal University Doctoral Topics (No. qd14156), Henan Higher School Funding Scheme for Young Teachers (No. 2012GGJS-063).
Recommended by Associate Editor Matjaz Gams

p, x_{ip} is the expression levels of p genes of the i-th observation, $y_i \in \{+1,\ -1\}$ represents the class label corresponding to input x_i. Similar to [16−18], we define the following notations:

$$Y = (y_1, \cdots, y_n)^{\mathrm{T}}$$
$$X = (x_1; x_2; \cdots; x_n) = \left(x_{(1)}, x_{(2)}, \cdots, x_{(p)}\right)$$

and assume that the response y is centered and x_{ij} are standardized, i.e.,

$$\sum_{i=1}^{n} y_i = 0, \qquad \frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0, \qquad \sum_{i=1}^{n} x_{ij}^2 = 1. \tag{1}$$

In this paper, we focus on fitting the data by using the regularized linear logistic regression model. Similar to [13], the logistic regression model represents the class-conditional probabilities through a linear function of the predictors

$$\Pr(Y = +1|x) = \frac{1}{(1 + \mathrm{e}^{-(\beta_0 + x^{\mathrm{T}}\beta)})}$$
$$\Pr(Y = -1|x) = \frac{1}{(1 + \mathrm{e}^{+(\beta_0 + x^{\mathrm{T}}\beta)})} =$$
$$1 - \Pr(Y = +1|x). \tag{2}$$

It can be easily obtained that

$$\log\left(\frac{\Pr(Y = +1|x)}{\Pr(Y = -1|x)}\right) =$$
$$\beta_0 + x^{\mathrm{T}}\beta = \log\Pr(Y = +1|x) - \log\Pr(Y = -1|x). \tag{3}$$

Let $p(x_i) = \Pr(Y = +1|x)$ be the probability (2) for observation i at a particular value for the parameter pairs (β_0, β). The maximized log-likelihood with penalty is as follows

$$\max_{(\beta_0,\beta)\in\mathbf{R}^{p+1}} \left[\frac{1}{n}\sum_{i=1}^{n}\{I(y_i = 1)\log\, p(x_i) + I(y_i = -1)\log(1 - p(x_i))\} - J(\lambda, \beta)\right] \tag{4}$$

where

$$I(y_i = 1) = \begin{cases} 1, & \text{if } y_i = 1 \\ 0, & \text{otherwise} \end{cases}$$
$$I(y_i = -1) = \begin{cases} 1, & \text{if } y_i = -1 \\ 0, & \text{otherwise.} \end{cases}$$

Note that

$$\frac{1}{n}\sum_{i=1}^{n}\{I(y_i = 1)\log\, p(x_i) + I(y_i = -1)\log(1 - p(x_i))\} =$$
$$\frac{1}{n}\sum_{i=1}^{n} I(y_i = 1)(\beta_0 + x_i^{\mathrm{T}}\beta) + \log(1 - p(x_i)).$$

Hence, the log-likelihood part of (4) can be rewritten as

$$l^*(\beta_0, \beta) = \frac{1}{n}\sum_{i=1}^{n}\left[I(y_i = 1)(\beta_0 + x_i^{\mathrm{T}}\beta) - \log(1 + \mathrm{e}^{\beta_0 + x_i^{\mathrm{T}}\beta})\right]. \tag{5}$$

Let $l(\beta_0, \beta) = -l^*(\beta_0, \beta)$. Since $l(\beta_0, \beta) \geq 0$. Hence, $l(\beta_0, \beta)$ can be defined as log-likelihood loss function.

According to [20], the elastic net can be represented as

$$\hat{\beta}(en) = \arg\min_{\beta} \|y - X\beta\|^2 + \lambda\left(\alpha\|\beta\|_1 + (1 - \alpha)\|\beta\|^2\right) \tag{6}$$

where $\|\beta\|^2 = \sum_{j=1}^{p}\beta_j^2$, $\|\beta\|_1 = \sum\limits_{j=1}^{p}|\beta_j|$, and $\lambda, \alpha > 0$ are the parameters. The advantage of this form is that the elastic net can be easily solved by using the pathwise coordinate descent algorithm. Hence, the elastic net is used as the initial estimator in the PAEN[18] to construct the partly adaptive elastic net penalty.

