doi:10. 16055/j. issn. 1672-058X. 2021. 0005. 018

基于预测变量图结构的高维逻辑回归模型*

黄文静, 邓　丹**, 杜杰琳, 吴明月
(重庆医科大学 公共卫生与管理学院,重庆 400016)

摘　要:针对高维数据集,提出一种利用预测变量之间的图结构信息来改进稀疏逻辑回归模型的方法。该方法通过利用高维图结构数据或者重叠组结构来进行逻辑回归建模,即使预测变量的图结构未知,该方法仍适用,当图结构为某些特殊形式时,目前流行的方法,如 Adaptive Lasso,(Overlapping) Group Lasso 和岭回归都可以看作是该模型方法的特例。数值模拟和实例分析应用表明:该方法能有效地利用预测变量图结构信息,提高模型在估计、预测以及变量选择等方面的表现,并且该模型在有限样本情形下是有效的;该模型方法克服了数据集的维数问题,利用高维数据的图结构提高了稀疏逻辑回归模型的性能,可广泛应用于高通量基因数据集的疾病分类研究中。

关键词:逻辑回归;高维数据;图结构;Lasso;稀疏性
中图分类号:C81　　**文献标志码**:A　　**文章编号**:1672-058X(2021)05-0107-07

0　引　言

随着信息和技术的快速发展,数据搜集能力越来越强,数据的规模和复杂程度都是前所未见的,特别是激增的变量维数给传统的统计学带来了巨大挑战。在高维情形下,变量个数 p 要大于样本量 n,而且高维变量之间通常具有很强的相依性,此时,设计矩阵 $\boldsymbol{X}$ 不是满秩的,回归参数的估计往往也不是唯一的[1]。因此,如何对高维变量建模和参数估计是统计学关注的重点问题。正则化方法是处理高维数据常用的一种方法。Lasso[2] 是人们熟知的正则化方法,但是当数据中存在某种结构(如组结构)时,Lasso 不能很好地利用这种结构。因此,作为 Lasso 的推广,随后进一步提出 Group Lasso[3] 和 Overlapping Group Lasso[4] 等,这些方法在高维数据分析中均得到了广泛的应用,但是这些方法通常需要提前给出预测变量的组结构,然而在有些应用中很难得到这种结构。相比组结构,预测变量的图结构更容易获得。在生物学研究中,每个个体包含了成千上万的基因,这些基因有的单独作用,有的相互作用,而基因之间相互作用的大量信息可以用来构建预测变量的图结构,其中基因表示图中的节点,调控关系表示图中的边。即使在没有先验信息的条件下,仍然可以通过协差阵(精度矩阵)的稀疏估计[5-7] 构造这些基因的图结构。Yu 和 Liu[8] 在线性模型的框架下,提出一种利用预测变量图结构的方法,该方法具有一定的普遍性。目前,逻辑回归是处理分类数据的有效工具,正则化方法也被广泛应用于处理大 p 小 n 的逻辑回归问题[9-10],本文将在文献[8]的基础上

收稿日期:2020-09-14;**修回日期**:2020-12-21.
* **基金项目**:重庆市基础研究与前沿探索专项课题(CSTC2018JCYJA0135).
作者简介:黄文静(1994—),女,湖北随州人,硕士研究生,从事生物统计和高维数据分析研究.
** **通讯作者**:邓丹(1977—),女,四川泸州人,博士,教授,从事人群健康与统计决策研究. Email:100079@ cqmu. edu. cn.

将图结构应用到逻辑回归模型中,为了评价这一方法的性能,将其与现有的许多方法进行比较,研究比较了不同图结构下的仿真实例,还将其应用到乳腺癌基因实例中,仿真实验结果和实际数据分析表明:该方法在估计、预测和模型的变量选择上均具有优越性。

1 基于预测变量图结构的逻辑回归模型

1.1 逻辑回归模型的改进

设$(\boldsymbol{X},\boldsymbol{Y})$为随机变量,其中$\boldsymbol{X}\in R^p$,$\boldsymbol{Y}\in\{0,1\}$。假设$\boldsymbol{X}$服从均值为$\boldsymbol{O}_{p\times1}$,协差阵为$\boldsymbol{\Sigma}$的多元正态分布。给定$\boldsymbol{X}=\boldsymbol{x}$,$\boldsymbol{Y}$的条件分布记为$P_{\boldsymbol{\beta}^*}(\boldsymbol{Y}=1\mid\boldsymbol{x})=p_{\boldsymbol{\beta}^*}(\boldsymbol{x})$。逻辑回归模型为

$$\log\left\{\frac{p_{\boldsymbol{\beta}^*}(x)}{1-p_{\boldsymbol{\beta}^*}(x)}\right\}=\boldsymbol{f}_{\boldsymbol{\beta}^*}(\boldsymbol{x})$$

