Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2014, Article ID 569501, 7 pages
http://dx.doi.org/10.1155/2014/569501

Research Article

Multinomial Regression with Elastic Net Penalty and Its Grouping Effect in Gene Selection

Liuyuan Chen,[1,2] Jie Yang,[1] Juntao Li,[2] and Xiaoyu Wang[2]

[1] *School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China*
[2] *School of Mathematics and Information Science, Henan Normal University, Xinxiang 453007, China*

Correspondence should be addressed to Liuyuan Chen; lkxbcly@126.com

Received 6 January 2014; Accepted 19 February 2014; Published 31 March 2014

Academic Editor: Yisheng Song

For the multiclass classification problem of microarray data, a new optimization model named multinomial regression with the elastic net penalty was proposed in this paper. By combining the multinomial likeliyhood loss and the multiclass elastic net penalty, the optimization model was constructed, which was proved to encourage a grouping effect in gene selection for multiclass classification.

1. Introduction

Support vector machine [1], lasso [2], and their expansions, such as the hybrid huberized support vector machine [3], the doubly regularized support vector machine [4], the 1-norm support vector machine [5], the sparse logistic regression [6], the elastic net [7], and the improved elastic net [8], have been successfully applied to the binary classification problems of microarray data. However, the aforementioned binary classification methods cannot be applied to the multiclass classification easily. Hence, the multiclass classification problems are the difficult issues in microarray classification [9–11].

Besides improving the accuracy, another challenge for the multiclass classification problem of microarray data is how to select the key genes [9–15]. By solving an optimization formula, a new multicategory support vector machine was proposed in [9]. It can be successfully used to microarray classification [9]. However, this optimization model needs to select genes using the additional methods. To automatically select genes during performing the multiclass classification, new optimization models [12–14], such as the L_1 norm multiclass support vector machine in [12], the multicategory support vector machine with sup norm regularization in [13], and the huberized multiclass support vector machine in [14], were developed.

Note that the logistic loss function not only has good statistical significance but also is second order differentiable. Hence, the regularized logistic regression optimization models have been successfully applied to binary classification problem [15–19]. Multinomial regression can be obtained when applying the logistic regression to the multiclass classification problem. The emergence of the sparse multinomial regression provides a reasonable application to the multiclass classification of microarray data that featured with identifying important genes [20–22]. By using Bayesian L_1 regularization, the sparse multinomial regression model was proposed in [20]. By adopting a data augmentation strategy with Gaussian latent variables, the variational Bayesian multinomial probit model which can reduce the prediction error was presented in [21]. By using the elastic net penalty, the regularized multinomial regression model was developed in [22]. It can be applied to the multiple sequence alignment of protein related to mutation. Although the above sparse multinomial models achieved good prediction results on the real data, all of them failed to select genes (or variables) in groups.

For the multiclass classification of the microarray data, this paper combined the multinomial likelihood loss function having explicit probability meanings [23] with multiclass elastic net penalty selecting genes in groups [14], proposed a multinomial regression with elastic net penalty, and proved that this model can encourage a grouping effect in gene selection at the same time of classification.

2. Problem Formulation and Preliminary

Given a training data set of K-class classification problem $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ represents the input vector of the ith sample and $y_i \in \{1, 2, \ldots, K\}$ represents the class label corresponding to x_i. For the microarray data, n and p represent the number of experiments and the number of genes, respectively. Restricted by the high experiment cost, only a few (less than one hundred) samples can be obtained with thousands of genes in one sample. Let $Y = (y_1, \ldots, y_n)^T$ and $X = (x_{(1)}, x_{(2)}, \ldots, x_{(p)})$, where $x_{(j)} = (x_{1j}, \ldots, x_{nj})^T$, $j = 1, \ldots, p$. Without loss of generality, it is assumed that

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

For the binary classification problem, the class labels are assumed to belong to $L = \{1, -1\}$. The logistic regression model represents the following class-conditional probabilities; that is,

$$\log \frac{\Pr(y_i = +1 \mid x)}{\Pr(y_i = -1 \mid x)} = b + w^T x, \tag{2}$$

and then

$$\Pr(y_i = +1 \mid x) = \frac{1}{1 + e^{-(b + w^T x)}},$$
$$\Pr(y_i = -1 \mid x) = \frac{1}{1 + e^{+(b + w^T x)}} = 1 - \Pr(y_i = +1 \mid x). \tag{3}$$

According to the common linear regression model, Y can be predicted as

$$\widehat{Y} = \widehat{w} X = \sum_{j=1}^{p} \widehat{w}_{(j)} x_{(j)} + \widehat{b}, \tag{4}$$

where $\widehat{b}$ represents bias and $\widehat{w} = (\widehat{w}_{(1)}, \ldots, \widehat{w}_{(p)})^T$ represents the parameter vector.