3 Adaptive logistic regression

3.1 Statistical model

Given a training pairs $\{(x_i, y_i)\}_{i=1}^{n}$ and a small constant α_0 (usually $\alpha_0 \leq 0.05$), where x_i^{T} is the input vector, and y_i indicates its class label as aforementioned. Similar to [18], we let $\hat{\beta}(\alpha_0)$ denote the optimal solution of the elastic net (6) which gives the smallest cross-validated prediction error. Since the magnitude of $\hat{\beta}_j(\alpha_0)$ implies the contribution of gene j to the classifier to some extent, we can use $|\hat{\beta}_j(\alpha_0)|$, $j = 1, \cdots, p$, to rank genes roughly. In the following, we suppose that the predictors $x_{(1)}, \cdots, x_{(p)}$ are ranked in the following way:

$$|\hat{\beta}_1(\alpha_0)| \geq |\hat{\beta}_2(\alpha_0)| \geq \cdots \geq |\hat{\beta}_p(\alpha_0)| \geq 0.$$

Let m_δ be the largest index number of the data set $\{j : |\hat{\beta}_j(\alpha_0)| \geq \delta\}$, without loss of generality, here we still let X denote the transformed model matrix. The following partly adaptive elastic net penalty was proposed[18] as

$$\lambda((1 - \alpha)\|\sqrt{W}\beta\|^2 + \alpha\|W\beta\|_1) \tag{7}$$

where $W = \mathrm{diag}\{w_1, \cdots, w_{m_\delta}, \frac{1}{\delta}\cdots, \frac{1}{\delta}\}$, $w_j = |\hat{\beta}_j(\alpha_0)|^{-1}$, $\|\sqrt{W}\beta\|^2 = \sum_{j=1}^{m_\delta} w_j\beta_j^2 + \frac{1}{\delta}\sum_{m_\delta+1}^{p}\beta_j^2$, $\|W\beta\|_1 = \sum_{j=1}^{m_\delta} w_j|\beta_j| + \frac{1}{\delta}\sum_{m_\delta+1}^{p}|\beta_j|$. Applying the partly adaptive elastic net penalty (7) to the log-likelihood loss (6), we propose the following adaptive logistic regression:

$$\hat{\beta} = \arg\min_{(\beta,\beta_0)}\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[I(y_i = 1)(\beta_0 + x_i^{\mathrm{T}}\beta) - \log(1 + \mathrm{e}^{\beta_0 + x_i^{\mathrm{T}}\beta})\right] + \lambda\left((1 - \alpha)\|\sqrt{W}\beta\|^2 + \alpha\|W\beta\|_1\right)\right\}. \tag{8}$$

3.2 Adaptive grouping effect

In the problem with large p and small n, the grouped variable selection is particularly important[1, 12, 17−18]. It is well-known that the elastic net penalized methods can encourage the grouped effect in gene selection[14−16]. In the following, it will be shown that the adaptive logistic regression can automatically identify the significant genes within each group, thus encouraging an adaptive grouped effect.

For the adaptive logistic regression (8), suppose that the predictors $x_{(j)}$, $j = 1, \cdots, p$, are standardized. If $\hat{\beta}_j \hat{\beta}_l > 0$ holds for $j, l \leq m_\delta$, then we have

$$|\hat{\beta}_j - \hat{\beta}_l| \leq \frac{1}{\sqrt{n}\lambda(1-\alpha)} \sqrt{\hat{\beta}_j{}^2(\alpha_0) + \hat{\beta}_l{}^2(\alpha_0)} \times \sqrt{1-\gamma\rho} \quad (9)$$

where

$$\rho = cor(x_{(j)}, x_{(l)}) = x_{(j)}^{\mathrm{T}} x_{(l)} = \sum_{i=1}^{n} x_{ij} x_{il}$$

$$\gamma = \frac{2|\hat{\beta}_j(\alpha_0)\hat{\beta}_l(\alpha_0)|}{\hat{\beta}_j^2(\alpha_0) + \hat{\beta}_l^2(\alpha_0)}.$$

Proof. Note that

$$\frac{(\partial I(y_i = 1)(\beta_0 + x_i^{\mathrm{T}}\beta) - \log(1 + \mathrm{e}^{\beta_0 + x_i^{\mathrm{T}}\beta}))}{\partial(\beta_0 + x_i^{\mathrm{T}}\beta)} = I(y_i = 1) - 1 + \frac{1}{(1 + \mathrm{e}^{\beta_0 + x_i^{\mathrm{T}}\beta})} \leq 1.$$

Hence, the log-likelihood loss function is Lipschitz continuous, i.e., for any $(\hat{\beta}_0, \hat{\beta})$ and $(\hat{\beta}_0^*, \hat{\beta}^*)$, the following inequality

$$|l(\hat{\beta}_0, \hat{\beta}) - l(\hat{\beta}_0^*, \hat{\beta}^*)| \leq \frac{1}{n} \sum_{i=1}^{n} \left| (\hat{\beta}_0 + x_i^{\mathrm{T}}\hat{\beta}) - (\hat{\beta}_0^* + x_i^{\mathrm{T}}\hat{\beta}^*) \right| \quad (10)$$

holds. Denote $(\hat{\beta}_0, \hat{\beta})$ be the solution of the adaptive logistic regression (8). Consider another set of coefficients

$$\hat{\beta}_0^* = \hat{\beta}_0$$

$$\hat{\beta}_{j'}^* = \begin{cases} \dfrac{\hat{\beta}_j w_j}{(w_j + w_l)} + \dfrac{\hat{\beta}_l w_l}{(w_j + w_l)}, \\ \qquad \text{if} \quad j' = j \ \text{ or } \ j' = l \\ \hat{\beta}_{j'}, \quad \text{otherwise.} \end{cases}$$

