doi:10.3772/j.issn.1002-0470.2020.04.004

弱分层交互 Lasso 罚 logistic 回归模型和改进坐标下降算法①

李 静②* 于 辉* 王金甲③**

(*燕山大学理学院 秦皇岛 066004)

(**燕山大学信息科学与工程学院,河北省信息传输与信号处理重点实验室 秦皇岛 066004)

摘 要 基于变量交互和分层思想,提出了一种弱分层交互 Lasso 罚 logistic 回归模型。首先给出了交互模型定义和弱分层约束条件,然后给出了凸松弛条件和基于坐标下降法的系数求解算法。在4个 UCI 机器学习数据集和1个日常生活活动识别数据集上进行实验,实验结果证明了变量交互对分类也有贡献,分层对分类也有贡献。分层交互 Lasso 兼具 Lasso 和交互 Lasso 的优点。

关键词 变量交互; 分层; Lasso; logistic 回归; 坐标下降算法

0 引 言

目前高维数据回归问题的稀疏性使线性方法如 Lasso 获得了巨大的成功[1]。Lasso 是 L1 罚的最小二乘回归,也可以推广到广义线性模型[2],例如 L1 罚的 logistic 回归用于分类[3]。响应变量是预测变量的线性加权和,加权系数可通过坐标下降法求解[4]。在分析高维数据时响应变量可能不能用预测变量的线性加权和来解释,那么就需要使用2次模型和高次模型。这有可能说明存在特征交互问题[5]。例如单核苷酸多态性(SNPs)间的交互被认为在癌症和其他疾病诊断中起着重要的作用[6]。线性模型可解释性好、计算简单的优点使得考虑特征交互的模型成为研究热点和难点[7]。特征交互的分层模型的方法可以分为3类:第1类是多步骤方法。一旦交互特征对应的预测变量在模型中,那么交互特征也必须在模型中[8]。或者先考虑变量选择后考虑交互[9],采用修正的最小角回归算法求解分层模型[10]。第2类是贝叶斯方法,例如改进随机搜索变量选择方法用于分层模型[11]。第3类是基于优化的方法,将稀疏交互分层模型用公式表示为非凸优化问题[12],进一步将非凸优化表达为凸优化问题如 Lasso[13] 和 group Lasso 问题[14]。在结构稀疏文献中[15],复合绝对处罚(composite absolute penalties ,CAP)也能获得分组和交互稀疏,但是交互特征系数被罚了2次[16]。文献[17]的方法解决了在非线性交互问题上的分层稀疏性。也有文献研究特征交互但不分层的方法,如考虑二值变量高阶交互的 logistic 回归方法[18],从高维数据中选择交互特征的研究[19],从高维数据中的多元数据图表示的交互特征中采用遗传算法优选特征[20]。

Bien 等人[13]提出了将交互分层 Lasso 方法用于回归,基于卡罗需库恩塔克(Karush-Kuhn-Tucker, KKT)条件和拉格朗日乘子法给出了系数解,并给出了参数稀疏和实际稀疏的概念。但是没有采用坐标下降法。本论文在已有交互特征研究基础上,提出了几何代数变量交互的概念,给出了弱分层交互 Lasso 方法用于 logistic 回归模型,并通过改进的坐标下降法求解加权系数。本论文创新在于:第1,用

① 国家自然科学基金(61473339, 61771420, 61501397, 81803958),京津冀基础研究合作专项(19JCZDJC65600, F2019203583),燕山大学青年教师自主研究计划课题(15LGA015)和燕山大学博士基金资助项目。

② 女,1977 年生,博士;研究方向:统计信号处理;E-mail: 01016888@sina.com

③ 通信作者,E-mail: wjj@ysu.edu.cn

(收稿日期:2019-06-06)

几何代数解释了变量交互;第 2,将文献[13]的弱分层交互 Lasso 推广到 logistic 回归;第 3,推导了改进的坐标下降算法用于模型求解。实验数据采用了 UCI 机器学习数据库的 4 种数据集和 1 个真实的日常生活活动识别数据集,并与多种方法进行了比较,实验结果表明分层交互 Lasso 的分类性能优于 Lasso、交互 Lasso 和传统方法。分层交互 Lasso 方法在处理数据时兼具 Lasso 和交互 Lasso 的优点。

1　几何代数的变量交互理论

定义 1　如果函数 $f(x, y)$ 不能被表示为独立的 $f_1(x)+f_2(y)$,则 x, y 在函数 f 上存在交互。

定义 2　通俗的解释如下,如果响应变量不能用预测变量的线性加权和来表示,这很可能说明变量间存在交互。

变量间交互可以很容易用几何代数理论加以解释。图 1 是几何代数各级子空间示意图,其中 1 阶向量可以表示原始数据的 p 维坐标子空间,即原始数据的 p 维变量是在 1 阶向量上的投影值。2 阶向量表示 2 种变量间的交互,最简单的 2 阶向量系数可以是 2 个 1 阶向量间的交互变量,文献[13]的变量交互采用的是面积特征,文献[20]的变量交互采用的是重心特征。依次类推,k 阶向量表示高阶交互。本文只研究了 1 阶主变量与 2 阶面积交互变量。本文方法也可以推广非线性复杂函数形式或高阶交互。

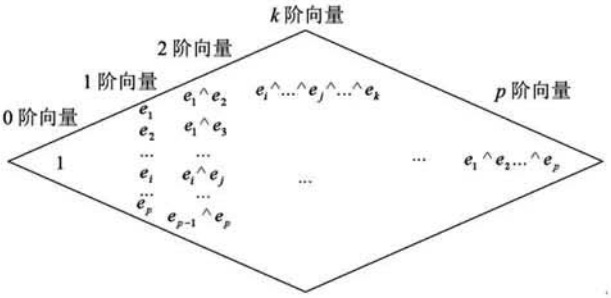

图 1　几何代数各级子空间示意图

2　考虑交互和分层的二值 logistic 回归模型

假设模型样本输出为 Y, 输入 X 为 p 维变量 $(x_1,\cdots x_j,\cdots, x_p)$,样本在每维输入变量间都存在几何代数面积交互项 x_jx_k。则含有 2 次交互项的 logistic 回归模型具有以下形式:

$$\operatorname{logit}(P(Y=1\mid X))=\sum_{j=0}^{p}\beta_jx_j+\frac{1}{2}\sum_{j\neq k}\Theta_{jk}x_jx_k+\varepsilon \tag{1}$$

