计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(12):37−45 [doi: 10.15888/j.cnki.csa.008195]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于自主学习与 SCAD-Net 正则化的回归模型①

刘　杰[1,2], 陈浩杰[1,2]

[1](中国科学技术大学 管理学院, 合肥 230026)
[2](中国科学技术大学 国际金融研究院, 合肥 230026)
通讯作者: 陈浩杰, E-mail: chenhj68@mail.ustc.edu.cn

摘　要: 众多基因生物标志物选择方法常因研究样本较少而不能直接用于临床诊断. 于是有学者提出整合不同基因表达数据同时保留生物信息完整性的方法. 然而, 由于存在批量效应, 导致直接整合不同基因表达数据可能会增加新的系统误差. 针对上述问题, 提出一个融合自主学习与 SCAD-Net 正则化的分析框架. 一方面, 自主学习方法能够先从低噪声样本中学习出基础模型, 然后再通过高噪声样本学习使得模型更加稳健, 从而避免批量效应; 另一方面, SCAD-Net 正则化融合了基因表达数据与基因间的交互信息, 可以实现更好的特征选择效果. 不同情形下的模拟数据以及在乳腺癌细胞系数据集上的结果表明, 基于自主学习与 SCAD-Net 正则化的回归模型在处理高维复杂网络数据集时具有更好的预测效果.

关键词: 自主学习; 图正则化; 变量选择; 基因表达; 回归

引用格式: 刘杰,陈浩杰.基于自主学习与 SCAD-Net 正则化的回归模型.计算机系统应用,2021,30(12):37–45. http://www.c-s-a.org.cn/1003-3254/8195.html

Regression Model with Self-Paced Learning and SCAD-Net Regularization

LIU Jie[1,2], CHEN Hao-Jie[1,2]

[1](School of Management, University of Science and Technology of China, Hefei 230026, China)
[2](International Institute of Finance, University of Science and Technology of China, Hefei 230026, China)

Abstract: Many methods for gene biomarker selection can not be directly used in clinical diagnosis because of a small number of research samples. Therefore, some scholars proposed methods of integrating different gene expression data while preserving the integrity of biological information. However, due to the batch effect, direct integration of different gene expression data may bring new systematic errors. In response to the above problems, an analysis framework integrating self-paced learning and SCAD-Net regularization is proposed. On the one hand, self-paced learning can learn the basic model from low-noise samples and then make the model more robust through high-noise samples to avoid batch effect. On the other hand, SCAD-Net regularization combines biological interaction information and gene expression data, which can achieve a better performance in feature selection. The simulation data in different cases and the results on the breast cancer cell line dataset show that the regression model based on self-paced learning and SCAD-Net regularization obtains better prediction results when dealing with high-dimensional complex network datasets.

Key words: self-paced learning; graph regularization; variable selection; gene expression; regression

① 基金项目: 国家自然科学基金 (71771201, 71874171, 71731010, 71631006, 71991464)
Foundation item: National Natural Science Foundation of China (71771201, 71874171, 71731010, 71631006, 71991464)
收稿时间: 2021-02-21; 修改时间: 2021-03-19; 采用时间: 2021-03-26

基因组学研究的一个关键问题是如何确定与疾病相关的基因及其生物途径, 常见的做法是通过将高维基因组数据 (如微阵列基因表达数据) 与各种临床结果联系起来构建疾病诊断预测模型. 然而, 迄今为止, 虽然许多基因生物标志物研究[1,2] 已经完成, 但目前提出的众多相关方法在临床应用中均难以得到令人满意的结果. 其原因主要在于研究样本量太小[3,4], 从而导致统计效能降低, 进而得到可信度较低甚至错误的结论. 因此, 充足的样本是产生有效统计分析和结论的必要条件. 另一方面, 数据收集技术的进步促使现行可用生物数据日益增多, 于是有学者提出了数据融合的思想, 即综合多个数据集或有关结果. 然而, 尽管一些基因表达研究有着相同的目标, 但所用数据集通常是来自不同的处理设备、不同的数据平台, 甚至彼此之间具有不同的数值尺度, 从而导致批量效应的存在. 因此, 直接整合不同的基因表达数据将会给统计分析带来巨大挑战.

