Adaptive Polynomial Regression and Its Application to Gene Selection of Rat Liver Regeneration

Juntao Li
Henan Engineering Laboratory for Big Data Statistical Analysis and Optimal Control, School of Mathematics and Information Science, Henan Normal University ,46 Jianshe Road, Xinxiang, 453007, P.R. China

Yimin Cao
Henan Engineering Laboratory for Big Data Statistical Analysis and Optimal Control, School of Mathematics and Information Science, Henan Normal University ,46 Jianshe Road, Xinxiang, 453007, P.R. China

Xiaoyu Wang
Henan Engineering Laboratory for Big Data Statistical Analysis and Optimal Control, School of Mathematics and Information Science, Henan Normal University ,46 Jianshe Road, Xinxiang, 453007, P.R. China

Cunshuan Xu
State Key Laboratory Cultivation Base for Cell Differentiation Regulation, Henan Normal University, 46 Jianshe Road, Xinxiang, 453007, P.R. China
E-mail: juntaolimail@126.com, cao_yimin@126.com, 49207543@qq.com, cellkeylab@126.com
www.henannu.edu.cn

Abstract

To deal with multi-class classification problem for gene expression data, this paper proposed an adaptive polynomial regression by incorporating multi-class adaptive elastic net penalty into polynomial likelihood loss. The adaptive polynomial regression was proved to adaptively select relevant genes in groups in performing multi-class classification. The proposed method was successfully applied to gene expression data for rat liver regeneration and the relevant genes were selected. The pathway relationships among the selected genes were also provided to verify their biological rationality.

Keywords: multi-class classification, polynomial regression, gene selection, rat liver regeneration

1 Introduction

Statistical machine learning methods, e.g., support vector machine, lasso and their extensions,[1-3] have been successfully applied to the microarray classification and gene selection. Note that multiple decision functions are required and each is decided by the coefficients corresponding to the different gene groups. Hence, it is a challenge to select the proper genes for the microarray multi-class classification problem.

To deal with the aforementioned problem, some new statistical learning methods have been proposed. For

example, Ref. 4 proposed L_1-norm multi-class support vector machine (MSVM), Ref. 5 proposed the adaptive sup-norm MSVM. In particular, by using the initial estimator, Ref. 6 proposed an adaptive MSVM which can adaptively select the related genes.

Recently, the polynomial regression have attracted much attention.[7-9] Ref. 7 proposed variational bayesian polynomial probit regression model and provide the R package. Ref. 8 proposed the regularized polynomial regression and applied it to multiple alignments of protein sequences. Ref. 9 proposed the multinomial regression and proved its performance of the grouped gene selection. However, these polynomial regression methods can not adaptively select the related genes.

Motivated by Ref. 6, this paper present a new polynomial regression method by introducing the multi-class adaptive elastic net penalty. The adaptive gene selection performance is analyzed and experiment results on the gene expression data for rat liver regeneration are provided which demonstrate the effectiveness of the proposed method.

2 Problem description

Given the microarray set $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$, where $K\geq 3$ represents the number of classes, $x_i=(x_{i1},x_{i2},\cdots x_{ip})$ is the input vector and $y_i\in\{1,2,\cdots K\}$ is sample label. Similar to Ref 6, Let $Y=(y_1,\cdots,y_n)^T$, $X=(x_1;x_2;\cdots;x_n)=(x_{(1)},x_{(2)},\cdots x_{(p)})$, $x_{(j)}=(x_{1j},\cdots x_{nj})^T$. Suppose that the predictors $x_{(j)},\ j=1,\cdots,p$ are standardized, the response has zero mean value. The objective of this paper is to perform classification and select genes by building the following classifier

$$\phi(x)=\arg\max_{k=1,2,\cdots,K} f_k(x).$$

For the sake of convenience, we adopt the notations in Ref. 9. $f_k(x)=b_k+w_k^T x,(k=1,2,\cdots,K)$ be the k th decision function, $w_k=(w_{k1},\cdots,w_{kp})^T$ be the k th row vector of matrix w, $w_{(j)}=(w_{1j},\cdots,w_{Kj})^T$ be the j th column vector of matrix w. By using the multi elastic net penalty and the polynomial likelihoods loss, Ref. 9 proposed the regularized polynomial regression

$$\min_{(b,w)}\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}y_{ik}\left(b_k+w_k^T x_i\right)-\log\sum_{k=1}^{K}e^{\left(b_k+w_k^T x_i\right)}\right]+\lambda_2\sum_{k=1}^{K}\sum_{j=1}^{p}w_{kj}^2+\lambda_1\sum_{k=1}^{K}\sum_{j=1}^{p}\left|w_{kj}\right|\right\}$$
$$s.t.\quad 1^T b=0,\ 1^T w_j=0,\ (j=1,2,\cdots,\mathrm{p}) \tag{1}$$

Although this method can select genes in groups during performing classification, it can not adaptively identify the important gene in the selected groups. This paper is devoted to solving the problem.