其中,x_0 为 1,x_j 表示 1 阶主变量,x_jx_k 表示几何代数 2 阶面积交互变量,1 阶主变量系数 $\beta\in R^{p+1}$,2 阶面积交互变量系数 $\Theta\in R^{p\times p}$ 是对称变量系数矩阵,$\Theta_{jj}=0$,$\varepsilon\sim N(0,\sigma^2)$。

假设训练样本为 $(\boldsymbol{x}_1, y_1),\cdots,(\boldsymbol{x}_i, y_i),\cdots,(\boldsymbol{x}_N, y_N)$,本文目的是从 p 个 1 阶主变量和 $p(p-1)/2$ 个 2 阶交互变量中通过 Lasso 方法进行子集选择,估计出模型系数中非零参数值。定义 $y_i=\begin{cases}1 & y_i=1\\ 0 & y_i\neq 1\end{cases}$,并定义这 2 类的概率分别如下:

$$\mathrm{P}(y_i=1\mid \boldsymbol{x}_i)=\frac{1}{1+\exp\left(-\sum_j\beta_jx_{ij}-\frac{1}{2}\sum_{j\neq k}\Theta_{jk}x_{ij}x_{ik}\right)}=p(\boldsymbol{x}_i) \tag{2}$$

$$\mathrm{P}(y_i=0\mid \boldsymbol{x}_i)=\frac{\exp\left(-\sum_j\beta_jx_{ij}-\frac{1}{2}\sum_{j\neq k}\Theta_{jk}x_{ij}x_{ik}\right)}{1+\exp\left(-\sum_j\beta_jx_{ij}-\frac{1}{2}\sum_{j\neq k}\Theta_{jk}x_{ij}x_{ik}\right)}=1-p(\boldsymbol{x}_i) \tag{3}$$

通过极大似然估计拟合模型式(1),使 N 次独立观测的似然函数 $L(\beta,\Theta)$ 最大:

$$L(\beta,\Theta)=\prod_{i=1}^{N}p(\boldsymbol{x}_i)^{y_i}[1-p(\boldsymbol{x}_i)]^{(1-y_i)} \tag{4}$$

对 $L(\beta,\Theta)$ 求取对数可得到对数似然函数:

$$\frac{1}{N}\ln[L(\beta,\Theta)]=\frac{1}{N}\sum_{i=1}^{N}\left[y_i\left(\boldsymbol{x}_i^{\mathrm{T}}\beta+\frac{1}{2}\boldsymbol{x}_i^{\mathrm{T}}\Theta x_i\right)+\ln(1-p(\boldsymbol{x}_i))\right] \tag{5}$$

对式(5)进行当前估计值 $(\tilde{\beta},\tilde{\Theta})$ 的二元 2 阶泰勒展开得到等价的目标函数 $l(\beta,\Theta)$,式(5)到式(6)证明见附录 A。

$$l(\beta,\Theta)=\frac{1}{2}\omega_i\left(z_i-\sum_j\tilde{\beta}_jx_{ij}-\frac{1}{2}\sum_{j\neq k}\tilde{\Theta}_{jk}x_{ij}x_{ik}\right)^2+\frac{(y_i-\tilde{p}(\boldsymbol{x}_i))^2}{2\times\tilde{p}^2(\boldsymbol{x}_i)[1-\tilde{p}(\boldsymbol{x}_i)]^2}$$

$$+\left[y_i\left(\sum_j \tilde{\beta}_j x_{ij}+\sum_{j\neq k}\tilde{\Theta}_{jk}x_{ij}x_{ik}\right)+\ln(1-\tilde{p}(\boldsymbol{x}_i))\right] \tag{6}$$

其中，$\omega_i=\tilde{p}(\boldsymbol{x}_i)[1-\tilde{p}(\boldsymbol{x}_i)]$，$z_i=\sum_j \beta_j x_{ij}+\frac{1}{2}\sum_{j\neq k}\Theta_{jk}x_{ij}x_{ik}+\frac{y_i-\tilde{p}(\boldsymbol{x}_i)}{\tilde{p}(\boldsymbol{x}_i)[1-\tilde{p}(\boldsymbol{x}_i)]}$。

为达到收缩模型系数的目的，得到 1 阶主变量系数与 2 阶交互变量系数的稀疏解，提高模型稳定性，本文为目标函数 $l(\beta,\Theta)$ 添加系数的 Lasso 罚函数，λ_1、λ_2 为超参数，则：

$$(\beta,\Theta)=\arg\min_{\beta\in R^{p+1},\Theta\in R^{p\times p}} -l(\beta,\Theta)+\lambda_1\|\beta\|_1+\lambda_2\|\Theta\|_1 \tag{7}$$

交互分层思想是指：如果模型中包含了交互变量对应的 2 个主变量，那么应该允许交互变量进入模型。有 2 种交互变量分层情况：强分层是指 $\hat{\Theta}_{jk}\neq 0\Rightarrow\hat{\beta}_j\neq 0$ 和 $\hat{\beta}_k\neq 0$，弱分层是指 $\hat{\Theta}_{jk}\neq 0\Rightarrow\hat{\beta}_j\neq 0$ 或 $\hat{\beta}_k\neq 0$。只考察包含大的 1 阶变量间交互更为高效。即 $\hat{\Theta}_{jk}\neq 0\Rightarrow\hat{\beta}_j\neq 0$ 或 $\hat{\beta}_k\neq 0$。为式(7)添加弱分层约束，得：

$$\begin{aligned}(\beta,\Theta)=&\min_{\beta\in R^{p+1},\Theta\in R^{p\times p}} -l(\beta,\Theta)+\lambda_1\|\beta\|_1+\lambda_2\|\Theta\|_1\\ &\text{s.t. }\|\Theta_j\|_1\leqslant|\beta_j|\quad j=1,\cdots,p\end{aligned} \tag{8}$$

其中 Θ_j 是 Θ 的第 j 列。如果 $\hat{\Theta}_{jk}\neq 0$，那么有 $\|\hat{\Theta}_j\|_1>0$ 和 $\|\hat{\Theta}_k\|_1>0$，因此有 $\hat{\beta}_j\neq 0$ 或 $\hat{\beta}_k\neq 0$。

新增的约束条件强制分层，导致式(8)并不是凸函数，将式(8)进行凸松弛变换，得：

$$\begin{aligned}(\beta^{\pm},\Theta)=&\min_{\beta_0\in R,\,\beta^{\pm}\in R^p,\Theta\in R^{p\times p}} -l(\beta^+-\beta^-,\Theta)+\lambda_1 1^{\mathrm{T}}(\beta^++\beta^-)+\lambda_2\sum_j\|\Theta_j\|\\ &\text{s.t.}\left.\begin{array}{l}\|\Theta_j\|_1\leqslant\beta_j^++\beta_j^-\\ \beta_j^+\geqslant 0,\beta_j^-\geqslant 0\end{array}\right\}\ j=1,\cdots,p\end{aligned} \tag{9}$$