Note that adaptive logistic regression (8) is a minimization problem. Hence, we have

$$-\frac{1}{n}\sum_{i=1}^{n}\left[I(y_i = 1)(\hat{\beta}_0^* + x_i^{\mathrm{T}}\hat{\beta}^*) - \log(\mathrm{e}^{\hat{\beta}_0^* + x_i^{\mathrm{T}}\hat{\beta}} + 1)\right] + \lambda\left((1-\alpha)\|\sqrt{W}\hat{\beta}^*\|^2 + \alpha\|W\hat{\beta}^*\|_1\right) - \left\{-\frac{1}{n}\sum_{i=1}^{n}\left[I(y_i = 1)(\hat{\beta}_0 + x_i^{\mathrm{T}}\hat{\beta}) - \log(\mathrm{e}^{\hat{\beta}_0 + x_i^{\mathrm{T}}\hat{\beta}} + 1)\right] + \lambda\left((1-\alpha)\|\sqrt{W}\hat{\beta}\|^2 + \alpha\|W\hat{\beta}\|_1\right)\right\} \geq 0. \quad (11)$$

Following the similar procedure in [18], we have

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{I(y_i = 1)(\hat{\beta}_0^* + x_i^{\mathrm{T}}\hat{\beta}^*) - \log(\mathrm{e}^{\hat{\beta}_0^* + x_i^{\mathrm{T}}\hat{\beta}} + 1) - I(y_i = 1)(\hat{\beta}_0 + x_i^{\mathrm{T}}\hat{\beta}) + \log(\mathrm{e}^{\hat{\beta}_0 + x_i^{\mathrm{T}}\hat{\beta}} + 1)\right\} = \frac{1}{(n(w_j + w_l))}|\hat{\beta}_j - \hat{\beta}_l| \times \| w_l x_{(j)} - w_j x_{(l)} \|_1. \quad (12)$$

For $j, l \leq m_\delta$, it can be easily obtained that

$$\|\sqrt{W}\hat{\beta}^*\|^2 - \|\sqrt{W}\hat{\beta}\|^2 = -\frac{w_j w_l}{(w_j + w_l)(\hat{\beta}_j - \hat{\beta}_l)^2} \quad (13)$$

$$\|W\hat{\beta}^*\|_1 - \|W\hat{\beta}\|_1 = |w_j\hat{\beta}_j + w_l\hat{\beta}_l| - w_j\hat{\beta}_j - w_l\hat{\beta}_l \leq 0. \quad (14)$$

Substituting (12)−(14) into (11) yields

$$\frac{|\hat{\beta}_j - \hat{\beta}_l|}{n(w_j + w_l)}\|w_l x_{(j)} - w_j x_{(l)}\|_1 - \lambda(1-\alpha)(\hat{\beta}_j - \hat{\beta}_l)^2 \frac{w_j w_l}{(w_j + w_l)} \geq 0. \quad (15)$$

It follows from (15) that

$$|\hat{\beta}_j - \hat{\beta}_l| \leq \frac{\|w_l x_{(j)} - w_j x_{(l)}\|_1}{(n w_j w_l \lambda(1-\alpha))}. \quad (16)$$

Note that

$$\| w_l x_j - w_j x_l \|_1 \leq \sqrt{n}\sqrt{\sum_{i=1}^{n}(w_l x_{ij} - w_j x_{il})^2} = \sqrt{n}\sqrt{w_j^2 + w_l^2}\sqrt{1-\gamma\rho}. \quad (17)$$

Hence, substituting (17) into (16) yields (9). □

It should be noted that Theorem 1 still holds for $j \geq m_\delta$ and $l \leq m_\delta$. The only difference is to substitute δ for $\hat{\beta}_l(\alpha_0)$. Hence, for $j \geq m_\delta$ and $l \geq m_\delta$, the following Corollary 1 holds.

Corollary 1. Suppose that the predictors $x_{(j)}$, $j = 1, \cdots, p$ are standardized. If $\hat{\beta}_j \hat{\beta}_l > 0$ holds for $j, l \geq m_\delta$, then we have

$$|\hat{\beta}_j - \hat{\beta}_l| \leq \frac{\sqrt{2(1-\rho)}\delta}{\sqrt{n}\lambda(1-\alpha)}. \quad (18)$$

This is a special situation of (9), when $w_j = w_l = \frac{1}{\delta}$.