其中,$\boldsymbol{f}_{\boldsymbol{\beta}^*}(\boldsymbol{x})=\boldsymbol{\beta}_0^*+\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{x}$。设$(\boldsymbol{X}_1,\boldsymbol{Y}_1),\cdots,(\boldsymbol{X}_n,\boldsymbol{Y}_n)$为来自总体$(\boldsymbol{X},\boldsymbol{Y})$的独立同分布的样本,其中$\boldsymbol{X}_i=(X_{i,1},\cdots,X_{i,p})^{\mathrm{T}}$,$i=1,\cdots,n$。考虑高维情形,即$p\gg n$,逻辑回归的似然函数为

$$\boldsymbol{L}(\boldsymbol{\beta})=\sum_{i=1}^{n}\{\boldsymbol{Y}_i\boldsymbol{f}_{\boldsymbol{\beta}}(\boldsymbol{X}_i)-\log(1+\exp(\boldsymbol{f}_{\boldsymbol{\beta}}(\boldsymbol{X}_i)))\}$$

损失函数为

$$\ell(\boldsymbol{\beta})=\ell(\boldsymbol{\beta};\boldsymbol{x},\boldsymbol{Y})=$$
$$-\boldsymbol{Y}\boldsymbol{f}_{\boldsymbol{\beta}}(\boldsymbol{x})+\log(1+\exp(\boldsymbol{f}_{\boldsymbol{\beta}}(\boldsymbol{x})))$$

相应的风险函数和经验风险函数分别记为

$$\boldsymbol{P}\ell(\boldsymbol{\beta}_0,\boldsymbol{\beta})=\boldsymbol{E}\ell(\boldsymbol{\beta}_0,\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{Y})$$

$$\boldsymbol{P}_n\ell(\boldsymbol{\beta}_0,\boldsymbol{\beta})=$$

$$\frac{1}{n}\sum_{i=1}^{n}\{-\boldsymbol{Y}_i\boldsymbol{f}_{\boldsymbol{\beta}}(\boldsymbol{X}_i)+\log(1+\exp(\boldsymbol{f}_{\boldsymbol{\beta}}(\boldsymbol{X}_i)))\}$$

考虑$\Xi=\{(\boldsymbol{\beta}_0,\boldsymbol{\beta})\in\boldsymbol{R}^{(p+1)}:\forall\boldsymbol{x}\in[-K,K]^p,\boldsymbol{\beta}_0+\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}\in\Theta\}$,其中$\Theta$为自然空间,显然有,$(\boldsymbol{\beta}_0^*,\boldsymbol{\beta}^*)=\arg\min_{(\boldsymbol{\beta}_0,\boldsymbol{\beta})\in\Xi}P\ell(\boldsymbol{\beta})$。

由文献[12]的定理2.1及文献[13]的条件3.1,有

$$E(\boldsymbol{X}\mid\boldsymbol{Y}=y)=E[E(\boldsymbol{X}\mid\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{X})\mid\boldsymbol{Y}=y]=$$
$$E\left[\left\{\boldsymbol{\mu}+\frac{\boldsymbol{\Sigma}\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\mathrm{T}}(\boldsymbol{X}-\boldsymbol{\mu})}{\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}\right\}\mid\boldsymbol{Y}=y\right]=$$
$$\boldsymbol{\mu}+\frac{\boldsymbol{\Sigma}\boldsymbol{\beta}^*E[\boldsymbol{\beta}^{*\mathrm{T}}(\boldsymbol{X}-\boldsymbol{\mu})\mid\boldsymbol{Y}=y]}{\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}=$$
$$\boldsymbol{\mu}+\boldsymbol{\Sigma}\boldsymbol{\beta}^*k(y)$$

其中,$\boldsymbol{\mu}=E(\boldsymbol{X})$,$\boldsymbol{\Sigma}=\mathrm{Var}(\boldsymbol{X})$,$k(y)=\frac{E[\boldsymbol{\beta}^{*\mathrm{T}}(\boldsymbol{X}-\boldsymbol{\mu})\mid\boldsymbol{Y}=y]}{\boldsymbol{\beta}^{*\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}^*}$。

令$\eta(y)=\boldsymbol{\mu}+\boldsymbol{\Sigma}\boldsymbol{\beta}^*k(y)$则

$$\boldsymbol{\beta}^*\propto\boldsymbol{\Sigma}^{-1}(\eta(y)-\boldsymbol{\mu})\tag{1}$$