这里用 β^+、$\beta^-\in R^{p+1}$ 代替向量 $\beta=\beta^+-\beta^-$，$\beta^{\pm}=\max\{\pm\beta,0\}$，则 $\|\beta\|_1=\beta^++\beta^-$。

3 坐标下降算法以及 KKT 条件

坐标下降算法的思想是首先给出初始系数解；然后每次更新一个坐标的系数值，重复坐标更新过程，就可以得到所有坐标的一次更新值；重复系数更新过程就可以得到系数向量最终解。如果坐标满足独立条件，整个迭代过程将很快完成。

因此根据坐标下降法对式(9)求解。式(9)的拉格朗日函数为

$$\begin{aligned}L(\beta^+,\beta^-,\Theta)=&-l(\beta^+-\beta^-,\Theta)\\ &+(\lambda_1 1-\hat{\alpha}-\gamma^+1)^{\mathrm{T}}\beta^+\\ &+(\lambda_1 1-\hat{\alpha}-\gamma^-1)^{\mathrm{T}}\beta^-\\ &+<\operatorname{diag}(\lambda_2 1+\hat{\alpha})U,\Theta>\end{aligned} \tag{10}$$

其中 $U_{jk}=\begin{cases}\in[-1,1] & \hat{\Theta}_{kj}=0\\ \operatorname{sign}(\hat{\Theta}_{jk}) & \hat{\Theta}_{jk}\neq 0\end{cases}$，$\hat{\alpha}$ 是分层约束的对偶变量，$\gamma^{\pm}$ 是非负约束的对偶变量。上式可以分解为 p 个子问题进行求解：

$$\begin{aligned}L(\beta_j^+,\beta_j^-,\Theta_j)=&-l(\beta_j^+-\beta_j^-,\Theta_j)\\ &+(\lambda_1-\hat{\alpha}_j-\gamma^+)^{\mathrm{T}}\beta_j^+\\ &+(\lambda_1-\hat{\alpha}_j-\gamma^-)^{\mathrm{T}}\beta_j^-\\ &+<(\lambda_2+\alpha)U_j,\Theta_j>\end{aligned} \tag{11}$$

利用式(11)计算出的估计值可以作为凸函数问题式(9)的解。这可以通过卡罗需库恩塔克(KKT)条件进行求解。平稳条件为 $\frac{\partial L}{\partial\hat{\beta}_j^{\pm}}=0$，$\frac{\partial L}{\partial\hat{\Theta}_{jk}}=0$ $(j=1,\cdots,p;k=1,\cdots,K)$，$K$ 是 Θ_j 中的所有元素个数，互补松弛条件为 $\gamma_j^{\pm}\hat{\beta}_j^{\pm}=0$，$\hat{\alpha}_j(\hat{\Theta}_{j1}-\hat{\beta}_j^+-\hat{\beta}_j^-)=0$，$\hat{\beta}^{\pm}\geqslant 0$，$\hat{\Theta}_{j1}\leqslant\hat{\beta}_j^++\hat{\beta}_j^-$，$\hat{\alpha},\hat{\gamma}^{\pm}\geqslant 0$。

本文分 3 种情况对并对 $\hat{\beta}_j^+-\hat{\beta}_j^-$ 进行讨论。

(1)假设 $\hat{\beta}_j^+\geqslant 0$，$\hat{\beta}_j^-=0$，则 $\hat{\beta}_j^+-\hat{\beta}_j^-=x_j^{\mathrm{T}}(z_i-A_i)+\hat{\beta}_j^+-\hat{\beta}_j^--\frac{\lambda_1+\hat{\alpha}_j}{\omega_i}$；

(2)假设 $\hat{\beta}_j^+=0$，$\hat{\beta}_j^-\geqslant 0$，则 $\hat{\beta}_j^+-\hat{\beta}_j^-=x_j^{\mathrm{T}}(z_i-A_i)+\hat{\beta}_j^+-\hat{\beta}_j^-+\frac{\lambda_1+\hat{\alpha}_j}{\omega_i}$；

(3)假设 $\hat{\beta}_j^+>0$，$\hat{\beta}_j^->0$：则与假设 $\hat{\beta}_j^+\hat{\beta}_j^-=0$ 矛盾。

其中，$A_i=\sum_j\hat{\beta}_j x_{ij}+\frac{1}{2}\sum_{j\neq k}\tilde{\Theta}_{jk}x_{ij}x_{ik}$，$z_i=\sum_j$

$$\hat{\beta}_j x_{ij} + \frac{1}{2}\sum_{j\neq k}\tilde{\Theta}_{jk}x_{ij}x_{ik} + \frac{y_i - \tilde{p}(x_i)}{\tilde{p}(x_i)[1-\tilde{p}(x_i)]}$$

模型系数 $\hat{\beta}_j = \hat{\beta}_j^+ - \hat{\beta}_j^-$ 更新公式如下:

$$\hat{\beta}_j^+ - \hat{\beta}_j^- = S[x_{ij}(z_i - A_i) + \hat{\beta}_j^+ - \hat{\beta}_j^-, \frac{\lambda_1 + \hat{\alpha}_j}{\omega_i}] \tag{12}$$

其中,S 为软阈值 $S(c,\lambda) = \text{sign}(c)(|c|-\lambda)_+$。

由平稳条件 $\frac{\partial L}{\partial \hat{\Theta}_{jk}} = 0$,得

$$\frac{1}{2}\omega_i \cdot 2(z_i - \sum_j \hat{\beta}_j x_{ij} - \frac{1}{2}\sum_{j\neq k}\tilde{\Theta}_{jk}x_{ij}x_{ik}) \cdot (-\frac{1}{2}x_{ij}x_{ik}) + (\lambda_2 + \alpha_j)U_{jk} = 0$$

令 $\gamma^{(-jk)} = z_i - \sum_j \hat{\beta}_j x_{ij} - \frac{1}{2}\sum_{j\neq k}\tilde{\Theta}_{jk}x_{ij}x_{ik} + \hat{\Theta}_{jk}x_{ij}x_{ik}$,整理上式得:

$$\hat{\Theta}_{jk} = \frac{\gamma^{(-jk)} \cdot x_{ij}x_{ik} - 2(\lambda_2 + \alpha_j)U_{jk}/\omega_i}{(x_{ij}x_{ik})^2}$$

下面分 3 种情况讨论 $\hat{\Theta}_{jk}$ 的值:

(1)假设 $\hat{\Theta}_{jk} > 0$,$U_{jk} = 1$,则 $\hat{\Theta}_{jk} = \frac{\gamma^{(-jk)} \cdot x_{ij}x_{ik} - 2(\lambda_2 + \alpha_j)/\omega_i}{(x_{ij}x_{ik})^2}$

(2)假设 $\hat{\Theta}_{jk} < 0$,$U_{jk} = -1$,则 $\hat{\Theta}_{jk} = \frac{\gamma^{(-jk)} \cdot x_{ij}x_{ik} + 2(\lambda_2 + \alpha_j)/\omega_i}{(x_{ij}x_{ik})^2}$

(3)假设 $\hat{\Theta}_{jk} = 0$,则 $\hat{\Theta}_{jk} = 0$,模型系数 $\hat{\Theta}_{jk}$ 更新公式为

$$\hat{\Theta}_{jk} = \frac{S[\gamma^{(-jk)} \cdot x_{ij}x_{ik}, 2(\lambda_2 + \alpha_j)/\omega_i]}{(x_{ij}x_{ik})^2} \tag{13}$$

现定义 $f(\hat{\alpha}_j) = \hat{\Theta}_{j1} - \hat{\beta}_j^+ - \hat{\beta}_j^-$,则有:

$$f(\hat{\alpha}_j) = \| \sum_{k=1}^{P} S[(x_{ij}x_{ik})^{\mathrm{T}}(z_i - A_i + \hat{\Theta}_{jk}x_{ij}x_{ik}), \frac{2(\lambda_2 + \hat{\alpha}_j)}{\omega_i}]/\| x_{ij}x_{ik} \|_2^2 \|_1 - \| S[x_{ij}(z_i - A_i) + \hat{\beta}_j^+ - \hat{\beta}_j^-, \frac{\lambda_1 + \hat{\alpha}_j}{\omega_i}] \|_1$$

,最后保存的 KKT 条件只包含 $\hat{\alpha}$:$\hat{\alpha}f(\hat{\alpha}) = 0$,$f(\hat{\alpha}) \leqslant 0$,$\hat{\alpha} \geqslant 0$。观察发现 f 关于 $\hat{\alpha}$ 是非增的,基于其是分段线性,可求解出 $\hat{\alpha}$。

弱分层交互 Lasso 罚 logistic 回归模型的改进坐标下降算法的总体思路如下。首先对样本进行了列归一化,即 $\frac{1}{N}\sum_{i=1}^{N}x_{ij}^2 = 1$,初始化模型系数 β_j、Θ_j 和超参数 λ_1,λ_2;其次计算 $\hat{\alpha}$,z_i,ω_i,A_i,利用式(12)和(13)更新模型系数 β_j 和 Θ_{jk},重复所有坐标和整个系数更新过程直至收敛。提出算法复杂度分析如下,模型系数 β_j 和 Θ_j 实际上是软阈值函数的计算复杂度,加上函数参数的计算复杂度,其中式(12)中第 1 个参数的计算复杂度为 1 次乘法和 3 次加法运算,A_i、ω_i 和 $\gamma^{(-jk)}$ 的计算复杂度大约为 $O(2p^2)$ 次乘法和 $O(p^2)$ 次加法运算。

4　实验结果与分析

4.1　UCI 数据的实验结果与分析

本文选用 UCI 学习数据库中 4 个数据集,即 Breast-cancer-wisconsin 数据集、Ionosphere 数据集、Liver-disorders 数据集和 Sonar 数据集,如表 1 所示。

表 1　UCI 数据集信息

数据集	样本数	变量维数	面积交互维数	类别数
Breast-cancer-wisconsin	683	9	36	2
Ionosphere	351	33	528	2
Liver-disorders	345	6	15	2
Sonar	208	60	1 770	2

本文对 4 个数据集进行 20 次 10 倍交叉验证,实验中选择 $\lambda_1 = 2\lambda_2 = \lambda$。完成了 logistic 回归的分层交互 Lasso 方法的实验,采用 R 编程,结果包括非 0 变量系数个数、平均错误率、标准差、CPU 时间以及所选取的 λ 值估计值 Lamhat,见表 2 所示。图 2 ~图 5 分别是本文所提出方法在 4 种数据集上 10 倍交叉验证的实验结果,下横坐标表示 λ 的对数值,上横坐标表示非零变量数,纵坐标是错误率。

表 2 Logistic 回归的分层交互 Lasso 方法 10 倍交叉验证的实验结果

性能指标 / 数据集	变量系数非 0 的数目	平均错误率(%)	标准差	时间(s)	Lamhat
Breast-cancer-wisconsin	18	3.0	0.01	9.49	3.14
Ionosphere	86	29.0	0.02	124.63	2.21
Liver-disorders	19	26.0	0.02	12.11	2.37
Sonar	117	14.0	0.02	150.69	0.73

图 2 所示为 Breast-cancer-wisconsin 数据实验结果,当所选择变量超过 11 个时,错误率开始平稳,最低为 3.0%。从图中可以看出标准差较小,说明了本文选用方法的高效性和稳定性。图 3 所示为 Ionosphere 数据实验结果,当所选择变量为 86 个时,错误率最低为 29.0%,标准差较小。错误率最低时的变量数目高于原始变量数目,故变量交互为分类提供了分类信息。图 4 为 Liver-disorder 数据实验结果,当所选择变量为 19 个时,错误率最低为26.0%,标准差为 0.02。图 5 为 Sonar 数据实验结果,错误率变化的趋势大致与所选用变量数呈分段线性关系。当所选择的变量大于 80 时,错误率较低,最低为 14.0%。