3 Main Result

3.1 Adaptive Polynomial Regression Model

Similar to Ref. 6, let $V=(\mathrm{v_{ik}})_{n*k}=[v_1^T;v_2^T;\cdots;v_n^T;]$ denote the index matrix, where v_i^T represents vector code corresponding to label y_i. By introducing multi-class adaptive elastic net penalty, the adaptive polynomial regression can be represented by

$$\min_{(b,w)}\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}y_{ik}\left(b_k+w_k^T x_i\right)-\log\sum_{k=1}^{K}e^{\left(b_k+w_k^T x_i\right)}\right]+\sum_{k=1}^{K}\sum_{j=1}^{p}t_j\left(\lambda_2 w_{kj}^2+\lambda_1\left|w_{kj}\right|\right)\right\}$$
$$s.t.\quad 1^T b=0,\ 1^T w_j=0,\ (j=1,2,\cdots,\mathrm{p}) \tag{2}$$

where
$$t_j=\left\|\tilde{w}_{(j)}\right\|^{-1},\quad \tilde{w}_{kj}=\frac{\left|\sum_{i=1}^{n}x_{ij}v_{ik}\right|}{\sum_{i=1}^{n}x_{ij}^2}. \tag{3}$$

$\tilde{w}_{kj}$ and t_j are the estimated contributions of the j th gene to the k th classifier and the whole classifier. Compared with the regularized polynomial regression (1), the weight $t_j\ (j=1,2,\cdots,\mathrm{p})$ can be viewed as leverage vector which controls shrinkage of parameters adaptively. Substituting the constraint condition into objective function, the constrained optimization problem (2) can be transformed into unconstrained optimization problem. The altered objective function is shown as follows:

$$\begin{aligned}\bar{L}(\lambda_1,\lambda_2,\mathrm{b},w)=&-\frac{1}{n}\sum_{i=1}^{n}\left\{y_{iK}\left[-\sum_{k=1}^{K-1}b_{k'}+\left(-\sum_{k=1}^{K-1}w_{k'}\right)x_i\right]+\sum_{k=1}^{K}y_{ik}(\mathrm{b}_k+w_k^T\mathrm{x}_i)\right.\\&\left.-\log\left(\sum_{k=1}^{K-1}e^{\mathrm{b}_K+w_k^T\mathrm{x}_i}+\mathrm{e}^{\left[-\sum_{k=1}^{K-1}b_{k'}+\left(-\sum_{k=1}^{K-1}w_{k'}\right)x_i\right]}\right)\right\}\\&+\lambda_2\sum_{k=1}^{K-1}\sum_{j=1}^{p}t_j w_{kj}^2+\lambda_2\sum_{j=1}^{p}t_j\left(-\sum_{k=1}^{K-1}w_{k'j}\right)^2\\&+\lambda_1\sum_{k}^{K-1}\sum_{j=1}^{p}t_j\left|w_{kj}\right|+\lambda_1\sum_{j=1}^{p}t_j\left|-\sum_{k=1}^{K-1}w_{k'j}\right|\end{aligned} \tag{4}$$

3.2 Gene Selection performance

Note that the parameters corresponding to the j th gene for each classifiers is assigned the same weight t_j in

adaptive polynomial regression. Hence, the weighted L_1 norm penalty can shrink the parameters corresponding to the non-significant genes converge to 0 adaptively. Meanwhile, it can also reduce shrinkage error of the parameters corresponding to important genes and emerge sparsity. In anther words, the adaptive polynomial regression can select the relevant genes in groups adaptively by estimating their contribution to K classifiers. This performance of gene selection can by described by the following Theorem.