Remark 1. Similar to [18], the adaptive logistic regression will assign similar coefficients to the predictors only when $\rho = 1$ and $|\hat{\beta}_j(\alpha_0)| = |\hat{\beta}_l(\alpha_0)|$. Hence, the more genes with similar ranking significance ($|\tilde{\beta}_i| \approx |\tilde{\beta}_j|$), the bigger size of the selected gene groups. This implies that the adaptive logistic regression can adaptively control the size of the selected groups and therefore automatically identify the significant genes within each group.

3.3 Algorithm

Note that the traditional convex optimization methods cannot be used to solve the adaptive logistic regression due to involving the concave function of the parameters. Motivated by [20], the Newton algorithm was used to solve it. Suppose that $(\tilde{\beta}_0, \tilde{\beta})$ are the current estimates of the parameters. Similar to [20], the following quadratic approximation to the log-likelihood (Taylor expansion about current estimates) was used

$$l_Q = -\frac{1}{2n}\sum_{i=1}^{n}\eta_i(\gamma_i - \beta_0 - x_i^{\mathrm{T}}\beta)^2 + C(\tilde{\beta}_0, \tilde{\beta})^2$$

where

$$\gamma_i = \frac{\tilde{\beta}_0 + x_i^{\mathrm{T}}\tilde{\beta} + (I(y_i = 1) - \tilde{p}(x_i))}{(\tilde{p}(x_i)(1 - \tilde{p}(x_i)))} \tag{19}$$

$$\eta_i = \tilde{p}(x_i)(1 - \tilde{p}(x_i)) \tag{20}$$

and $\tilde{p}(x_i)$ is evaluated at current parameters. The Newton update is obtained by minimizing l_Q. For each value of λ, an outer loop which computes the quadratic approximation l_Q about the current parameters $(\tilde{\beta}_0, \tilde{\beta})$ was created in [20]. We can get the solving algorithm of the weighted least-squares with partly adaptive elastic net penalty

$$\min_{(\beta_0,\beta)\in\mathbf{R}^{p+1}} \left[-l_Q + \lambda\left((1-\alpha)\|\sqrt{W}\beta\|^2 + \alpha\|W\beta\|_1\right)\right]. \tag{21}$$

The algorithm solving the adaptive logistic regression proceeds as follows:

Algorithm 1.

Compute the adaptive weight matrix W.

1) Select α_0 $(0.02 \leq \alpha_0 \leq 0.05)$, scalars k_1 and k_2.
2) Compute the entire regularization solution path $\hat{\beta}(\alpha_0, \lambda)$ by using pathwise coordinate descent algorithm in [20].
3) Determine the optimal model by cross-validation and give its solution $\hat{\beta}(\alpha_0)$.
4) Determine δ, m_δ according to $\hat{\beta}(\alpha_0)$, k_1 and k_2.
5) Compute the weight matrix W.

Solve the adaptive logistic regression.

1) Let the weighting coefficients be penalty factor. Select new $\alpha(\alpha \geq 0.5)$ and lambda sequence.
2) Solve (21) by using pathwise coordinate descent algorithm.

Outer loop: Decrement λ.

Middle loop: Update the quadratic approximation l_Q using the current parameters $(\tilde{\beta}_0, \tilde{\beta})$.

Inner loop: Run the coordinate descent algorithm on the weighted least-squares with partly adaptive elastic net penalty (21).

3) Determine the optimal model by cross-validation.

Extract the non-zero coefficients and determine their corresponding genes.

Remark 2. α_0 is the model parameter of the initial elastic net and its value determines the nonzero coefficients whose reciprocals are used to construct the weights. To guarantee the enough valid weights, the initial α_0 should be smaller. So, we let $\alpha_0 \leq 0.05$. If α_0 is too small, all of the nonzero coefficients tend to the same value and the constructed weights are meaningless. So, we let $\alpha_0 \geq 0.02$.

4 Experiments on microarray data

To illustrate the effectiveness of the proposed method, we use the gene expression data of rat liver regeneration to perform our experiments. The gene expression data of rat liver regeneration are successfully produced in the cell differentiation regulation and control of Henan provincial and ministerial jointly built State-Class Key Lab. This microarray data have not been published. The brief description of the data is provided as follows.