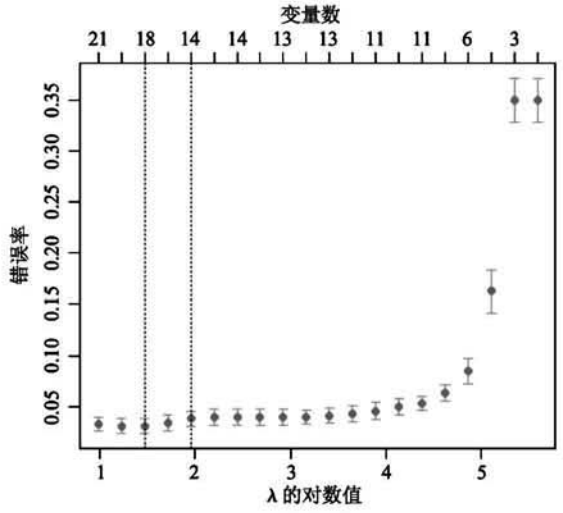

图 2 Wisconsin-breast-cancer 实验结果

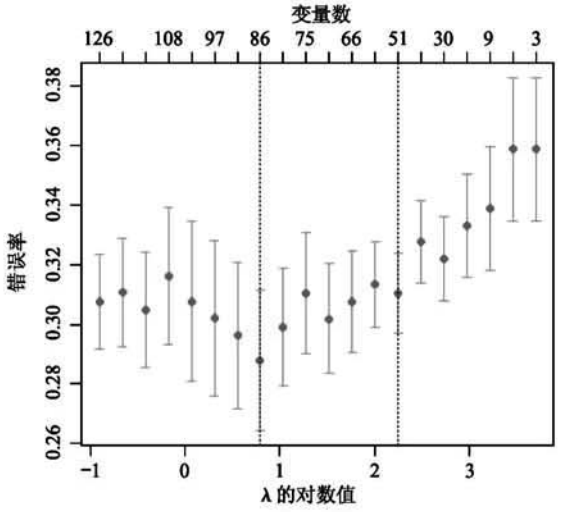

图 3 Ionosphere 实验结果

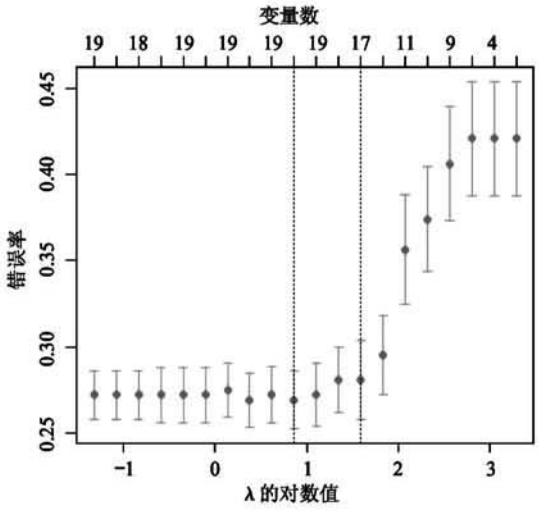

图 4 Liver-disorders 实验结果

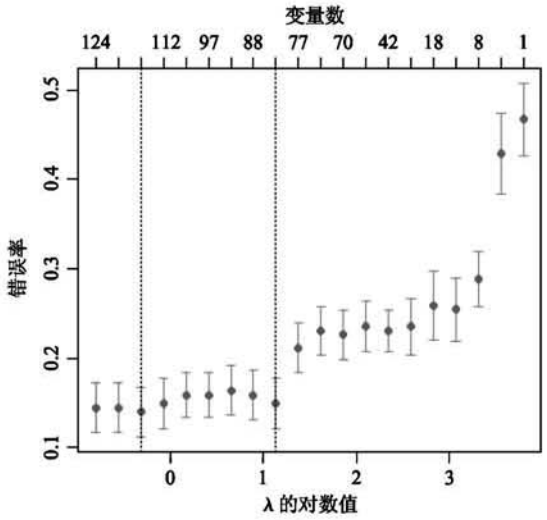

图 5 Sonar 实验结果

4.2　实验讨论

分层交互 Lasso 定义为考虑 1 阶变量和 2 阶交互变量也考虑分层。Lasso 方法定义为只考虑 1 阶变量不考虑 2 阶交互变量,交互 Lasso 方法定义为考虑 1 阶变量和 2 阶交互变量但不考虑分层。Lasso、交互 Lasso 方法和传统模式识别方法在 4 种数据集上 20 次 10 倍交叉验证的实验结果分别为表 3、表 4 和表 5 所示。实验结果说明了本文所选用分层交互 logistic 回归方法分类效果好,模型稳定,突出展现了几何代数变量交互思想和弱分层思想的优势。

表 3　Lasso 方法在 4 种数据集上 10 倍交叉验证的实验结果

数据集	平均错误率(%)	方差	时间(s)
Breast-cancer-wisconsin	3.22	0.0001	0.3744
Ionosphere	31.15	0.0075	0.4863
Liver-disorders	30.77	0.0077	0.0908
Sonar	22.13	0.0121	15.6806

表 4　交互 Lasso 方法在 4 种数据集上 10 倍交叉验证的实验结果

数据集	平均错误率(%)	方差	时间(s)
Breast-cancer-wisconsin	2.93	0.0001	12.1213
Ionosphere	27.00	0.0089	3.1734
Liver-disorders	26.36	0.0058	4.9190
Sonar	20.63	0.0115	4.1041

表 5　传统模式识别方法在 4 种数据集上 10 倍交叉验证的实验结果

数据集	分类器	平均错误率(%)	方差	时间(s)
Breast-cancer-wisconsin	支持向量机	3.42	0.0019	1.3092
	线性判别分析	3.96	0.0007	0.2429
	二次判别分析	4.92	0.0014	0.2309
	最近邻	4.61	0.0030	0.6636
	决策树	4.99	0.0041	1.6201
Ionosphere	支持向量机	35.83	0.0063	2.9149
	线性判别分析	33.10	0.0090	0.6168
	二次判别分析	32.01	0.0076	0.4516
	最近邻	28.30	0.0090	0.6363
	决策树	38.23	0.0408	3.0865
Liver-disorders	支持向量机	30.80	0.0100	1.1401
	线性判别分析	31.31	0.0089	0.2474
	二次判别分析	40.19	0.0119	0.1958
	最近邻	36.89	0.0124	0.6089
	决策树	36.94	0.0187	1.9921
Sonar	支持向量机	25.70	0.0197	1.2181
	线性判别分析	25.13	0.0145	0.5897
	二次判别分析	23.80	0.0205	0.4594
	最近邻	13.09	0.0101	0.5345
	决策树	35.91	0.0303	0.7153