Theorem 1. *Given the sample set* (X,Y) *and regularized parameters* (λ_1,λ_2). *Let* $\hat{\mathrm{b}},\hat{\mathrm{w}}$ *be the solution of adaptive polynomial regression (2),* $x_{(m)}\in R^n, x_{(l)}\in R^n$ *be the columns of* X *corresponding to* $\hat{w}_{(m)},\hat{w}_{(l)}$. *If predictors* $\hat{w}_{(m)}$ *and* $\hat{w}_{(l)}$ *are standardized, then*

$$\left\|\hat{w}_{(m)}-\hat{w}_{(l)}\right\|_2 \le \frac{2\sqrt{K}}{\sqrt{n}\lambda_2}\sum_{i=1}^{n}\left|t_m^{-1}x_{im}-t_l^{-1}x_{il}\right| \tag{5}$$

holds**.**

Proof. Construct the following coefficient matrices

$$b^*=\hat{b}\quad w_{kj'}^*=\begin{cases}\dfrac{t_m}{t_m+t_l}\hat{w}_{km}+\dfrac{t_l}{t_m+t_l}\hat{w}_{kl}, & j'=m,l\\ \hat{w}_{kj'}, & j'\ne m,l\end{cases}$$

Note that $(\hat{\mathrm{b}},\hat{\mathrm{w}})$ be the solution of adaptive polynomial regression. Hence, we have

$$0\le \bar{L}(\lambda_1,\lambda_2,b^*,\mathrm{w}^*)-\bar{L}(\lambda_1,\lambda_2,\hat{b},\hat{\mathrm{w}}) \tag{6}$$

Following the similar procedure in Ref. 6, we have

$$\left\|\hat{w}_{(m)}-\hat{w}_{(l)}\right\|_2 \le \frac{2\sqrt{K}}{\sqrt{n}\lambda_2}\sum_{i=1}^{n}\left|t_m^{-1}x_{im}-t_l^{-1}x_{il}\right| \tag{7}$$

□

Let $\rho=x_{(m)}^T x_{(l)}=\sum_{i=1}^{n}x_{im}x_{il}$, $\gamma=2t_m t_l/t_m^2+t_l^2$. By some simple calculations, we have

$$\frac{2\sqrt{K}}{n\lambda_2}\left\|t_m^{-1}x_{(m)}-t_l^{-1}x_{(l)}\right\|_1=\frac{2\sqrt{K}}{n\lambda_2}\sqrt{n}\left\|t_m^{-1}x_{(m)}-t_l^{-1}x_{(l)}\right\|_2 \tag{8}$$

$$=\frac{2\sqrt{K}}{n\lambda_2}\sqrt{t_m^{-2}+t_l^{-2}}\cdot\sqrt{1-\gamma\rho}$$

When $K=2$, it can yield that

$$\left|\hat{w}_{(1m)}-\hat{w}_{(1l)}\right|\le\frac{2}{\sqrt{n}\lambda_2}\sqrt{2(1-\rho)}$$

Note that $\gamma=1$ if and only if $t_m=t_l$. Hence, the adaptive polynomial regression will assign the same coefficients to the corresponding genes only if $x_{(m)}$ and $x_{(l)}$ are highly correlated and these genes show the same overall importance to the all K classifiers, i.e. $\|\tilde{w}_{(m)}\|_1=\|\tilde{w}_{(l)}\|_1$. In other words, adaptive polynomial regression can adaptively identify the important gene in the selected gene groups. According the terms in Ref.6, we call it the adaptive grouping gene selection p. Different from Ref.6, the pathwise coordinate descent algorithm (PCD) presented in Ref. 10 can be improved to solve the adaptive polynomial regression model.

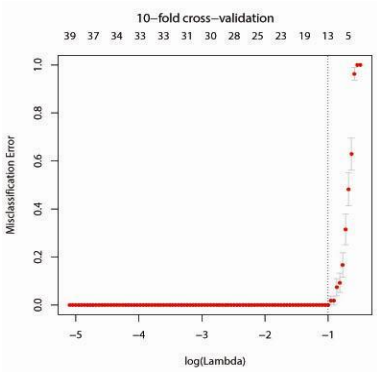
4 Experiment

To demonstrate the performance of the proposed model, we apply the adaptive polynomial regression to the gene chip data of rat liver regeneration. This data set is produced by the State Key Laboratory Cultivation Base for Cell Differentiation Regulation, Henan Normal University, which is available in the NCBI database with accession number: GSE55434.