In this paper, we pay attention to the multiclass classification problems, which imply that $K \geq 3$. Let $f = (f_1, f_2, \ldots, f_K)$ be the decision function, where $f_k(x) = b_k + w_k^T x, (k = 1, 2, \ldots, K)$. The multiclass classifier can be represented as

$$\phi(x) = \arg \max_{k=1,2,\ldots,K} f_k(x). \tag{5}$$

Let $b = (b_1, \cdots, b_K)^T$ and

$$w = \begin{Bmatrix} w_{11} & w_{12} & \cdots & w_{1p} \\ w_{21} & w_{22} & \cdots & w_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ w_{K1} & w_{K2} & \cdots & w_{Kp} \end{Bmatrix}. \tag{6}$$

For convenience, we further let $w_k = (w_{k1}, \ldots, w_{kp})^T$ and $w_{(j)} = (w_{1j}, \ldots, w_{Kj})^T$ represent the kth row vector and jth column vector of the parameter matrix w. Then extending the class-conditional probabilities of the logistic regression model to $(K-1)$-logits, we have the following formula:

$$\log \frac{\Pr(y_i = k \mid x)}{\Pr(y_i = K \mid x)} = b_k + w_k^T x \quad (k = 1, 2, \ldots, K-1), \tag{7}$$

where (b_k, w_k) represent a pair of parameters which corresponds to the sample $(Y = k \mid x)$, and $b_k \in R^1$, $w_k \in R^p$. Similarly, we can construct the Kth as

$$\begin{aligned} \log \frac{\Pr(y_i = K \mid x)}{\Pr(y_i = K \mid x)} &= \log 1 = 0 \\ &= b_K + w_K^T x \end{aligned} \tag{8}$$

holds if and only if $(b_K, w_K) = (0, \vec{0})$. It can be easily obtained that

$$\begin{aligned} 1 &= \Pr(y_i = 1 \mid x) + \Pr(y_i = 2 \mid x) + \cdots + \Pr(y_i = K \mid x) \\ &= \left[e^{b_1 + w_1^T x} + e^{b_2 + w_2^T x} + \cdots + e^{b_K + w_K^T x}\right] \cdot \Pr(y_i = K \mid x), \end{aligned} \tag{9}$$

that is,

$$\Pr(y_i = K \mid x) = \frac{1}{\sum_{k=1}^{K} e^{(b_k + w_k^T x)}}. \tag{10}$$

It should be noted that $(b_K, w_K) = (0, \vec{0})$ if $k = K$. Therefore, the class-conditional probabilities of multiclass classification problem can be represented as

$$\Pr(y_i = k \mid x) = \frac{e^{b_k + w_k^T x}}{\sum_{k=1}^{K} e^{(b_k + w_k^T x)}}. \tag{11}$$

3. Main Results

3.1. Multinomial Regression with the Multiclass Elastic Net Penalty. Following the idea of sparse multinomial regression [20–22], we fit the above class-conditional probability model by the regularized multinomial likelihood. Let $p_k(x_i) = \Pr(y_i = k \mid x_i)$. It is easily obtained that

$$p_{y_i}(x_i) = \frac{e^{(b_{y_i} + w_{y_i}^T x_i)}}{\sum_{k=1}^{K} e^{(b_k + w_k^T x_i)}}. \tag{12}$$