为解决上述问题, 研究者们做了大量的工作, 主要分为以下两类: 元分析和融合分析[5] 方法. 元分析即利用统计的概念与方法去收集、整理以及分析之前学者针对某个主题所做的众多实证研究. 然而, 元分析对一些必要条件较为敏感, 稍加违反就可能造成错误性结论[6]. 融合分析是对不同的数据集进行整合并以此作为研究数据集. 相比元分析, 融合分析具有更多的样本从而更具统计效用. 近年来, 基于融合分析的方法层出不穷, 如 Benito 等[7] 提出的距离加权判别法 (DWD), Johnson 等[8] 提出的经验贝叶斯方法 (EB), Shabalin 等[9] 提出的跨平台标准化方法 (XPN), Deshwar等[10] 提出的 PLIDA 方法以及 Deng 等[11] 提出的 WaveICA 方法. 然而, 由于批量效应的存在, 且其来源复杂无法消除, 导致以上方法均可能给融合数据集带来新的系统误差, 使其变得更加复杂. 因此, 直接分析融合后的数据可能会引起一些问题[12,13], 需要提出一种新的方法来解决数据融合问题.

Kumar 等[14] 提出的自主学习 (Self-Paced Learning, SPL) 方法可以根据模型已经学习的内容自适应地识别简单和困难样本, 并且随着模型训练的不断进行, 越来越多的困难样本进入模型. SPL 方法可以在很大程度上克服批量效应, 并且其应用较为广泛, 目前已成功应用于各种机器学习问题[15]. 此外, Ma 等[16] 还对 SPL 方法的收敛性质进行了补充和讨论, 使其在理论上更加丰富.

除样本规模问题之外, 样本维度是另一研究热点. 许多研究中的样本维数通常远远大于样本数量, 即常见的高维度低样本问题. 这在生物统计中尤为常见, 如基因表达数据. 为解决该问题, 研究者们提出了许多正则化方法, 用于在回归框架中识别与临床表型相关的基因, 如 Lasso[17]、SCAD[18]、Elastic-Net [19]、Fused Lasso[20]、Lars[21]、adaptive Lasso[22]、Group Lasso[23] 以及$L_{1/2+2}$混合正则化方法[24–26]. 然而, 以上正则化方法都存在共同的局限性, 即这些方法仅是从计算或算法的角度出发, 没有利用任何先验知识或信息. 但对于许多复杂的疾病尤其是癌症, 许多生物学途径信息对于了解治疗疾病具有较大的效用, 并且该信息可以从多年的生物医学研究中获得, 故将此种先验信息纳入模型考虑应该会有更好的预测效果.

本文将基于 SPL 方法构建一个更精确的基因表达预测模型. 首先我们将不同的基因表达数据集融合到一个统一的数据集中, 紧接着在线性回归的背景下将 SPL 方法与 SCAD 网络惩罚相结合得到最终的回归预测模型. 具体来说, 该模型由 3 部分组成: (1) SCAD 罚函数. 利用 SCAD 惩罚来增强模型的稀疏性, 该惩罚不仅为大系数提供了无偏估计, 并且具有较高的理论价值, 例如 Oracle 性质[18]; (2) 基于网络的惩罚, 利用网络惩罚来实现基因调控网络上相邻节点系数之间的平滑; (3) SPL 方法, 促使模型自适应地从简单样本 (高置信度样本) 向复杂样本 (低置信度样本) 上过渡. SPL 方法对于分析融合数据是至关重要的, 因为融合数据往往存在较大的噪声以及异常值点.

本文接下来内容安排如下: 第 1 节提出了一个基于 SCAD 网络惩罚的线性回归模型, 紧接着介绍了自主学习 (SPL) 方法并将其与 SCAD 网络惩罚相结合从而得到最终的预测模型; 第 2 节首先对 SCAD 网络惩罚函数的理论性质进行简单分析, 包括群组效应以及渐近性质; 然后给出一种求解本文所提出模型的有效算法; 在第 3 节中, 通过不同情形下的模拟数据以及在乳腺癌细胞系数据集上的分析结果来评估本文所提出模型的预测效果. 第 4 节是结论与展望.

1 SCAD 网络正则化与自主学习方法

1.1 SCAD 网络正则化

假设数据集$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$, 其中$x_i=(x_{i1},x_{i2},\cdots,x_{ip})^{\mathrm{T}}$表示第$i$个样本, y_i为对应的响应变

量, 记$X=(x_1,x_2,\cdots,x_p), Y=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$. 进一步, 假设各个预测因子$x_i$ $(i=1,2,\cdots,p)$经过标准化处理, 响应变量y经过去中心化处理, 从而有:

$$\sum_{i=1}^{n} y_i=0, \sum_{i=1}^{n} x_{ij}=0, \sum_{i=1}^{n} {x_{ij}}^2=1,\ j=1,2,\cdots,p$$