Adult male Sprague-Dawley rats (12-weeks old), each weighted 230±20 g, were obtained from animal center of Henan Normal University. A total of 114 rats were randomly divided into 9 groups for two-third hepatectomy (PH), 9 groups for sham operation (SO) and one control group with 6 rats per group. Rats in PH groups were subjected to PH following the method of Higgins and Anderson[21]. Isolation and identification of hepatocytes (HCs) from rat regenerating liver were obtained according to the method previously described by Xu et al.[22]. Total RNA was extracted, purified[23] and detected by Rat Genome 230 2.0 microarray following the protocols previously described. To minimize the technical errors from microarray experiments, isolated hepatocytes from control groups and PH groups were detected by Rat Genome 230 2.0 Array for at least three times[24−26].

Each chip contained 31 099 genes. After removal of the duplicate data, 24 618 genes are left. In our experiments, we let the label of the samples (chips) from partial hepatectomy be 1, and the label of the samples (chips) from sham operation be −1. We randomly select two-third samples for the training, and the rest for testing. We compute the regularization solution path of the adaptive logistic regression and select the corresponding genes according to the algorithm in the above section. The entire process is repeated 10 times and the genes which could be selected in common are considered to be the correlated genes for liver regeneration. Table 1 lists the top 10 genes which are believed to be highly relevant to the rat liver regeneration. we compare the partly adaptive elastic net[18] with the adaptive logistic regression. The test accuracy and the number of the selected genes are summarized in Table 2. Compared with the adaptive logistic regression, the partly adaptive elastic net (PAEN) method[18] could not perform well. In addition, we do not find the regulatory relations among the genes selected by the partly adaptive elastic net. The probable reason why the adaptive logistic regression is superior to the partly adaptive elastic net is the former represents the class-conditional probabilities through a linear function of the predictors which fits better such data.

In our experiment, Genes (in this case, gene means the gene symbol) M6PR, IGF2 and IGF2R are selected as a group, and genes MCM5 and STAT1 are selected as another group. To demonstrate the rationality of the obtained gene groups, we also construct gene regulatory networks among the selected genes by using the business software Pathway studio 8.0. Fig. 1 shows that the 23 genes selected by adaptive logistic regression have the regulatory relations (direct regulation or promoter binding). Fig. 2 shows the two pathways related to cell proliferation.

Table 1 Some genes selected via the adaptive logistic regression

Gene symbol	Gene title	Possible function
STAT1	Signal transducer and activator of transcription 1	STAT1 regulates cell proliferation and survival. It also suppresses liver regeneration and hepatocyte proliferation in mice[27−29].
MCM5	Minichromosome maintenance complex component 5	The MCM5 protein is essential for the induction of Stat1 target gene expression in response to IFN-γ stimulation, and it is important in cell cycle and DNA replication[30].
HSPD1	Heat shock 60kDa protein 1	HSP60 is crucial for cell survival, and whole-body Hsp60 deficiency leads to cellular apoptosis and early embryonic death[31].
M6PR/IGF2R	Mannose-6-phosphate Receptor/insulin-like growth factor-II receptor	M6P/IGF2R may enhance activation of TGF-β to regulate cell proliferation and cell growth. And M6PR plays an important role in the intracellular transport of lysosomal enzymes[32−34].
IGF2	Insulin-like growth factor-II	The insulin-like growth factors possess growth-promoting activity.
LRP1	Low density lipoprotein Receptor-related Protein 1	LRP1 mediates the endocytotic clearance of a multitude of extracellular ligands and regulates diverse signaling processes such as growth factor signaling, inflammatory signaling pathways, a poptosis, and phagocytosis in liver[35].
STIP1	Stress-induced-Phosphoprotein 1	STIP1 mediates the association of the molecular chaperones HSC70 and HSP90 (HSPCA and HSPCB). And the STI1-PrP(C) complex may play a critical role in neural progenitor/stem cells self-renewal via the modulation of cell proliferation[36].
NOTCH1	Notch 1	Notch1 is one member of notch ligands, and the notch pathway is important for cell fate determination, tissue patterning and morphogenesis, and cell differentiation, proliferation and death[37].
CDKN1A	Cyclin-dependent Kinase Inhibitor 1A (P21, Cip1)	p21(Cip1) protein can play a vital role in cell cycle progression, pro-proliferative and survival[38].