Hence,

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\log p_{y_i}(x_i)\\
&\quad=\frac{1}{n}\left[\log p_{y_1}(x_1)+\log p_{y_2}(x_2)+\cdots+\log p_{y_n}(x_n)\right]\\
&\quad=\frac{1}{n}\left[\log\frac{e^{(b_{y_1}+w_{y_1}^{T}x_1)}}{\sum_{k=1}^{K}e^{(b_k+w_k^{T}x_1)}}+\log\frac{e^{(b_{y_2}+w_{y_2}^{T}x_2)}}{\sum_{k=1}^{K}e^{(b_k+w_k^{T}x_2)}}\right.\\
&\qquad\left.+\cdots+\log\frac{e^{(b_{y_n}+w_{y_n}^{T}x_n)}}{\sum_{k=1}^{K}e^{(b_k+w_k^{T}x_n)}}\right]\\
&\quad=\frac{1}{n}\left[\left(b_{y_1}+w_{y_1}^{T}x_1\right)+\cdots+\left(b_{y_n}+w_{y_n}^{T}x_n\right)\right.\\
&\qquad\left.-\log\sum_{k=1}^{K}e^{(b_k+w_k^{T}x_1)}-\cdots-\log\sum_{k=1}^{K}e^{(b_k+w_k^{T}x_n)}\right]\\
&\quad=\frac{1}{n}\left[\left(b_{y_1}+w_{y_1}^{T}x_1\right)+\cdots+\left(b_{y_n}+w_{y_n}^{T}x_n\right)\right]\\
&\qquad-\frac{1}{n}\sum_{i=1}^{n}\log\sum_{k=1}^{K}e^{(b_k+w_k^{T}x_i)}.
\end{aligned}
\tag{13}
$$

Let

$$
y_{ik}=I\left(y_i=k\right)=\begin{cases}1, & y_i=k,\\ 0, & y_i\neq k.\end{cases}
\tag{14}
$$

Then (13) can be rewritten as

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}y_{ik}\left(b_k+w_k^{T}x_i\right)-\frac{1}{n}\sum_{i=1}^{n}\log\sum_{k=1}^{K}e^{(b_k+w_k^{T}x_i)}\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}y_{ik}\left(b_k+w_k^{T}x_i\right)-\log\sum_{k=1}^{K}e^{(b_k+w_k^{T}x_i)}\right].
\end{aligned}
\tag{15}
$$

Note that

$$
\begin{aligned}
\log p_{y_i}(x_i)&=\log\frac{e^{(b_{y_i}+w_{y_i}^{T}x_i)}}{\sum_{k=1}^{K}e^{(b_k+w_k^{T}x_i)}}<\log 1=0,\\
&-\frac{1}{n}\sum_{i=1}^{n}\log p_{y_i}(x_i)>0.
\end{aligned}
\tag{16}
$$

Hence, the multinomial likelihood loss function can be defined as

$$
\begin{aligned}
l\left(\{b_k,w_k\}_1^{K}\right)=-\frac{1}{n}\sum_{i=1}^{n}&\left[\sum_{k=1}^{K}y_{ik}\left(b_k+w_k^{T}x_i\right)\right.\\
&\left.-\log\sum_{k=1}^{K}e^{(b_k+w_k^{T}x_i)}\right].
\end{aligned}
\tag{17}
$$

In order to improve the performance of gene selection, the following elastic net penalty for the multiclass classification problem was proposed in [14]

$$
J(b,w)=\lambda_2\sum_{k=1}^{K}\sum_{j=1}^{p}w_{kj}^{2}+\lambda_1\sum_{k=1}^{K}\sum_{j=1}^{p}\left|w_{kj}\right|.
\tag{18}
$$

By combing the multiclass elastic net penalty (18) with the multinomial likelihood loss function (17), we propose the following multinomial regression model with the elastic net penalty:

$$
\begin{aligned}
\underset{(b,w)}{\arg\min}&\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}y_{ik}\left(b_k+w_k^{T}x_i\right)-\log\sum_{k=1}^{K}e^{(b_k+w_k^{T}x_i)}\right]\right.\\
&\left.+\lambda_2\sum_{k=1}^{K}\sum_{j=1}^{p}w_{kj}^{2}+\lambda_1\sum_{k=1}^{K}\sum_{j=1}^{p}\left|w_{kj}\right|\right\},
\end{aligned}
\tag{19}
$$

where λ_1,λ_2 represent the regularization parameter. Note that $(b_K,w_K)=(0,\vec{0})$. Hence, the optimization problem (19) can be simplified as

$$
\begin{aligned}
\underset{(b,w)}{\arg\min}&\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K-1}y_{ki}\left(b_k+w_k^{T}x_i\right)\right.\right.\\
&\qquad\left.-\log\left(1+\sum_{k=1}^{K-1}e^{(b_k+w_k^{T}x_i)}\right)\right]\\
&\left.+\lambda_2\left(\sum_{k=1}^{K-1}\sum_{j=1}^{p}w_{kj}^{2}\right)+\lambda_1\left(\sum_{k=1}^{K-1}\sum_{j=1}^{p}\left|w_{kj}\right|\right)\right\}.
\end{aligned}
\tag{20}
$$

3.2. Grouping Effect. For the microarray classification, it is very important to identify the related gene in groups. In the section, we will prove that the multinomial regression with elastic net penalty can encourage a grouping effect in gene selection. To this end, we must first prove the inequality shown in Theorem 1.