本文方法和其他文献方法区别较大，将其同文献[13]进行了对比，logistic 回归方法的平均错误率低于文献[13]回归方法，提出的坐标下降法的训练时间更少。

4.3 数值仿真实验与讨论

本文取样本 $n = 200, p = 20$，考虑 2 阶交互变量，基于式(1)给出下面 3 种仿真情况：(1) 真实模型是分层 $\Theta_{jk} \neq 0 \Rightarrow \beta_j \neq 0$ 或 $\beta_k \neq 0$ $(j,k = 1,\cdots,p)$，β 中有 10 个元素是非零的，Θ 中有 20 个元素是非零的；(2) 真实模型是仅有 2 阶几何代数变量交互：$\beta_j = 0$ $(j = 1,\cdots,p)$，Θ 中有 20 个元素是非零的；(3) 真实模型是仅有 1 阶变量：$\Theta_{jk} = 0$ $(j,k = 1,\cdots,p)$，β 中有 10 个元素是非零的。1 阶变量信噪比为 1.5，2 阶交互变量信噪比为 1。100 次仿真实验的结果如图 6 所示。当真实模型是分层时，分层交互 Lasso 完成最好，Lasso 完成最差，如图 6(a)所示。当真实模型仅有 2 阶交互时，交互 Lasso 完成最好，分层交互 Lasso 居中，Lasso 完成最差，如图 6(b)所示。这与所期待的结果有所偏差，原因可能是分层交互模型将部分交互项当做 1 阶变量进行模型拟合。当真实模型是仅有 1 阶变量时，Lasso 完成最好，分层交互 Lasso 居中，交互 Lasso 完成最差，如图 6(c)所示。

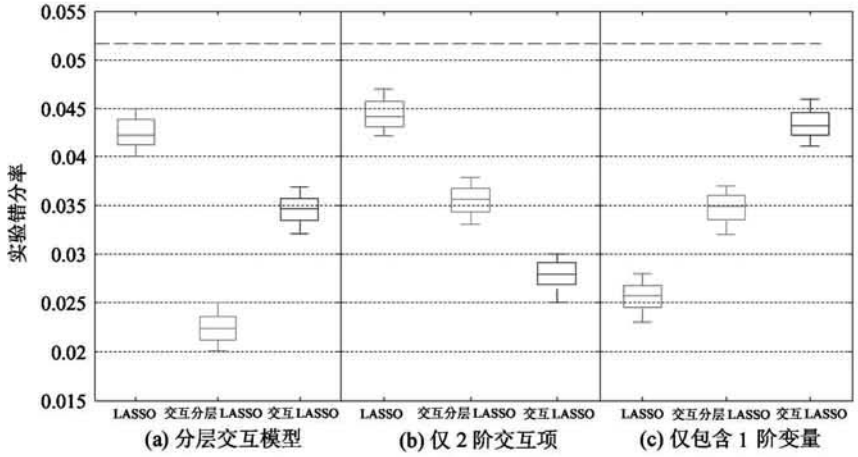

图 6 3 种 Lasso 在仿真实验中的错误率比较

实际数据真实模型可能是分层交互模型，即包含 1 阶变量，又包含几何代数 2 阶交互变量。不但如此，分层也不可忽视，因为对模型影响大的几何代数 2 阶交互变量预示着对应的 1 阶变量的贡献也重要。故分层交互 Laaso 方法最大限度地拟合了这种情况。

4.4 基于智能手机传感器数据的日常生活活动识别

Anguita 等人[21]采集智能手机传感器数据，利用硬件友好的多类支持向量机方法进行日常活动识别。这对失能或老年人的日常活动监护具有重要意义。日常活动数据集可以从文献[21]下载，这里被用来评价分层交互 Lasso。数据集实验结果采自 30 个 19～48 岁的志愿者，每人进行 6 次实验。所有人在执行活动之前都要将智能手机放在腰部。为了方便进行数据标签，实验进行了视频录制。实验采用三星 Galaxy S2 手机作为终端，因为内置了用于测量 3 维线性加速度和角速度的加速度计和陀螺仪，其采样频率为 50 Hz，这对于捕捉人体运动有效。本文实验采用了上楼和下楼 2 个活动类，其训练集分别为 986 个样本和 1 073 个样本，测试集分别为 420 个样本和 471 个样本，变量维数都是 561，由采集传感器信号的时域和频域特征组成。3 种 Lasso 方法和常用模式识别方法的实验结果如表 7。可见 Lasso 方法比常用模式识别方法更好考虑了变量选择和变量交互问题，而分层交互 Lasso 分类结果最好，训练时间和测试时间都较少。

表 7　活动识别数据的实验结果

方法 \ 性能指标	训练集平均错误率(%)	训练时间(s)	测试集平均错误率(%)	测试时间(s)
Lasso	0.00	0.2649	1.46	0.0312
交互 Lasso	0.00	0.8807	1.35	0.0953
分层交互 Lasso	0.00	0.6012	1.12	0.0556
线性支持向量机	0.00	0.9516	1.23	0.0936
最近邻	0.00	0.0000	9.20	0.7956
3-近邻	0.00	0.0000	8.98	0.7488
二次判别分析	6.61	1.4196	4.04	0.2184
决策树	0.00	0.7332	14.93	0.1092

5　结　论

从考虑变量间的交互性出发,基于分层 Lasso 思想,本文提出了 logistic 回归的分层交互 Lasso 模型和算法,给出了模型定义、约束条件、凸松弛条件、凸优化方法和坐标下降算法的系数解。4 组 UCI 数据集和 1 组真实的日常生活活动识别数据集的实验结果揭示了交互性在模型分类中广泛存在,变量交互对响应类别变量有贡献。而不考虑交互变量的 Lasso 方法分类性能欠佳,不考虑分层而仅考虑交互变量的交互 Lasso 方法分类性能也不够好。这说明变量交互和分层思想是广义线性模型能够成功应用的 2 个重要原因。进一步研究方向包括广义梯度下降法或交替方向乘子法等其他模型凸优化方法,多类 logistic 回归方法的分层交互 Lasso 方法,将新方法用于活动识别问题的多传感器交互研究中。

参考文献

[1] Tibshirani R. Regression shrinkage and selection via the lasso[J]. *Journal of the Royal Statistical Society*, 1996, 58(1):267-288

[2] Park M Y, Hastie T. L1-regularization path algorithm for generalized linear models[J]. *Journal of the Royal Statistical Society B*, 2007, 69(4):659-677

[3] Wu T T, Chen Y F, Hastie T, et al. Genome-wide association analysis by lasso penalized logistic regression[J]. *Bioinformatics*, 2009, 25(6): 714-721

[4] Friedman J, Hastie T, Tibshirani B. Regularization paths for generalized linear models via coordinate descent[J]. *Journal of Statistical Software*, 2010, 33(1):1-22

[5] Lim M, Hastie T. Learning interactions via hierarchical group-lasso regularization[J]. *Journal of Computational and Graphical Statistics*, 2015, 24(3):627-654

[6] Schwender H, Ickstadt K. Identification of snp interactins using logic regression[J]. *Biostatistics*, 2008, 9(1): 187-198

[7] Jiang H, Dong Y. Structural regularization in quadratic logistic regression model[J]. *Knowledge-Based Systems*, 2019, 163(1):842-857