Table 2 Comparison of the test error and the selected genes

Method	Test error	Number of the selected genes
Partly adaptive elastic net	69.35%	309
Adaptive logistic regression	76.23%	321

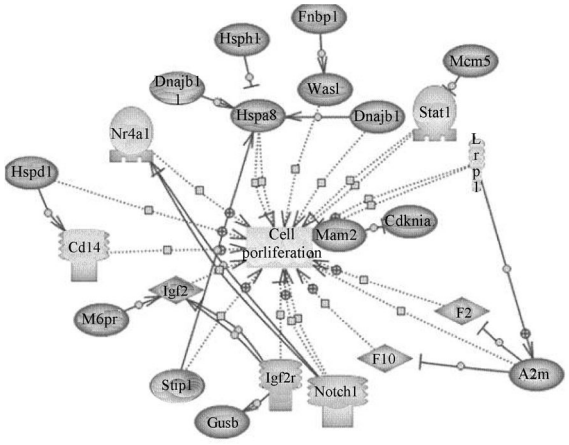

Fig. 1 The regulatory relations of the selected genes

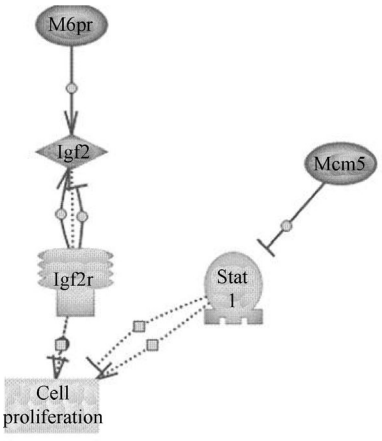

Fig. 2 Two pathways related to cell proliferation

The first pathway is M6PR→IGF2R. This pathway is insulin-like growth factor receptor signaling pathway. Insulin-like growth factors (IGFs) family includes IGF1, IGF2, IGF1R, IGF2/M6PR, and IGF binding proteins (IGFBPs). IGFs are synthesized and secreted mainly by liver from endocrine pathway and play important roles in growth, development and metabolism processes, whereas most other tissues of body can also secrete IGFs by autocrine or paracrine. Down-regulated expression of IGF2/M6PR results in up-regulated expression of the IGF2 in early liver cancer, which implies that IGFs system may play a crucial role in regulating hepatocyte proliferation in rat liver regeneration.

The second pathway is MCM5→STAT1. DNA replication licensing factor MCM5 is a protein involved in the initiation of DNA replication. The expression level of MCM5 has been considered as a criterion to reflect cell proliferation. MCM5 is up-regulated in the transition from the G0 to G1/S phase and may actively participate in cell cycle regulation. MCM5 is essential for Stat1-mediated transcriptional activation. Signal transducer and activator of transcription 1 (STAT1) is a member of the signal transducers and activators of transcription family. Literature [27] shows that STAT1 can regulate liver cells proliferation by INF-gamma.

It should be noted that the genes in the two pathway are in accord with the genes selected as groups by adaptive logistic regression. These genes are known as the key genes for liver regeneration since they are highly correlated to the physiological activity of cell proliferation in the process of liver regeneration. This successfully illustrates not only the effectiveness of the adaptive logistic regression but also the biological rationality of the selected genes. Hence, these genes are regarded as the biomarkers of rat liver regeneration.

5 Conclusions and future works

The adaptive logistic regression is proposed for identifying the biomarkers of rat liver regeneration. It has been shown that the adaptive logistic regression can encourage an adaptive grouping effect in the process of automatic gene selection. Particularly, the selected genes are verified to be highly correlated to the physiological activity of cell proliferation. It is important to compare the adaptive logistic regression with other existing methods, to find the regulatory relations among the selected genes and then to build the gene regulatory network. We leave these issues for future research.

References

[1] I. Guyon, J. Weston, S. Barnhill, V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, vol. 286, no. 5439, pp. 531–537, 1999.

[3] H. Choi, D. Yeo, S. Kwon, Y. Kim. Gene selection and prediction for cancer classification using support vector machines with a reject option. *Computational Statistics & Data Analysis*, vol. 55, no. 5, pp. 1897–1908, 2011.

[4] M. Y. You, G. Z. Li. Feature selection for multi-class problems by using pairwise-class and all-class techniques. *International Journal of General Systems*, vol. 40, no. 4, pp. 381–394, 2011.

[5] L. Liu, F. Yang, P. Zhang, J. Y. Wu, L. Hu. SVM-based ontology matching approach. *International Journal of Automation and Computing*, vol. 9, no. 3, pp. 306–314, 2012.

[6] P. Bromová, P. Škoda, J. Vážný. Classification of spectra of emission line stars using machine learning techniques. *International Journal of Automation and Computing*, vol. 11, no. 3, pp. 265–273, 2014.

[7] G. Z. Li, H. H. Meng, M. Q. Yang, J. Y. Yang. Combining support vector regression with feature selection for multivariate calibration. *Neural Computing & Applications*, vol. 18, no. 7, pp. 813–820, 2009.

[8] G. Z. Li, J. Yang, C. Z. Ye, D. Y. Geng. Degree prediction of malignancy in brain glioma using support vector machines. *Computers in Biology and Medicine*, vol. 36, no. 3, pp. 315–325, 2006.

[9] M. Y. Park, T. Hastie. Penalized logistic regression for detecting gene interactions. *Biostatistics*, vol. 9, no. 1, pp. 30–50, 2008.