Theorem 1. *Let* $(\widehat{b},\widehat{w})$ *be the solution of the optimization problem* (19) *or* (20). *For any new parameter pairs which are selected as* $(b^*,w^*)=([b_-^*;\vec{0}],[w_k;\vec{0}])$, *the following inequality*

$$
\begin{aligned}
&\left|\log\left(1+\sum_{k=1}^{K-1}e^{(b_k^*+w_k^{*T}x_i)}\right)-\log\left(1+\sum_{k=1}^{K-1}e^{(\widehat{b}_k+\widehat{w}_k^{T}x_i)}\right)\right|\\
&\quad\leq\sum_{k=1}^{K-1}\left|\left(b_k^*+w_k^{*T}x_i\right)-\left(\widehat{b}_k+\widehat{w}_k^{T}x_i\right)\right|
\end{aligned}
\tag{21}
$$

holds, where b_-^* *and* $\widehat{b}_-$ *represent the first* $K-1$ *rows of vectors* b^* *and* $\widehat{b}$ *and* w_-^* *and* $\widehat{w}_-$ *represent the first* $K-1$ *rows of matrices* w^* *and* $\widehat{w}$.

Proof. Note that the inequality $n + |m - n| \geq m$ holds for the arbitrary real numbers m and n. Hence, the following inequality

$$\left(\widehat{b}_t + \widehat{w}_t^T x_i\right) + \left|\left(b_t^* + w_t^{*T} x_i\right) - \left(\widehat{b}_t + \widehat{w}_t^T x_i\right)\right| \geq \left(b_t^* + w_t^{*T} x_i\right) \tag{22}$$

holds for any pairs $(b_t^*, w_t^*), (\widehat{b}_t, \widehat{w}_t)$. From (22), it can be easily obtained that

$$\begin{aligned}\left(\widehat{b}_t + \widehat{w}_t^T x_i\right) + \sum_{k=1}^{K-1} \left|\left(b_k^* + w_k^{*T} x_i\right) - \left(\widehat{b}_k + \widehat{w}_k^T x_i\right)\right| \\ \geq \left(b_t^* + w_t^{*T} x_i\right).\end{aligned} \tag{23}$$

that is,

$$e^{(\widehat{b}_t+\widehat{w}_t^T x_i)+\sum_{k=1}^{K-1}|(b_k^*+w_k^{*T}x_i)-(\widehat{b}_k+\widehat{w}_k^T x_i)|} \geq e^{(b_t^*+w_t^{*T}x_i)}. \tag{24}$$

Note that

$$e^{\sum_{k=1}^{K-1}|(b_k^*+w_k^{*T}x_i)-(\widehat{b}_k+\widehat{w}_k^T x_i)|} \geq e^0 = 1. \tag{25}$$

Hence, from (24) and (25), we can get

$$\frac{\Delta_1}{\Delta_2} \leq 1, \tag{26}$$

where

$$\begin{aligned}\Delta_1 &= 1 + e^{(b_1^*+w_1^{*T}x_i)} + \cdots + e^{(b_t^*+w_t^{*T}x_i)} \\ &\quad + \cdots + e^{(b_{K-1}^*+w_{K-1}^{*T}x_i)}, \\ \Delta_2 &= e^{\sum_{k=1}^{K-1}|(b_k^*+w_k^{*T}x_i)-(\widehat{b}_k+\widehat{w}_k^T x_i)|} \\ &\quad + e^{(\widehat{b}_1+\widehat{w}_1^T x_i)+\sum_{k=1}^{K-1}|(b_k^*+w_k^{*T}x_i)-(\widehat{b}_k+\widehat{w}_k^T x_i)|} \\ &\quad + \cdots + e^{(\widehat{b}_t+\widehat{w}_t^T x_i)+\sum_{k=1}^{K-1}|(b_k^*+w_k^{*T}x_i)-(\widehat{b}_k+\widehat{w}_k^T x_i)|} \\ &\quad + \cdots + e^{(\widehat{b}_{K-1}+\widehat{w}_{K-1}^{*T}x_i)+\sum_{k=1}^{K-1}|(b_k^*+w_k^{*T}x_i)-(\widehat{b}_k+\widehat{w}_k^T x_i)|}.\end{aligned} \tag{27}$$