令$\boldsymbol{\Sigma}^{-1}=\boldsymbol{\Omega}=(\omega_{ij})_{i,j=1,2,\cdots,p}$,其中$\boldsymbol{\Omega}$为精度矩阵,由式(1)可知,$\boldsymbol{\beta}^*$可表示为$\boldsymbol{\beta}^*=\boldsymbol{\Omega}\boldsymbol{\gamma}$,$\boldsymbol{\gamma}=\eta(y)-\boldsymbol{\mu}$。因此,文献[10]的方法可以推广到逻辑回归模型,于是有

$$\begin{aligned}\beta_1^*&=\gamma_1\omega_{11}+\gamma_2\omega_{12}+\cdots+\gamma_j\omega_{1j}+\cdots+\gamma_p\omega_{1p}\\\beta_2^*&=\gamma_1\omega_{21}+\gamma_2\omega_{22}+\cdots+\gamma_j\omega_{2j}+\cdots+\gamma_p\omega_{2p}\\&\vdots\\\beta_p^*&=\gamma_1\omega_{p1}+\gamma_2\omega_{p2}+\cdots+\gamma_j\omega_{2j}+\cdots+\gamma_p\omega_{pp}\end{aligned}\tag{2}$$

注意到式(2)中,$\boldsymbol{\beta}_j^*$可表示为$\{(\gamma_j\omega_{1j},\gamma_j\omega_{2j},\cdots,\gamma_j\omega_{pj})^{\mathrm{T}}:1\leqslant j\leqslant p\}$ p个部分的和。其中第j部分$(\gamma_j\omega_{1j},\gamma_j\omega_{2j},\cdots,\gamma_j\omega_{pj})^{\mathrm{T}}$有共同因子$\gamma_j$。若$\gamma_j=0$,则$\boldsymbol{\beta}^*$的第$j$部分所有元素将同时为0;若$\gamma_j\neq0$且$\boldsymbol{\Omega}$定义了预测变量的图结构,则$(\gamma_j\omega_{1j},\gamma_j\omega_{2j},\cdots,\gamma_j\omega_{pj})^{\mathrm{T}}$的支撑即为预测变量$\boldsymbol{X}_j$及图结构中与其相连的变量集合,记为$\boldsymbol{N}_j$。上述方法可以扩展到任意图结构中。由于图结构可利用先验信息构造,也可通过数据进行估计,因此,为叙述方便,假设预测变量的图结构已知为G,定义$p\times p$邻接矩阵$\boldsymbol{E}=(E_{ij})_{i,j=1,\cdots,p}$,其中若预测变量$i$和$j$相连,则$E_{ij}=1$;否则,$E_{ij}=0$。令$E_{jj}=1$,$j=1,\cdots,p$,于是,邻域$\boldsymbol{N}_j$可定义为$\boldsymbol{N}_j=\{k:E_{jk}=1\}$。假设$\boldsymbol{\beta}^*$可进行如下分解:

$$\begin{aligned}\beta_1^*&=\Gamma_1^{(1)}E_{11}+\Gamma_1^{(2)}E_{12}+\cdots+\Gamma_1^{(j)}E_{1j}+\cdots+\Gamma_1^{(p)}E_{1p}\\\beta_2^*&=\Gamma_2^{(1)}E_{21}+\Gamma_2^{(2)}E_{22}+\cdots+\Gamma_2^{(j)}E_{2j}+\cdots+\Gamma_2^{(p)}E_{2p}\\&\vdots\\\beta_p^*&=\Gamma_p^{(1)}E_{p1}+\Gamma_p^{(2)}E_{p2}+\cdots+\Gamma_p^{(j)}E_{pj}+\cdots+\Gamma_p^{(p)}E_{pp}\end{aligned}\tag{3}$$

由E_{ij}的定义不难看出,第j部分的候选非零分量$(\Gamma_1^{(j)}E_{1j},\Gamma_2^{(j)}E_{2j},\cdots,\Gamma_p^{(j)}E_{pj})^{\mathrm{T}}$即为$\{\Gamma_k^{(j)}E_{kj}:k\in\boldsymbol{N}_j\}$。注意到每部分中的因子$\{\Gamma_k^{(j)}:k\in\boldsymbol{N}_j\}$可以看作第$j$个预测变量和响应变量之间的边际相关效应。如果它们之间不相关,对于每个$k\in\boldsymbol{N}_j$,$\Gamma_k^{(j)}$都为0,也就是说,集合$\{\Gamma_k^{(j)}E_{kj}:k\in\boldsymbol{N}_j\}$中的每个分量都同时为0。因此,可以将每个邻域$\boldsymbol{N}_1,\boldsymbol{N}_2,\cdots,\boldsymbol{N}_p$

看作一组并用 Group Lasso 惩罚函数进行变量选择[10]。基于上述想法,提出如下的基于预测变量图结构的高维逻辑回归模型:

基于预测变量的图结构 G 确定领域 $\boldsymbol{N}_1,\boldsymbol{N}_2,\cdots,\boldsymbol{N}_p$(注意到 $j\in\boldsymbol{N}_j$, $j\in\boldsymbol{N}_j$,$j=1,\cdots,p$)。

求解如下优化问题:

$$\min_{(\boldsymbol{\beta}_0,\boldsymbol{\beta}),\boldsymbol{\Gamma}^{(1)},\cdots,\boldsymbol{\Gamma}^{(P)}} P_n\ell(\boldsymbol{\beta}_0,\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p} d_j\|\boldsymbol{\Gamma}^{(j)}\|_2 \quad (4)$$

使得 $\sum_{j=1}^{p}\boldsymbol{\Gamma}^{(j)}=\boldsymbol{\beta}$ 且 $supp(\boldsymbol{\Gamma}^{(j)})\subset\boldsymbol{N}_j$, $\forall j=1,\cdots,p$,其中 $supp(\boldsymbol{\Gamma}^{(j)})$ 表示 $\boldsymbol{\Gamma}^{(j)}$ 的支撑集,$\|\cdot\|_2$ 表示 ℓ_2 范数。注意到式(4)中 λ 为调节参数,可以通过交叉验证的方法确定,d_j 是第 j 组的权重参数,d_j 的确定相对比较复杂,将在下一节讨论。

1.2　优化求解及权重参数的选取

求解优化问题式(4),可以利用 Obozinski[13] 等提出的复制变量的方法。具体而言,令 $\boldsymbol{\Gamma}^{(j)}_{N_j}$ 和 $\boldsymbol{X}_{iN_j}$ 表示 $|\boldsymbol{N}_j|\times1$ 的向量,其分量分别为指标在邻域 N_j 中的 $\boldsymbol{\Gamma}^{(j)}$ 和 $\boldsymbol{X}_i$ 的子向量,其中 $i=1,\cdots,n;j=1,\cdots,p$。设 $\widetilde{\boldsymbol{X}}_i=(\boldsymbol{X}^{\mathrm{T}}_{iN_1},\boldsymbol{X}^{\mathrm{T}}_{iN_2},\cdots,\boldsymbol{X}^{\mathrm{T}}_{iN_p})^{\mathrm{T}}$ 及 $\widetilde{\boldsymbol{\Gamma}}=(\boldsymbol{\Gamma}^{(1)\mathrm{T}}_{N_1},\boldsymbol{\Gamma}^{(2)\mathrm{T}}_{N_2},\cdots,\boldsymbol{\Gamma}^{(P)\mathrm{T}}_{N_P})^{\mathrm{T}}$,则容易验证 $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}_i=\widetilde{\boldsymbol{\Gamma}}^{\mathrm{T}}\widetilde{\boldsymbol{X}}_i$。于是,优化问题式(4)等价于普通的 Group Lasso 问题:

$$\min_{(\boldsymbol{\beta}_0,\widetilde{\boldsymbol{\Gamma}})}\frac{1}{n}\sum_{i=1}^{n}[-\boldsymbol{Y}_i(\boldsymbol{\beta}_0+\widetilde{\boldsymbol{\Gamma}}^{\mathrm{T}}\widetilde{\boldsymbol{X}}_i)+\varphi(\boldsymbol{\beta}_0+\widetilde{\boldsymbol{\Gamma}}^{\mathrm{T}}\widetilde{\boldsymbol{X}}_i)]+\lambda\sum_{j=1}^{p}d_j\|\boldsymbol{\Gamma}^{(j)}_{N_j}\|_2 \quad (5)$$

目前,有很多 R 包可以求解优化问题式(5),如 grpLasso[9], grpreg[15] 以及 gglasso[16]。本文利用 Zeng 和 Breheny[9] 提供的 R 包求解优化问题式(5),该方法能够直接求解重叠组结构的 Group Lasso 问题。

对于组结构中有重叠变量的情形,每组权重参数的选取比没有重叠的情形更加重要,也更加复杂。关于此问题,Bach 等[14] 进行了详细的讨论并提出了如何选取权重的方法。他们建议应该选取如下形式的权重参数,即 $d_j=m_j^{\tau}$,其中 $m_j=|\boldsymbol{N}_j|$ 表示邻域 $\boldsymbol{N}_j$ 中预测变量的个数,$\tau\in\left(0,\frac{1}{2}\right)$,当 $\tau=0$ 或 $\tau=\frac{1}{2}$ 时分别对应两种极端的情形,即仅有最大的组或仅有最小的组有可能被选中。本文选取 $\tau=\frac{\log 2}{2\log 3}$,$\lambda$ 为调节参数,利用 10 折交叉验证的方法确定 λ 的值。