本文考虑最简单的线性回归模型:

$$y_i={x_i}^{\mathrm{T}}\beta+\varepsilon_i$$

式中, $\beta=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$ 为要估计的参数, ε_i表示均值为 0, 方差为σ^2的误差项. 上述模型的平方损失函数可以表示为:

$$l(\beta)=\sum_{i=1}^{n}(y_i-{x_i}^{\mathrm{T}}\beta)^2$$

在许多研究当中, 样本维度通常远远大于样本数量, 即高维度低样本问题. 在这种情形下, 线性回归并不能够直接用来估计回归参数. 由此, 引入了正则化方法, 即:

$$\ell(\beta,\lambda)=l(\beta)+P(\beta)$$

其中, $P(\beta)$表示正则化项. 高维变量选择中常用的正则化方法为L_1约束, 即 Lasso 方法, 具体可以表示为$P_{\lambda,\mathrm{Lasso}}(\beta)=\sum_{i=1}^{p}\left|\beta_j\right|_1$, 这里$\lambda$表示任意非负数, 一般可使用$k$折交叉验证方法确定. 由于$L_1$罚函数具有奇异性, 故基于$L_1$惩罚的线性回归模型可以将一些系数较小的参数压缩为 0 从而达到变量选择的效果. 但当λ过大时, β估计量中系数较大的参数会存在较大偏差, 而当λ过小时, β估计量则不够稀疏. 为克服这一问题, Fan 等[18] 提出了 SCAD 惩罚函数, 其具体形式为:

$$P_\lambda(\beta)=\begin{cases}\lambda|\beta|, & |\beta|\le\lambda\\ \dfrac{(a^2-1)\lambda^2-(|\beta|-a\lambda)^2}{2(a-1)}, & \lambda<|\beta|\le a\lambda\\ \dfrac{(a+1)\lambda^2}{2}, & |\beta|>a\lambda\end{cases} \tag{1}$$

其中, a为一个大于 2 的超参数, 根据文献 [18], 可将其设为 3.7, 当然也可以通过交叉验证的方法加以确定. 从表达式 (1) 可以发现, 当$|\beta|$较小时, 惩罚函数为线性函数; 当$|\beta|$较大时, 惩罚函数为二次惩罚; 当$|\beta|$很大时, 惩罚项为常数. SCAD 惩罚函数关于β的一阶导函数为:

$$\begin{aligned}P_\lambda'(\beta)&=\begin{cases}\lambda, & |\beta|\le\lambda\\ \dfrac{a\lambda-\beta}{a-1}, & \lambda<|\beta|\le a\lambda\\ 0, & |\beta|>a\lambda\end{cases}\\ &=\lambda\{I(\beta\le\lambda)+\frac{(a\lambda-\beta)_+}{(a-1)\lambda}I(\beta>\lambda)\}\end{aligned} \tag{2}$$

在非 0 处, 对任意$\beta_j\approx z_j$, 由二阶泰勒展开可得:

$$P_\lambda(\beta_j)\approx P_\lambda(|z_j|)+\frac{1}{2}\left\{\frac{P_\lambda'(|z_j|)}{|z_j|}\right\}(\beta_j^2-z_j^2) \tag{3}$$

此外, 对于协变量之间存在高相关性的问题, Zou 等[19] 提出了 Elastic-Net 惩罚函数, 其具体表达为$P_{\lambda_1,\lambda_2,\mathrm{enet}}(\beta)=\sum\lambda_1\left|\beta_j\right|_1+\lambda_2\left|\beta_j\right|_2$. Zeng 等[27] 提出了 SCAD-$L_2$惩罚, 该惩罚同时结合 SCAD 和$L_2$惩罚. 以上几种方法都可以实现群组效应, 即具有强相关性的预测因子要么同时被选中, 要么同时被剔除. 然而, 上述几种方法都是从计算的角度出发, 都没有考虑先验信息, 如网络拓扑信息.

我们知道, 生物统计中的基因交互信息对于识别基因组模式具有重要价值. 该先验信息可以用一个加权图$G=(N,E,W)$来表示, 其中N是网络节点集, 表示p个预测因子, $E=\{u\sim v\}$是图中边的集合, 表示节点u和v之间有边相连, $W=\{w(u,v)\}$则表示边上的权重. 近年来, 网络惩罚出现在大量现实应用中, 例如, Li 等[28], Chen 等[29] 以及 Wang 等[30] 利用基于网络的L_1惩罚对基因组数据进行回归分析并进行变量选择. 在这些研究当中, 网络惩罚函数被定义为拉普拉斯矩阵的二次型. 然而, 在某些情况下, L_1惩罚存在偏差并且可能导致结果不够稀疏, 而 SCAD 惩罚可以避免过度惩罚并且具有良好的统计性质. 因此, 本文给出基于网络的 SCAD-Net 惩罚函数 (SCAD Network-based penalized fuction, SCAD-Net):

$$P_{\lambda_1,\lambda_2,\text{scad-net}}(\beta)=P_{\lambda_1,\mathrm{scad}}(\beta)+\lambda_2\beta^{\mathrm{T}}L\beta \tag{4}$$