Equation (26) is equivalent to the following inequality:

$$\begin{aligned}&\left(1 + e^{(b_1^*+w_1^{*T}x_i)} + \cdots + e^{(b_{K-1}^*+w_{K-1}^{*T}x_i)}\right) \\ &\times\left(\left[1 + e^{(\widehat{b}_1+\widehat{w}_1^T x_i)} + \cdots + e^{(\widehat{b}_{K-1}+\widehat{w}_{K-1}^T x_i)}\right]\right. \\ &\quad\left.\cdot e^{\sum_{k=1}^{K-1}|(b_k^*+w_k^{*T}x_i)-(\widehat{b}_k+\widehat{w}_k^T x_i)|}\right)^{-1} \leq 1 \\ &\Longleftrightarrow \log\frac{1 + e^{(b_1^*+w_1^{*T}x_i)} + \cdots + e^{(b_{K-1}^*+w_{K-1}^{*T}x_i)}}{1 + e^{(\widehat{b}_1+\widehat{w}_1^T x_i)} + \cdots + e^{(\widehat{b}_{K-1}+\widehat{w}_{K-1}^T x_i)}} \\ &\quad \leq \sum_{k=1}^{K-1}\left|\left(b_k^* + w_k^{*T}x_i\right) - \left(\widehat{b}_k + \widehat{w}_k^T x_i\right)\right| \\ &\Longleftrightarrow \left|\log\left(1 + \sum_{k=1}^{K-1} e^{(b_k^*+w_k^{*T}x_i)}\right) - \log\left(1 + \sum_{k=1}^{K-1} e^{(\widehat{b}_k+\widehat{w}_k^T x_i)}\right)\right| \\ &\quad \leq \sum_{k=1}^{K-1}\left|\left(b_k^* + w_k^{*T}x_i\right) - \left(\widehat{b}_k + \widehat{w}_k^T x_i\right)\right|.\end{aligned} \tag{28}$$

Hence, inequality (21) holds. This completes the proof. □

Using the results in Theorem 1, we prove that the multinomial regression with elastic net penalty (19) can encourage a grouping effect.

Theorem 2. *Give the training data set* $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ *and assume that the matrix* X *and vector* Y *satisfy* (1). *If the pairs* $(\widehat{b}, \widehat{w})$ *are the optimal solution of the multinomial regression with elastic net penalty* (19), *then the following inequality*

$$\left\|\widehat{w}_{(m)} - \widehat{w}_{(l)}\right\|_2 \leq \frac{2\sqrt{K}}{\sqrt{n}\lambda_2}\sqrt{2(1-\rho)}. \tag{29}$$

holds, where $\rho = x_{(m)}^T x_{(l)} = \sum_{i=1}^n x_{im}x_{il}$, $\widehat{w}_{(m)}$ *is the* m*th column of parameter matrix* $\widehat{w}$, *and* $\widehat{w}_{(m)}$ *is the* l*th column of parameter matrix* $\widehat{w}$.

Proof. First of all, we construct the new parameter pairs $(b^*, w^*) = ([b_-^*; \vec{0}], [w_k^*; \vec{0}])$, where

$$b_-^* = \widehat{b}_-, \qquad w_{kj'}^* = \begin{cases}\dfrac{1}{2}\left(\widehat{w}_{km} + \widehat{w}_{kl}\right), & j' = m, l, \\ \widehat{w}_{kj'}, & j' \neq m, l.\end{cases} \tag{30}$$

Let

$$\begin{aligned}\overline{L} = -\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K-1} y_{ki}\left(b_k + w_k^T x_i\right)\right. \\ \left. - \log\left(1 + \sum_{k=1}^{K-1} e^{(b_k+w_k^T x_i)}\right)\right] \\ + \lambda_2\sum_{k=1}^{K-1}\sum_{j=1}^{p} w_{kj}^2 + \lambda_1\sum_{k=1}^{K-1}\sum_{j=1}^{p}\left|w_{kj}\right|.\end{aligned} \tag{31}$$