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

[11] J. Zhu, S. Rosset, T. Hastie, R. Tibshirani. 1-norm support vector machines. *Advances in Neural Information Processing Systems 16*, Cambridge, USA: MIT Press, pp. 49–56, 2004.

[12] G. C. Cawley, N. L. C. Talbot. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics*, vol. 22, no. 19, pp. 2438–2355, 2006.

[13] Z. Liu, F. Jiang, G. Tian, S. Wang, F. Sato, S. J. Meltzer, M. Tan. Sparse logistic regression with Lp penalty for biomarker identification. *Statistical Applications in Genetics and Molecular Biology*, vol. 6, Article number 6, 2007.

[14] H. Zou, T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[15] L. Wang, J. Zhu, H. Zou. The Doubly regularized support vector machine. *Statistica Sinica*, vol. 16, pp. 589–615, 2006.

[16] L. Wang, J. Zhu, H. Zou. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, vol. 24, no. 3, pp. 412–419, 2008.

[17] J. T. Li, Y. M. Jia. An improved elastic net for cancer classification and gene selection. *Acta Automatica Sinica*, vol. 36, no. 7, pp. 976–981, 2010.

[18] J. T. Li, Y. M. Jia, Z. H. Zhao. Partly adaptive elastic net and its application to microarray classification. *Neural Computing and Application*, vol. 22, no. 6, pp. 1193–1200, 2013.

[19] D. Y. Meng, Y. M. Jia, J. P. Du, F. S. Yu. Data-driven control for relative degree systems via iterative learning. *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2213–2225, 2011.

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

[21] G. M. Higgins, R. M. Anderson. Experimental pathology of the liver. I. Restoration of the liver of the white rat following partial surgical removal. *Archives of Pathology & Laboratory Medicine*, vol. 12, pp. 186–202, 1931.

[22] C. S. Xu, Y. J. Yang, J. Y. Yang, X. G. Chen, G. P. Wang. Analysis of the role of the integrin signaling pathway in hepatocytes during rat liver regeneration. *Cellular & Molecular Biology Letters*, vol. 17, no. 2, pp. 274–288, 2012.

[23] C. S. Xu, X. G. Chen, C. F. Chang, G. P. Wang, W. B. Wang, L. X. Zhang, Q. S. Zhu, L. Wang, F. C. Zhang. Transcriptome analysis of hepatocytes after partial hepatectomy in rats. *Development Genes and Evolution*, vol. 220, no. 9–10, pp. 263–274, 2010.

[24] R. R. Amon, D. C. DuBois, K. E. Pearson, D. A. Stephan, W. J. Jusko. Gene arrays and temporal patterns of drug response: Corticosteroid effects on rat liver. *Functional & Integrative Genomics*, vol. 3, no. 4, pp. 171–179, 2003.

[25] A. Nikitin, S. Egorov, N. Daraselia, I. Mazo. Pathway studio-the analysis and navigation of molecular networks. *Bioinformatics*, vol. 19, no. 16, pp. 2155–2157, 2003.

[26] L. Mulrane, E. Rexhepaj, V. Smart, J. J. Callanan, D. Orhan, T. Eldem, A. Mally, S. Schroeder, K. Meyer, M. Wendt, D. OShea, W. M. Gallagher. Creation of a digital slide and tissue microarray resource from a multi-institutional predictive toxicology study in the rat: An initial report from the PredTox group. *Experimental and Toxicologic Pathology*, vol. 60, no. 4–5, pp. 235–245, 2008.

[27] W. I. Jeong, Q. Park, S. Radaeva, B. Gao. STAT1 inhibits liver fibrosis in mice by inhibiting stellate cell proliferation and stimulating NK cell cytotoxicity. *Hepatology*, vol. 44, no. 6, pp. 1441–1451, 2007.

[28] R. Sun, O. Park, N. Horiguchi, S. Kulkarni, W. I. Jeong, H. Y. Sun, S. Radaeva, B. Gao. STAT1 contributes to dsRNA inhibition of liver regeneration after partial hepatectomy in mice. *Hepatology*, vol. 44, no. 4, pp. 955–966, 2006.

[29] G. F. Chen, H. H. Wang, S. L. Xie, J. Ma, G. Y. Wang. STAT1 negatively regulates hepatocellular carcinoma cell proliferation. *Oncology Reports*, vol. 29, no. 6, pp. 2303–2310, 2013.

[30] M. Snyder, W. He, J. J. Zhang. The DNA replication factor MCM5 is essential for Stat1-mediated transcriptional activation. In *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 41, pp. 14539–14544, 2005.