[8] Wu J, Devlin B, Ringquist S, et al. Screen and clean: a tool for identifying interactions in genome-wide association studies[J]. *Genetic Epidemiology*, 2010, 34(3):275-285

[9] Nardi Y, Rinaldo A. The log-linear group-lasso estimator and its asymptotic properties[J]. *Bernoulli*, 2012, 18(3):945-974

[10] Yuan M, Joseph V R, Lin Y. An efficient variable selection approach for analyzing designed experiments[J]. *Technometrics*, 2007, 49(4):430-439

[11] Chipman H. Bayesian variable selection with related predictors[J]. *The Canadian Journal of Statistics*, 1996, 24(1):17-36

[12] Choi N H, Li W, Zhu J. Variable selection with the strong heredity constraint and its oracle property[J]. *Journal of the American Statistical Association*, 2010, 105(489):354-364

[13] Bien J, Taylor J, Tibshirani R. A lasso for hierarchical interactions[J]. *The Annals of Statistics*, 2013,41(3): 1111-1141

[14] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables[J]. *Journal of the Royal Statistical Society: Series B*, 2006,68(1):49-67

[15] Jenatton R, Audibert J Y, Bach F. Structured variable selection with sparsity-inducing norms[J]. *Journal of Machine Learning Research*, 2011, 12(10): 2777-2824

[16] Radchenko P, James G M. Variable selection using adaptive nonlinear interaction structures in high dimensions[J]. *Journal of the American Statistical Association*, 2010, 105 (492): 1541-1553

[17] Bach F, Jenatton R, Mairal J, et al. Structured sparsity through convex optimization[J]. *Statistical Science*, 2012, 27(4):450-468

[18] Ruczinski I, Kooperberg C L M. Logic regression[J]. *Journal of Computational and Graphical Statistics*, 2003, 12(3):475-511

[19] Peter H, Xue J. On selecting interacting features from high-dimensional data[J]. *Computational Statistics and Data Analysis*, 2014,71(1):694-708

[20] 王金甲,李静,张涛. 二次映射和遗传算法用于鉴别可视化特征提取[J]. 系统仿真学报, 2009, 21(16): 5080-5083

[21] Anguita D, Ghio A, Oneto L, et al. Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine[C]// Proceedings of the International Workshop of Ambient Assisted Living (IWAAL 2012), Vitoria-Gasteiz, Spain, 2012, 7657: 216-223

Weak hierarchical interactive Lasso penalized logistic regression model and improved coordinate descent algorithms

Li Jing*, Yu Hui*, Wang Jinjia**

(*School of Science, Yanshan University, Qinhuangdao 066004)

(**School of Information Science and Engineering, Hebei Provincial Key Laboratory of Information Transmission and Signal Processing, Yanshan University, Qinhuangdao 066004)

Abstract

Based on hierarchy theory and variable interactions, the hierarchical interactive Lasso penalized logistic regression model is proposed. First, an interaction model and the hierarchical constraint conditions are defined, and then the coordinate descent algorithms are used to solve model coefficients. The experiments on four UCI data sets and activities of daily living recognition are performed. The experimental results prove that the variable interaction makes great contribution to the classification results, and hierarchy contributes to the classification results. Hierarchical interactive Lasso shows more advantages than the Lasso and interactive Lasso.

Key words: variables interaction, hierarchy, Lasso, logistic regression, coordinate descent

附录 A 式(5)到式(6)证明过程。

为简便定义 $p(x_i) = p_i$, 令

$$L_Q = \frac{1}{N}\ln[L(\beta,\Theta)] = \frac{1}{N}\sum_{i=1}^{N}[y_i(x_i^{\mathrm{T}}\beta + \frac{1}{2}x_i^{\mathrm{T}}\Theta x_i) + \ln(1-p_i)] \tag{A-1}$$

首先求出 L_Q 关于 β,Θ 的一阶以及二阶偏导数和混合偏导数:

$$\frac{\partial L_Q}{\partial \beta} = \frac{\partial \frac{1}{N}\sum_{i=1}^{N}[y_i(x_i^{\mathrm{T}}\beta + \frac{1}{2}x_i^{\mathrm{T}}\Theta x_i) + \ln(1-p_i)]}{\partial \beta} = \frac{1}{N}\sum_{i=1}^{N}x_i(y_i - p_i)$$

$$\frac{\partial^2 L_Q}{\partial \beta^2} = \frac{\partial}{\partial \beta}\frac{1}{N}\sum_{i=1}^{N}x_i^{\mathrm{T}}(y_i - p_i) = -\frac{1}{N}\sum_{i=1}^{N}x_i^{\mathrm{T}}\frac{\partial p_i}{\partial \beta} = -\frac{1}{N}\sum_{i=1}^{N}(1-p_i)\cdot p_i \cdot x_i^{\mathrm{T}}\cdot x_i$$

$$\frac{\partial L_Q}{\partial \Theta}=\frac{\partial \frac{1}{N}\sum_{i=1}^{N}\left[y_i\left(x_i^{\mathrm{T}}\beta+\frac{1}{2}x_i^{\mathrm{T}}\Theta x_i\right)+\ln(1-p_i)\right]}{\partial \Theta}=\frac{1}{2N}\sum_{i=1}^{N}x_i^{\mathrm{T}}x_i(y_i-p_i)$$

$$\frac{\partial^2 L_Q}{\partial\,\Theta^2}=\frac{\partial}{\partial\Theta}\frac{1}{2N}\sum_{i=1}^{N}x_i^{\mathrm{T}}x_i(y_i-p_i)=-\frac{1}{2N}\sum_{i=1}^{N}x_i^{\mathrm{T}}x_i\frac{\partial p_i}{\partial\Theta}=-\frac{1}{4N}\sum_{i=1}^{N}(1-p_i)\cdot p_i\cdot\|x_i\|^2$$

$$\frac{\partial^2 L_Q}{\partial\beta\partial\Theta}=\frac{\partial}{\partial\beta}\frac{1}{N}\sum_{i=1}^{N}x_i^{\mathrm{T}}(y_i-p_i)=-\frac{1}{N}\sum_{i=1}^{N}x_i^{\mathrm{T}}\frac{\partial p_i}{\partial\Theta}=-\frac{1}{2N}\sum_{i=1}^{N}(1-p_i)\cdot p_i\cdot x_i^{\mathrm{T}}\cdot x_i\cdot x_i^{\mathrm{T}}$$