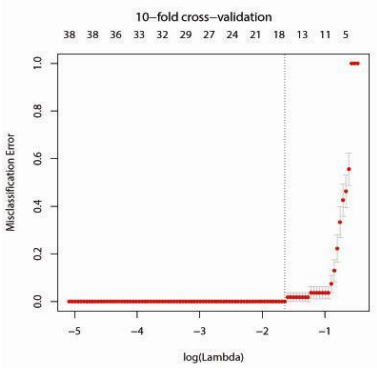
Note that rat liver regeneration process has different physiological activity at every moment. Hence, we treat it as 9-classes classification problem which has 81 sample points, where each class has 9 sample points and each point contains 24618 genes. In our experiments, 54 sample points are used to train the regression model and the rest 27 for testing. In order to guarantee the balance of different classes, we take 6 samples out of each class for training and the rest 3 for testing. Fig.1 (a) shows the tenfold cross-validation error of the adaptive polynomial regression on operation group training set. From Fig.1(a) we can see that when $\ln\lambda\le -1$, i.e. $\lambda\le 0.3678794$ the cross-validation error is the least. Thus, we choose the optimal $\lambda=0.220036243$ and determine the solution of adaptive polynomial regression model. Then we use the obtained coefficients to predict on test set and get 100% classification accuracy. Further, we determine 196 relevant genes using nonzero coefficients.

In order to eliminate the operation error and the influence of other noisy points, we also perform experiments on sham operation group data. Fig.1 (b) shows the tenfold cross-validation error of the adaptive polynomial regression model on sham operation group training set. Similar to the case of operation group data, we choose $\lambda=0.1826835$ and obtain 83 genes. By eliminating the genes existing in sham operation group from 196 genes in operation group, the rest 113 genes are considered to be the relevant genes of rat liver

regeneration. To verify the biological rationality, we conduct pathway relationships among the selected genes by using the pathway studio 8. Fig.2 shows there exists two gene pathways.



(a)



(b)

Fig.1, (a) $\alpha = 0.5$ tenfold cross-validation error of the adaptive polynomial regression on operation group training set. (b) $\alpha = 0.5$ tenfold cross-validation error of the adaptive polynomial regression on shame operation group training set.

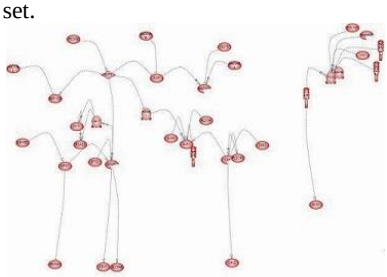


Fig.2, Regulative relation of relevant genes of rat liver Regeneration.

5 Conclusion

By combining polynomial likelihood loss and the multi-class adaptive elastic net penalty, an adaptive polynomial regression was proposed in this paper. The proposed model can adaptively select genes in groups. PCD algorithm can be improved to solve the proposed model. The proposed method was successfully applied to the gene expression data for rat liver regeneration and the relevant genes were selected. Furthermore, the pathway relationships among the selected genes were analyzed by using the pathway studio 8, which verifies their biological rationality.

Acknowledgements

This work is supported by Natural Science Foundation of China (61203293), Program for Science and Technology Innovation Talents in Universities of Henan Province (13HASTIT040).

References

1. I. Guyon, et al., Gene selection for cancer classification using support vector machine, Machine Learning **46**(1) (2002) 389-422.
2. L. Wang, et al., Hybrid huberized support vector machine for microarray classification and gene selection, Bioinformatics **24**(3) (2008) 412-419.
3. R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B (Statistical Methodology) **58**(1) (1996) 267-288.
4. L. Wang and X. Shen, On L1-norm multi-class support vector machine: method and theory, Journal of American Statistical Association **102**(478) (2007) 583-594.
5. H. H. Zhang, et al., Variable selection for the multicategory SVM via adaptive sup-norm regularization, Electronic Journal of Statistics **2**(2008) 149-167.
6. J. Li, et al., Adaptive Multi-class Support Vector Machine for Microarray Classification and Gene Selection, ICROS-SICE International Joint Conference, (Japan, Fukuoka, 2009), pp. 2658-2663.
7. L. Nicola and M. Girolami, vbmp: Variational bayesian multinomial probit regression for multi-class classification in R, Bioinformatics **24**(1) (2008) 135-136.
8. J. Sreekumar et al., Correlated mutations via regularized multinomial regression, BMC Bioinformatics **12**(22) (2011) 444-456.
9. L. Chen, et al., Multinomial regression with elastic net penalty and its grouping effect in gene selection, Abstract and Applied Analysis **7**(1) (2014) 558-577.
10. J. Friedman, et al., Regularization path for generalized linear models via coordinate descent, Journal of Statistical Software **33**(1) (2010) 1-22.