[31] A. Kleinridders, H. P. M. M. Lauritzen, S. Ussar, J. H. Christensen, M. A. Mori, P. Bross, C. R. Kahn. Leptin regulation of Hsp60 impacts hypothalamic insulin signaling. *Journal of Clinical Investigation*, vol. 123, no. 11, pp. 4667–4680, 2013.

[32] G. J. Moser, D. C. Wolf, R. Harden, A. M. Standeven, J. Mills, R. L. Jirtle, T. L. Goldsworthy. Cell proliferation and regulation of negative growth factors in mouse liver foci. *Carcinogenesis*, vol. 17, no. 9, pp. 1835–1840, 1996.

[33] S. Waguri, M. Kohmura, S. Kanamori, T. Watanabe, Y. Ohsawa, M. Koike, Y. Tomiyama, M. Wakasugi, E. Kominami, Y. Uchiyama. Different distribution patterns of the two mannose 6-phosphate receptors in rat liver. *Journal of Histochemistry & Cytochemistry*, vol. 49, no. 11, pp. 1397–1405, 2001.

[34] L. Villevalois-Cam, C. Rescan, D. Gilot, F. Ezan, P. Loyer, B. Desbuquois, C. Guguen-Guillouzo, G. Baffet. The hepatocyte is a direct target for transforming-growth factor activation via the insulin-like growth factor II/mannose 6-phosphate receptor. *Journal of Hepatology*, vol. 38, no. 2, pp. 156–163, 2003.

[35] U. Pieper-Fürst, F. Lammert. Low-density lipoprotein receptors in liver: Old acquaintances and a newcomer. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, vol. 1831, no. 7, pp. 1191–1198, 2013.

[36] T. G. Santos, I. R. Silva, B. Costa-Silva, A. P. Lepique, V. R. Martins, M. H. Lopes. Enhanced neural progenitor/stem cells self-renewal via the interaction of stress-inducible protein 1 with the prion protein. *Stem Cells*, vol. 29, no. 7, pp. 1126–1136, 2011.

[37] S. Artavanis-Tsakonas, M. D. Rand, R. J. Lake. Notch signaling: Cell fate control and signal integration in development. *Science*, vol. 284, no. 5415, pp. 770–776, 1999.

[38] L. Wierød, C. M. Rosseland, B. Lindeman, M. P. Oksvold, H. Grøsvik, E. Skarpen, H. S. Huitfeldt. Activation of the p53-p21Cip1 pathway is required for CDK2 activation and S-phase entry in primary rat hepatocytes. *Oncogene*, vol. 27, no. 19, pp. 2763–2771, 2008.

Liu-Yuan Chen received his B. Sc. and M. Sc. degrees in applied mathematics from the Henan Normal University, China in 2003 and 2009, respectively. Currently, he is a Ph. D. candidate of Wuhan University of Technology, China.

His research interests include machine learning and data mining.

E-mail: lkxbcly@126.com (Corresponding author)

ORCID iD: 0000-0001-8160-5447

Jie Yang received her B. Sc. degree in communication engineering from Xidian University, China in 1982 and M. Sc. degree in computer and automation from Wuhan Transportation University, China in 1988. She received Ph. D. degree in electronic engineering from the Shanghai Jiao Tong University, China in 1999. Since 1999, she is a professor and doctoral supervisor in the School of Electronics and Information at Wuhan University of Technology, China.

Her research interests include image processing, information hiding, cryptography and multimedia communication.

E-mail: jieyang509@163.com

Guo-Guo Xu received her B. Sc. degree in biological sciences from Shangqiu Normal University, China in 2011. Currently, she is a master student of School of Life Sciences at Henan Normal University, China.

Her research interests include rat liver regeneration and planaria head regeneration.

E-mail: dzzxgg-88430@163.com

Yun-Qing Liu received his B. Sc. degree in mechanical design manufacturing and automation from College of Air Defence Force of Chinese People′s Liberation Army, China in 2009. Currently, he is a master student of School of Life Sciences at Henan Normal University, China.

His research interests include rat liver regeneration and planaria head regeneration.

E-mail: qingyun891@126.com

Jun-Tao Li received his B. Sc. and M. Sc. degrees in applied mathematics from Henan Normal University, China in 2001 and 2004, respectively. He received Ph. D. degree in control theory and control engineering from Beihang University, China in 2010. Since 2011, he is an associate professor of Henan Normal University, China.

His research interests include machine learning and its applications.

E-mail: juntaolimail@126.com

Cun-Shuan Xu received his B. Sc. degree from Henan Normal University, China in 1982 and M. Sc. degree from Beijing Normal University, China in 1985. He received Ph. D. degree from Bremen University, Germany in 1995. Since 1995, he is a professor of School of Life Sciences at Henan Normal University, China.

His research interests include regenerating biology and medicine.

E-mail: cellkeylab@126.com