Since the pairs $(\widehat{b}, \widehat{w})$ are the optimal solution of the multinomial regression with elastic net penalty (19), it can be easily obtained that

$$0 \leq \overline{L}\left(\lambda_1, \lambda_2, b_-^*, w_-^*\right) - \overline{L}\left(\lambda_1, \lambda_2, \widehat{b}_-, \widehat{w}_-\right). \tag{32}$$

Note that the function $L_{ki}(\widehat{b}_k, \widehat{w}_k, x_i) = b_k + w_k^T x_i$ is Lipschitz continuous. Hence, we have

$$\left|L_{ki}\left(b_k^*, w_k^*, x_i\right) - L_{ki}\left(\widehat{b}_k, \widehat{w}_k, x_i\right)\right| \le \left|\left(w_k^* - \widehat{w}_k\right)^T x_i\right|. \quad (33)$$

From (33) and (21) and the definition of the parameter pairs (b^*, w^*), we have

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K-1} L_{ki}\left(b_k^*, w_k^*, x_i\right) - L_{ki}\left(\widehat{b}_k, \widehat{w}_k, x_i\right)\right.\\
&\qquad - \log\left(1 + \sum_{k=1}^{K-1} e^{\left(b_k^* + w_k^{*T} x_i\right)}\right)\\
&\qquad \left. + \log\left(1 + \sum_{k=1}^{K-1} e^{\left(\widehat{b}_k + \widehat{w}_k^T x_i\right)}\right)\right]\\
&\le \left|-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K-1} L_{ki}\left(b_k^*, w_k^*, x_i\right) - L_{ki}\left(\widehat{b}_k, \widehat{w}_k, x_i\right)\right|\\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\left|\log\left(1 + \sum_{k=1}^{K-1} e^{\left(b_k^* + w_k^{*T} x_i\right)}\right)\right.\\
&\qquad \left. - \log\left(1 + \sum_{k=1}^{K-1} e^{\left(\widehat{b}_k + \widehat{w}_k^T x_i\right)}\right)\right|\\
&\le \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left|L_{ki}\left(b_k^*, w_k^*, x_i\right) - L_{ki}\left(\widehat{b}_k, \widehat{w}_k, x_i\right)\right|\\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left|\left(b_k^* + w_k^{*T} x_i\right) - \left(\widehat{b}_k + \widehat{w}_k^T x_i\right)\right|\\
&\le \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left|\left(w_k^* - \widehat{w}_k\right)^T x_i\right|\\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left|\left(w_k^* - \widehat{w}_k\right)^T x_i\right|\\
&= \frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left|\left(\widehat{w}_{km} - \widehat{w}_{kl}\right)\left(x_{im} - x_{il}\right)\right|\\
&\quad + \frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left|\left(\widehat{w}_{km} - \widehat{w}_{kl}\right)\left(x_{im} - x_{il}\right)\right|\\
&= \frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left|\left(\widehat{w}_{km} - \widehat{w}_{kl}\right)\left(x_{im} - x_{il}\right)\right|\\
&\quad + \frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left|\left(\widehat{w}_{km} - \widehat{w}_{kl}\right)\left(x_{im} - x_{il}\right)\right|\\
&= \frac{1}{2n}\sum_{i=1}^{n}\left|\left(x_{im} - x_{il}\right)\right| \cdot \sum_{k=1}^{K-1}\left|\left(\widehat{w}_{km} - \widehat{w}_{kl}\right)\right|\\
&\quad + \frac{1}{2n}\sum_{i=1}^{n}\sum_{k=1}^{K-1}\left|\left(\widehat{w}_{km} - \widehat{w}_{kl}\right)\left(x_{im} - x_{il}\right)\right|\\
&= \frac{1}{n}\left\|x_{(m)} - x_{(l)}\right\|_1 \cdot \left\|\widehat{w}_{(m)} - \widehat{w}_{(l)}\right\|_1.
\end{aligned} \quad (34)$$

Analogically, we have

$$\begin{aligned}
&\sum_{k=1}^{K-1}\sum_{j=1}^{p}\left(\left|w_{kj}^*\right| - \left|\widehat{w}_{kj}\right|\right)\\
&\quad = \sum_{k=1}^{K-1}\left(2\left|\frac{\widehat{w}_{km} + \widehat{w}_{kl}}{2}\right| - \left|\widehat{w}_{km}\right| - \left|\widehat{w}_{kl}\right|\right) \le 0,\\
&\sum_{k=1}^{K-1}\sum_{j=1}^{p}\left(w_{kj}^*\right)^2 - \left(\widehat{w}_{kj}^2\right)\\
&\quad = -\frac{1}{2}\sum_{k=1}^{K-1}\left(\widehat{w}_{km} - \widehat{w}_{kl}\right)^2 = -\frac{1}{2}\left\|\widehat{w}_{(m)} - \widehat{w}_{(l)}\right\|_2^2.
\end{aligned} \quad (35)$$