相比于目前流行的正则化方法(如 Lasso, Group Lasso 等),本文的方法更具一般性。当预测变量的图结构没有边时,本文方法等价于 Adaptive Lasso;当预测变量的图结构由若干个独立的完全图组成,本文方法等价于 Group Lasso;当预测变量的图结构为完全图时,本文方法与岭回归有相同的非零解[11]。因此,本文的方法是一种更加一般的方法,Adaptive Lasso, Group Lasso 和岭回归都可以看作是本文所提方法的特例。另外,本文的方法也可以用来处理 Overlapping Group Lasso 问题。

2　数值模拟

逻辑回归模型中响应变量 $\boldsymbol{Y}\in\{0,1\}$ 由 $P(\boldsymbol{Y}=1|\boldsymbol{X})=\frac{\exp(\boldsymbol{X}\boldsymbol{\beta}^*)}{1+\exp(\boldsymbol{X}\boldsymbol{\beta}^*)}$ 生成。将数据集 $\boldsymbol{X}$ 分为 3 部分:训练集、验证集和测试集,3 个部分的样本量分别为 120,120,400。训练集用来拟合模型,验证集用来选取调节参数,测试集用来对不同的方法进行比较。预测变量的图结构由训练集中的数据通过 Graphical Lasso 方法[8] 估计得到,模拟结果为 100 次重复的平均值。

例 1　($\boldsymbol{\Omega}$ 分块对角)$p=100$,$s^*=15$,真实的系数向量 $\boldsymbol{\beta}^*=(0.3,0.3,\cdots,0.3,0,0,\cdots,0)^{\mathrm{T}}$,预测变量按如下方式生成:

$$X_j=Z_1+0.4\varepsilon_j,\quad Z_1\sim N(0,1),\quad 1\leqslant j\leqslant5$$
$$X_j=Z_2+0.4\varepsilon_j,\quad Z_2\sim N(0,1),\quad 6\leqslant j\leqslant10$$
$$X_j=Z_3+0.4\varepsilon_j,\quad Z_3\sim N(0,1),\quad 11\leqslant j\leqslant15$$
$$X_j\sim N(0,1),\quad 16\leqslant j\leqslant100$$

其中 $\varepsilon_j\overset{\text{iid}}{\sim}N(0,1)$,$j=1,2,\cdots,15$。

例 2　($\boldsymbol{\Omega}$ 带状)$p=100$,$\boldsymbol{\beta}^*$ 与例 1 相同,预测变量 $(X_1,X_2,\cdots,X_p)^{\mathrm{T}}\sim N(0,\Sigma)$,其中 $\Sigma_{ij}=0.5^{|i-j|}$。此时,若 $|i-j|=1$,有 $\omega_{ii}=1.333$,$\omega_{ij}=-0.677$;若 $|i-j|>1$,则 $\omega_{ij}=0$。

例 3　($\boldsymbol{\Omega}$ 稀疏)$p=100$,预测变量 $(X_1,X_2,\cdots,X_p)^{\mathrm{T}}\sim N(0,\boldsymbol{\Omega}^{-1})$,其中 $\boldsymbol{\Omega}^{-1}=\boldsymbol{B}+\delta\boldsymbol{I}$。$\boldsymbol{B}$

中非对角线元素以概率0.05取0.5,以概率0.95取0,对角线元素为0。选取δ使得$\boldsymbol{\Omega}$的条件数为p。令$\boldsymbol{\beta}^*=\boldsymbol{\Omega\gamma}$,其中$\boldsymbol{\gamma}=(\gamma_1,\gamma_2,\cdots,\gamma_p)^{\mathrm{T}}$,且$\gamma_i=0.1$,$i=1,2,3,4$,否则,$\gamma_i=0$。

选取如下指标对不同方法进行比较:l_2距离(l_2 distance):$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*\|_2$;预测误差(Prediction error):$\frac{1}{N_{\text{test}}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)^{\mathrm{T}}\boldsymbol{X}_{\text{test}}^{\mathrm{T}}\boldsymbol{X}_{\text{test}}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)$,其中$\boldsymbol{X}_{\text{test}}$为测试集中的样本,$N_{\text{test}}$为测试集的样本量;$R_{\text{FPR}}$为将不重要变量错误的识别为重要变量的比率$R_{\text{FNR}}$为将重要变量错误的识别为不重要变量的比率;非零匹配率(R_{NMR}):$R_{\text{NMR}}=\frac{|\{(i,j):\Omega_{ij}\neq0,\hat{\beta}_i\neq0,\hat{\beta}_j\neq0\}|}{|\{(i,j):\Omega_{ij}\neq0,\beta_i^*\neq0,\beta_j^*\neq0\}|}$;零匹配率($R_{\text{ZMR}}$):$R_{\text{ZMR}}=\frac{|\{(i,j):\Omega_{ij}\neq0,\hat{\beta}_i=0,\hat{\beta}_j=0\}|}{|\{(i,j):\Omega_{ij}\neq0,\beta_i^*=0,\beta_j^*=0\}|}$,其中$R_{\text{NMR}}$(或$R_{\text{ZMR}}$)是用来检验图模型中与重要(或不重要)的预测变量相连的变量的估计值是否同时非零(或者同时为零)。