对(A-1)式进行二元泰勒展开:

$$\begin{aligned}
l(\beta,\Theta)&=L_Q+(\beta-\tilde{\beta})\cdot\frac{\partial L_Q}{\partial\tilde{\beta}}+(\Theta-\tilde{\Theta})\cdot\frac{\partial L_Q}{\partial\tilde{\Theta}}\\
&\quad+\frac{1}{2}\left[(\beta-\tilde{\beta})\cdot\frac{\partial L_Q}{\partial\tilde{\beta}^2}+2(\beta-\tilde{\beta})^{\mathrm{T}}(\Theta-\tilde{\Theta})\frac{\partial^2 L_Q}{\partial\beta\partial\Theta}+(\Theta-\tilde{\Theta})\cdot\frac{\partial L_Q}{\partial\tilde{\Theta}^2}\right]\\
&=\frac{1}{N}\sum_{i=1}^{N}\left[y_i\left(x_i^{\mathrm{T}}\tilde{\beta}+\frac{1}{2}x_i^{\mathrm{T}}\tilde{\Theta}x_i\right)+\ln(1-\tilde{p}_i)\right]+(\beta-\tilde{\beta})\cdot\frac{1}{N}\sum_{i=1}^{N}x_i(y_i-\tilde{p}_i)+(\Theta-\tilde{\Theta})\\
&\quad\cdot\frac{1}{2N}\sum_{i=1}^{N}x_i^{\mathrm{T}}x_i(y_i-\tilde{p}_i)-\frac{1}{2}(\beta-\tilde{\beta})\cdot\frac{1}{N}\sum_{i=1}^{N}(1-\tilde{p}_i)\cdot\tilde{p}_i\cdot x_i^{\mathrm{T}}\cdot x_i-2(\beta-\tilde{\beta})^{\mathrm{T}}(\Theta-\tilde{\Theta})\\
&\quad\cdot\frac{1}{2N}\sum_{i=1}^{N}(1-\tilde{p}_i)\cdot\tilde{p}_i\cdot x_i^{\mathrm{T}}\cdot x_i\cdot x_i^{\mathrm{T}}-(\Theta-\tilde{\Theta})\cdot\frac{1}{4N}\sum_{i=1}^{N}(1-\tilde{p}_i)\cdot\tilde{p}_i\cdot\|x_i\|^2\\
&=-\frac{1}{2N}\sum_{i=1}^{N}(1-\tilde{p}_i)\cdot\tilde{p}_i\left\{\left[x_i^{\mathrm{T}}(\beta-\tilde{\beta})\right]^2+(\beta-\tilde{\beta})^{\mathrm{T}}(\Theta-\tilde{\Theta})x_i\cdot x_i^{\mathrm{T}}x_i+\frac{1}{4}\left[x_i^{\mathrm{T}}(\Theta-\tilde{\Theta})x_i\right]^2\right\}\\
&\quad+\frac{1}{N}\sum_{i=1}^{N}\left[y_i\left(x_i^{\mathrm{T}}\tilde{\beta}+\frac{1}{2}x_i^{\mathrm{T}}\tilde{\Theta}x_i\right)+\ln(1-\tilde{p}_i)\right]+\frac{1}{N}\sum_{i=1}^{N}x_i^{\mathrm{T}}(\beta-\tilde{\beta})(y_i-\tilde{p}_i)\\
&\quad+\frac{1}{2N}\sum_{i=1}^{N}x_i^{\mathrm{T}}(\Theta-\tilde{\Theta})x_i(y_i-\tilde{p}_i)\\
&=-\frac{1}{2N}\sum_{i=1}^{N}(1-\tilde{p}_i)\cdot\tilde{p}_i\left\{x_i^{\mathrm{T}}(\beta-\tilde{\beta})+\frac{1}{2}x_i^{\mathrm{T}}(\Theta-\tilde{\Theta})x_i\right\}^2+\frac{1}{N}\sum_{i=1}^{N}x_i^{\mathrm{T}}(\beta-\tilde{\beta})(y_i-\tilde{p}_i)\\
&\quad+\frac{1}{2N}\sum_{i=1}^{N}x_i^{\mathrm{T}}(\Theta-\tilde{\Theta})x_i(y_i-\tilde{p}_i)+\frac{1}{N}\sum_{i=1}^{N}\left[y_i\left(x_i^{\mathrm{T}}\tilde{\beta}+\frac{1}{2}x_i^{\mathrm{T}}\tilde{\Theta}x_i\right)+\ln(1-\tilde{p}_i)\right]\\
&=-\frac{1}{2N}\sum_{i=1}^{N}(1-\tilde{p}_i)\cdot\tilde{p}_i\left\{x_i^{\mathrm{T}}(\beta-\tilde{\beta})+\frac{1}{2}x_i^{\mathrm{T}}(\Theta-\tilde{\Theta})x_i+\frac{y_i-\tilde{p}_i}{\tilde{p}_i[1-\tilde{p}_i]}\right\}^2\\
&\quad+\frac{1}{N}\sum_{i=1}^{N}\left[y_i\left(x_i^{\mathrm{T}}\tilde{\beta}+\frac{1}{2}x_i^{\mathrm{T}}\tilde{\Theta}x_i\right)+\ln(1-\tilde{p}_i)\right]-\frac{1}{2N}\sum_{i=1}^{N}\left[\frac{y_i-\tilde{p}_i}{\tilde{p}_i[1-\tilde{p}_i]}\right]^2
\end{aligned}$$

令 $\omega_i=\tilde{p}(x_i)[1-\tilde{p}(x_i)]$,$z_i=\sum_j\beta_j x_{ij}+\frac{1}{2}\sum_{j\neq k}\Theta_{jk}x_{ij}x_{ik}+\frac{y_i-\tilde{p}(x_i)}{\tilde{p}(x_i)[1-\tilde{p}(x_i)]}$,$(\tilde{\beta},\tilde{\Theta})$ 为展开点,则上式整理为:

$$\begin{aligned}
l(\beta,\Theta)&=-\frac{1}{2N}\sum_{i=1}^{N}\omega_i\left(z_i-\sum_j\tilde{\beta}_j x_{ij}-\frac{1}{2}\sum_{j\neq k}\tilde{\Theta}_{jk}x_{ij}x_{ik}\right)^2+\frac{1}{N}\sum_{i=1}^{N}\left[y_i\left(x_i^{\mathrm{T}}\tilde{\beta}+\frac{1}{2}x_i^{\mathrm{T}}\tilde{\Theta}x_i\right)+\ln(1-\tilde{p}_i)\right]\\
&\quad-\frac{1}{2N}\sum_{i=1}^{N}\left[\frac{y_i-\tilde{p}_i}{\tilde{p}_i[1-\tilde{p}_i]}\right]^2
\end{aligned}\tag{A-2}$$