Substituting (34) and (35) into (32) gives

$$0 \le \frac{1}{n}\left\|x_{(m)} - x_{(l)}\right\|_1 \cdot \left\|\widehat{w}_{(m)} - \widehat{w}_{(l)}\right\|_1 - \frac{\lambda_2}{2}\left\|\widehat{w}_{(m)} - \widehat{w}_{(l)}\right\|_2^2. \quad (36)$$

that is,

$$\begin{aligned}
0 &\le \frac{2}{n\lambda_2}\left\|x_{(m)} - x_{(l)}\right\|_1 \cdot \left\|\widehat{w}_{(m)} - \widehat{w}_{(l)}\right\|_1 - \left\|\widehat{w}_{(m)} - \widehat{w}_{(l)}\right\|_2^2\\
&\le \frac{2\sqrt{K}}{n\lambda_2}\left\|x_{(m)} - x_{(l)}\right\|_1 \cdot \left\|\widehat{w}_{(m)} - \widehat{w}_{(l)}\right\|_2 - \left\|\widehat{w}_{(m)} - \widehat{w}_{(l)}\right\|_2^2.
\end{aligned} \quad (37)$$

From (37), it can be easily obtained that

$$\begin{aligned}
&\left\|\widehat{w}_{(m)} - \widehat{w}_{(l)}\right\|_2^2\\
&\quad \le \frac{2\sqrt{K}}{n\lambda_2}\left\|x_{(m)} - x_{(l)}\right\|_1 \le \frac{2\sqrt{K}}{n\lambda_2} \cdot \sqrt{n}\left\|x_{(m)} - x_{(l)}\right\|_2\\
&\quad = \frac{2\sqrt{K}}{\sqrt{n}\lambda_2}\sqrt{2 - 2x_{(m)}^T x_{(l)}} = \frac{2\sqrt{K}}{\sqrt{n}\lambda_2}\sqrt{2\left(1 - \rho\right)},
\end{aligned} \quad (38)$$

where $\rho = x_{(m)}^T x_{(l)} = \sum_{i=1}^{n} x_{im} x_{il}$. This completes the proof. □

According to the inequality shown in Theorem 2, the multinomial regression with elastic net penalty can assign the same parameter vectors (i.e., $\widehat{w}_{(m)} = \widehat{w}_{(l)}$) to the high correlated predictors $x_{(m)}, x_{(l)}$ (i.e., $\rho = 1$). This means that the multinomial regression with elastic net penalty can select genes in groups according to their correlation. According to the technical term in [14], this performance is called grouping effect in gene selection for multiclass classification. Particularly, for the binary classification, that is, $K = 2$, inequality (29) becomes

$$\left|\widehat{w}_{1m} - \widehat{w}_{1l}\right| \le \frac{2}{\sqrt{n}\lambda_2}\sqrt{2\left(1-\rho\right)}. \tag{39}$$

This corresponds with the results in [7].

3.3. Solving Algorithm. Microarray is the typical small n, large p problem. Because the number of the genes in microarray data is very large, it will result in the curse of dimensionality to solve the proposed multinomial regression. To improve the solving speed, Friedman et al. proposed the pairwise coordinate decent algorithm which takes advantage of the sparse property of characteristic. Therefore, we choose the pairwise coordinate decent algorithm to solve the multinomial regression with elastic net penalty. To this end, we convert (19) into the following form:

$$\begin{aligned} \arg\min_{(b,w)} \Biggl\{ -\frac{1}{n}\sum_{i=1}^{n}\Biggl[\sum_{k=1}^{K} y_{ik}\left(b_k + w_k^T x_i\right) &\\ - \log\sum_{k=1}^{K} e^{(b_k + w_k^T x_i)}\Biggr] & \qquad \qquad (40)\\ + \lambda\left[\alpha\sum_{k=1}^{K}\sum_{j=1}^{p} w_{kj}^2 + (1-\alpha)\sum_{k=1}^{K}\sum_{j=1}^{p}\left|w_{kj}\right|\right]\Biggr\}. & \end{aligned}$$

Equation (40) can be easily solved by using the R package "glmnet" which is publicly available.