表1—表3给出了不同方法在不同的图结构中的表现,其中l_2距离和预测误差衡量的是模型的估计和预测能力,R_{FPR}和R_{FNR}衡量是模型的变量选择能力。LG-O和LG表示本文所提方法分别利用真实图结构和估计图结构得到的结果。表1—表3中的结果表明:当图结构比较简单时,本文方法与其他方法相当(表1所示);当图结构比较复杂时,本文方法无论从估计和预测方面还是从模型选择方面都明显优于其他方法,与具有Oracle性质的SCAD和MCP方法比较仍具有优越性(表2,表3所示)。

表4给出了不同方法R_{NMR}和R_{ZMR}的表现。表4结果表明:即使图结构比较简单时,其他方法的R_{NMR}和R_{ZMR}也很差。本文的方法在图结构较简单时明显优于其他方法,当图结构较复杂时,也要优于其他方法,这表明本文方法有效地利用了预测变量图结构的信息。

表1 例1中不同方法在模型估计、预测和模型选择能力的比较

Table 1 Comparison of different methods in estimation, prediction and model selection for Example 1

方 法	l_2 distance	Prediction error	R_{FPR}	R_{FNR}
Lasso	1.977(0.172)	7.810(2.951)	0.282(0.013)	0.468(0.010)
Ridge	1.372(0.039)	2.535(0.102)	1.000(0.000)	0.000 (0.000)
Alasso	2.614(0.245)	13.450(4.519)	0.194(0.010)	0.533(0.011)
Enet	1.543(0.103)	4.203(1.047)	0.421(0.015)	0.191(0.010)
SCAD	1.683(0.059)	1.873(0.162)	0.074(0.005)	0.731(0.012)
MCP	1.678(0.035)	1.603(0.051)	0.036(0.003)	0.801(0.001)
LG-O	0.836(0.026)	0.924(0.039)	0.156(0.008)	0.000(0.000)
LG	0.836(0.026)	0.924(0.039)	0.156(0.008)	0.000(0.000)

表2 例2中不同方法在模型估计、预测和模型选择能力的比较

Table 2 Comparison of different methods in estimation, prediction and model selection for Example 2

方 法	l_2 distance	Prediction error	R_{FPR}	R_{FNR}
Lasso	2.003(0.229)	10.101(3.446)	0.294(0.013)	0.311(0.011)
Ridge	1.421(0.042)	2.363(0.123)	1.000(0.000)	0.000(0.000)
Alasso	2.439(0.296)	16.050(5.674)	0.195(0.011)	0.395(0.011)
Enet	1.764(0.163)	6.091(1.947)	0.409(0.016)	0.193(0.010)
SCAD	5.859(2.583)	700.7(468.2)	0.163(0.005)	0.466(0.012)
MCP	6.958(2.489)	795.3(497.3)	0.096(0.004)	0.598(0.008)
LG-O	1.269(0.055)	2.075(0.162)	0.485(0.017)	0.109(0.008)
LG	1.301(0.050)	2.004(0.128)	0.437(0.015)	0.187(0.010)

表 3　例 3 中不同方法在模型估计、预测和模型选择能力的比较

Table 3　Comparison of different methods in estimation, prediction and model selection for Example 3

方　法	l_2 distance	Prediction error	R_{FPR}	R_{FNR}
Lasso	0.382(0.032)	0.297(0.082)	0.083(0.013)	0.911(0.015)
Ridge	0.338(0.000)	0.208(0.000)	1.000(0.000)	0.000(0.000)
Alasso	0.436(0.062)	0.654(0.369)	0.055(0.011)	0.937(0.012)
Enet	0.539(0.126)	2.104(1.371)	0.118(0.020)	0.871(0.021)
SCAD	1.020(0.561)	49.65(48.76)	0.049(0.007)	0.948(0.008)
MCP	0.802(0.309)	15.38(14.19)	0.031(0.004)	0.964(0.005)
LG-O	0.380(0.032)	0.317(0.078)	0.278(0.030)	0.743(0.031)
LG	0.370(0.042)	0.390(0.160)	0.337(0.031)	0.720(0.028)