4. Conclusion

By combining the multinomial likelihood loss function having explicit probability meanings with the multiclass elastic net penalty selecting genes in groups, the multinomial regression with elastic net penalty for the multiclass classification problem of microarray data was proposed in this paper. The proposed multinomial regression is proved to encourage a grouping effect in gene selection. In the next work, we will apply this optimization model to the real microarray data and verify the specific biological significance.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by Natural Science Foundation of China (61203293, 61374079), Key Scientific and Technological Project of Henan Province (122102210131, 122102210132), Program for Science and Technology Innovation Talents in Universities of Henan Province (13HASTIT040), Foundation and Advanced Technology Research Program of Henan Province (132300410389, 132300410390, 122300410414, and 132300410432), Foundation of Henan Educational Committee (13A120524), and Henan Higher School Funding Scheme for Young Teachers (2012GGJS-063).

References

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

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

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

[4] L. Wang, J. Zhu, and H. Zou, "The doubly regularized support vector machine," *Statistica Sinica*, vol. 16, no. 2, pp. 589–615, 2006.

[5] J. Zhu, R. Rosset, and T. Hastie, "1-norm support vector machine," in *Advances in Neural Information Processing Systems*, vol. 16, pp. 49–56, MIT Press, New York, NY, USA, 2004.

[6] G. C. Cawley and N. L. C. Talbot, "Gene selection in cancer classification using sparse logistic regression with Bayesian regularization," *Bioinformatics*, vol. 22, no. 19, pp. 2348–2355, 2006.

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

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

[9] Y. Lee, Y. Lin, and G. Wahba, "Multicategory support vector machines: theory and application to the classification of microarray data and satellite radiance data," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 67–81, 2004.

[10] X. Zhou and D. P. Tuck, "MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data," *Bioinformatics*, vol. 23, no. 9, pp. 1106–1114, 2007.

[11] S. Student and K. Fujarewicz, "Stable feature selection and classification algorithms for multiclass microarray data," *Biology Direct*, vol. 7, no. 33, pp. 133–140, 2012.

[12] L. Wang and X. Shen, "On L_1-norm multiclass support vector machines: methodology and theory," *Journal of the American Statistical Association*, vol. 102, no. 478, pp. 583–594, 2007.

[13] H. H. Zhang, Y. Liu, Y. Wu, and J. Zhu, "Variable selection for the multicategory SVM via adaptive sup-norm regularization," *Electronic Journal of Statistics*, vol. 2, pp. 149–167, 2008.

[14] J.-T. Li and Y.-M. Jia, "Huberized multiclass support vector machine for microarray classification," *Acta Automatica Sinica*, vol. 36, no. 3, pp. 399–405, 2010.

[15] M. You and 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.

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

[17] K. Koh, S.-J. Kim, and S. Boyd, "An interior-point method for large-scale L_1-regularized logistic regression," *Journal of Machine Learning Research*, vol. 8, pp. 1519–1555, 2007.

[18] C. Xu, Z. M. Peng, and W. F. Jing, "Sparse kernel logistic regression based on $L_{1/2}$ regularization," *Science China Information Sciences*, vol. 56, no. 4, pp. 1–16, 2013.

[19] Y. Yang, N. Kenneth, and S. Kim, "A novel k-mer mixture logistic regression for methylation susceptibility modeling of CpG dinucleotides in human gene promoters," *BMC Bioinformatics*, vol. 13, supplement 3, article S15, pp. 1471–1480, 2012.

[20] G. C. Cawley, N. L. C. Talbot, and M. Girolami, "Sparse multinomial logistic regression via Bayesian L1 regularization," in *Advances in Neural Information Processing Systems*, vol. 19, pp. 209–216, MIT Press, New York, NY, USA, 2007.

[21] N. Lama and M. Girolami, "vbmp: variational Bayesian multinomial probit regression for multi-class classification in R," *Bioinformatics*, vol. 24, no. 1, pp. 135–136, 2008.

[22] J. Sreekumar, C. J. F. ter Braak, R. C. H. J. van Ham, and A. D. J. van Dijk, "Correlated mutations via regularized multinomial regression," *BMC Bioinformatics*, vol. 12, article 444, 2011.

[23] 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.