表 4　不同模型的 NMR 和 ZMR 的比较

Table 4　Comparison of different models in R_{NMR} and R_{ZMR}

方　法	R_{NMR}			R_{ZMR}		
	例 1	例 2	例 3	例 1 *	例 2	例 3
Lasso	0.258(0.011)	0.460(0.017)	0.029(0.008)	…	0.522(0.019)	0.865(0.020)
Ridge	1.000(0.000)	1.000(0.000)	1.000(0.000)	…	0.000(0.000)	0.000(0.000)
Alasso	0.193(0.011)	0.349(0.015)	0.020(0.006)	…	0.680(0.018)	0.913(0.016)
Enet	0.648(0.017)	0.661(0.017)	0.056(0.016)	…	0.381(0.019)	0.823(0.026)
SCAD	0.057(0.010)	0.251(0.016)	0.007(0.002)		0.732(0.010)	0.915(0.012)
MCP	0.000(0.000)	0.085(0.007)	0.003(0.001)		0.865(0.007)	0.943(0.008)
LG-O	1.000(0.000)	0.826(0.011)	0.168(0.027)	…	0.417(0.019)	0.610(0.037)
LG	1.000(0.000)	0.681(0.016)	0.216(0.024)	…	0.445(0.018)	0.533(0.041)

* 注:…表示不可用值,没有用的预测变量之间没有相连的边。

以上模拟结果表明:即使图结构未知,本文方法仍然能够有效地利用预测变量图结构信息,从而提高模型在估计、预测以及变量选择等方面的表现。

3　实例分析与应用

本节将上述方法应用于公开的乳腺癌实际数据,包括 133 名受试者 22 283 个基因表达水平,其中 34 名为病理完全反应(pCR)受试者,99 名为残留病(RD)受试者。该数据可以通过网址 http://Bioinformatics. mdanderson. org/pubdata. html 下载。考虑到迭代算法运行的速度以及计算机的限制,将主要比较本文的方法与 Lasso 方法、岭回归方法、Adaptive Lasso 方法以及 Elastic Net 方法的效果。

为了估计精度矩阵,将数据集和测试集分别随机划分为大小为 112 和 21 的两部分,然后将整个过程重复 100 次。每次都采用分层抽样的方法从对应的组中选取 5 个 pCR 个体和 16 个 RD 个体作为测试集(大体相当于每组的 1/6),其余的个体作为训练集。在每个训练集上,利用两样本 t 检验选取最显著的 113 个基因作为预测变量。注意到,训练集的样本量为 $n=112$,比变量的维度 $p=113$ 稍小,这一点可以用来检验 $p>n$ 时各种方法的表现。利用训练集估计精度矩阵 $\boldsymbol{\Omega}$,基于该精度矩阵利用 Graphical Lasso 估计预测变量的图结构 G,其图结构如图 1 所示,图中包含 113 个节点,190 条边。训练

集用来拟合模型。利用测试集数据计算均方误差值(F_{MSE})来评估模型,利用10折交叉验证选取调节参数[7,17]。

实例分析结果如图2所示,图2给出了不同方法在对乳腺癌数据进行建模分析时的平均均方误差箱线图。结果可以看出本文的方法在MSE的表现上优于Lasso方法、Adaptive Lasso和Elastic Net方法,比岭回归的结果稍差。实际数据分析结果与数值模拟结果一致,说明该方法有效。

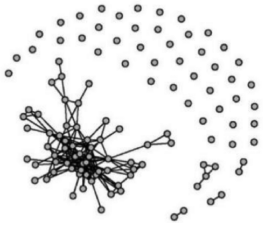

图1 乳腺癌数据的图结构

Fig. 1 Graphical structure of breast cancer data

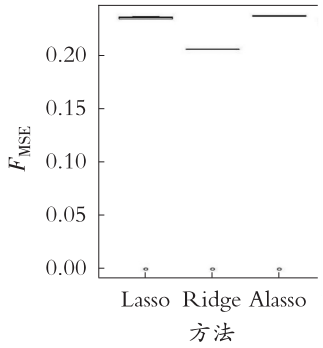

图2 乳腺癌数据集中不同方法的F_{MSE}

Fig. 2 Comparison of F_{MSE} for different methods on breast cancer data set

4 结论与讨论

提出一种基于预测变量图结构的高维逻辑回归模型,该模型可以用来对高维图结构数据或者重叠组结构数据进行逻辑回归建模,即使预测变量的图结构未知,本研究的方法仍然能够利用这种结构提高模型在估计、预测以及变量选择等方面的表现。另外,目前流行的方法,如Lasso方法、Group Lasso和岭回归方法等都可以看作是本文模型的特例。数值模拟和实例分析应用表明:该模型在有限样本情形下是有效的,并且可广泛应用于高通量基因数据中存在图结构的疾病分类研究中。

参考文献(References):

[1] HASTIE T, TIBSHIRANI R, FRIEDMAN J. The Elements of Statistical learning[J]. New York: Springer Series in Statistics, 2001,10(1):

[2] TIBSHIRANI R. Regression Shrinkage and Selection via the Lasso[J]. Journal of the Royal Statistical Society. 1996, 58(1): 267—288

[3] YUAN M, LIN Y. Model Selection and Estimation in Regression with Grouped Variables[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology),2006, 68(1): 49—67

[4] JACOB L, OBOZINSKI G, Vert J P. Group Lasso with Overlap and Graph Lasso[C]// Proceedings of the 26th Annual International Conference on Machine Learning. ACM, 2009: 433-440

[5] YUAN M, LIN Y. Model Selection and Estimation in the Gaussian Graphical Model [J]. Biometrika, 2007, 94(1): 19—35

[6] FRIEDMAN J, HASTIE T, TIBSHIRANI R. Sparse Inverse Covariance Estimation with the Graphical Lasso[J]. Biostatistics, 2008, 9(3): 432-441

[7] CAI T, LIU W, LUO X. A Constrained L_1 Minimization Approach to Sparse Precision Matrix Estimation [J]. Journal of the American Statistical Association, 2011, 106(494): 594—607

[8] YU G, LIU Y. Sparse Regression Incorporating Graphical Structure Among Predictors[J]. Journal of the American Statistical Association, 2016, 111(514): 707—720

[9] SHEVADE S K, KEERTHI S S. A Simple and Efficient Algorithm for Gene Selection Using Sparse Logical Regression[J]. Bioinformatics, 2003, 19(17): 2246—2253

[10] MEIER L, VAN DE GEER S, BJHLMANN P. The Group Lasso for Logical Regression[J]. Journal of the Royal Statistical Society: Series B (Statistical

Methodology), 2008, 70(1):53-71

[11] ZENG Y, BREHENY P. Overlapping Group Logical Regression with Applications to Genetic Pathway Selection[J]. Cancer Informatics, 2016, 15:179—187

[12] DUAN N, LI K C. Slicing Regression: A Link-free Regression Method[J]. The Annals of Statistics, 1991, 19(2): 505—530

[13] LI K C. Sliced Inverse Regression for Dimension Reduction[J]. Journal of the American Statistical Association, 1991, 86(414): 316—327

[14] BACH F, JENATTON R, MAIRAL J, et al. Structured Sparsity Through Convex Optomization [J]. Statistical Science, 2012, 27(4):450—468

[15] BREHENY P, HUANG J. Penalized Methods for Bi-level Variable Selection[J]. Statistics and Its Interface, 2009, 2(3): 369—380

[16] YANG Y, ZOU H. A Fast Unified Algorithm for Computing Group-lasso Penalized Learning Problems[J]. Statistics and Computing, 2015, 25(6): 1129—1141

[17] FAN J, FENG Y, WU Y. Network Exploration via the Adaptive LASSO And SCAD Penalties[J]. The Annals of Applied Statistics, 2009, 3(2): 521-541

High-dimensional Logic Regression Model Based on Graph Structure of Predictive Variables

HUANG Wen-jing, DENG Dan, DU Jie-lin, WU Ming-yue

(School of Public Health and Management, Chongqing Medical University, Chongqing 400016, China)

Abstract: For high-dimensional data sets, we propose a method to improve sparse logic regression model by using the graph structure information between predictive variables. In this method, logic regression modeling is carried out by using high-dimensional graph structure data or overlapping group structure, it is still applicable even if the graph structure of predictive variables is unknown. When graph structure is some special forms, all current popular methods such as Adaptive Lasso, (Overlapping) Group Lasso and ridge regression can be regarded as special cases of this method. Numerical simulation and real data analysis show that the proposed method can effectively use the graph structure information of predictive variables to improve the performance of the model in estimation, prediction, variable selection and so on. Moreover, the model is effective in the case of limited samples and overcomes the problem of the dimensionality of data sets, improves the performance of the sparse logic regression model by using the graph structure of high-dimensional data, and can be widely used in disease classification of high-throughput gene data sets.

Key words: logic regression; high-dimensional data; graph structure; Lasso; sparseness

责任编辑:李翠薇

引用本文/Cite this paper:

黄文静,邓丹,杜杰琳,等. 基于预测变量图结构的高维逻辑回归模型[J]. 重庆工商大学学报(自然科学版),2021,38(5):107—113

HUANG W J, DENG D, DU J L, et al. High-dimensional Logic Regression Model Based on Graph Structure of Predictive Variables[J]. Journal of Chongqing Technology and Business University (Natural Science Edition),2021,38(5):107—113