其中, L 表示拉普拉斯矩阵, 根据文献 [31], 可将其定义为:

$$L(u,v)=\begin{cases}1-w(u,v)/d_u, & u=v, d_u\ne 0\\ -w(u,v)/\sqrt{d_u d_v}, & u,v\text{相邻接}\\ 0, & \text{其他}\end{cases}$$

其中, $d_u=\sum_{u\sim v}w(u,v)$表示节点u的度, $d_u=0$说明节点u是孤立的点. L 融合了网络的结构信息, 且$\beta^{\mathrm{T}}L\beta$项使得β的估计更加光滑. 进一步, 根据 L 的性质, 表达式 (4) 可以写成:

$$P_{\lambda_1,\lambda_2,\text{scad-net}}(\beta)=P_{\lambda_1,\mathrm{scad}}(\beta)+\lambda_2\sum_{i\sim j}\left(\frac{\beta_i}{\sqrt{d_i}}-\frac{\beta_j}{\sqrt{d_j}}\right)^2 w(i,j) \tag{5}$$

其中, 参数λ_1与λ_2 分别控制参数估计的稀疏度与光滑度. 进一步, 在线性回归背景下, 可以得到基于 SCAD-Net 惩罚的线性回归模型 (SCAD-Net penalized Linear regression, SNL):

$$\hat{\beta}=\min_{\beta}\left\{l(\beta)+P_{\lambda_1,\text{scad}}(\beta)+\lambda_2\sum_{i\sim j}\left(\frac{\beta_i}{\sqrt{d_i}}-\frac{\beta_j}{\sqrt{d_j}}\right)^2 w(i,j)\right\} \tag{6}$$

其中, 第 1 项表示线性回归的损失函数; 第 2 项表示 SCAD 惩罚函数, 保证参数估计的稀疏性, 并且强化结果的可解释性; 最后一项表示基于网络的惩罚函数, 保证参数估计的光滑性, 并且将网络结构信息与基因表达数据相融合.

1.2 自主学习策略

尽管正则化方法在基因数据分析, 变量选择等方面扮演非常重要的角色, 但最终得到的结论却鲜少在临床中得到应用. 这是因为上述结论都是基于小样本数据, 导致结果的可信度较低. 为解决这一问题, 有人提出通过整合不同的数据集来生成人工大样本数据. 然而, 这些数据整合的方法并不能消除内部偏差, 甚至可能会增加新的误差.

受人类学习机制的启发, Kumar 等[14] 提出了自主学习 (SPL) 方法, 该方法首先通过低噪声样本学习一个基础模型, 然后通过高噪声样本学习使模型变得更加稳健, 该方法可以显著提高融合数据集的统计分析效用. 并且 Kumar 表示, 通过引入一个惩罚项, 可以将自主学习方法视为优化模型, 具体可以表示为:

$$\ell(\beta,v)=\sum_{i=1}^{n}\left\{v_i l_i(y_i,d(x_i,\beta))+f(v_i,\tau)\right\} \tag{7}$$

其中, $\ell(\beta,v)=\sum_{i=1}^{n}\left\{v_i l_i(y_i,d(x_i,\beta))+f(v_i,\tau)\right\}$为每个样本的加权损失, $l_i(y_i,d(x_i,\beta)$表示一般损失函数, $d(x_i,\beta)$ 为决策函数, $v=(v_1,v_2,\cdots,v_n)^{\mathrm{T}}$ 为样本集上的加权向量, τ 为超参数, 用来调整学习步长. $f(v_i,\tau)$ 表示施加在样本权重上的惩罚, 一般情况下, 我们可以取$f(v,\tau)=-v\tau$. 显然, 当τ 较小时, 为使式 (7) 最小, 训练集中只能包括少量的样本, 而当τ增大时, 训练集中又必须加入更多的样本, 样本由少变多, 模型从简单到复杂.

为了加强对融合数据分析的准确性与鲁棒性, 本文将 SPL 方法与 SCAD-Net 正则化在线性回归的背景下相结合, 从而得到最终的回归模型 (Self-paced learning and SCAD-Net penalized Linear regression, SSNL):

$$\begin{aligned}\ell(\beta,v)&=\frac{1}{2}\sum_{i=1}^{n}(v_i l_i(\beta)-v_i\tau)+\sum_{j=1}^{p}P_{\lambda_1}(|\beta_j|)+\frac{\lambda_2}{2}\beta^{\mathrm{T}}L\beta\\&=\frac{1}{2}\sum_{i=1}^{n}(v_i(y_i-x_i{}^{\mathrm{T}}\beta)^2-v_i\tau)+\sum_{j=1}^{p}P_{\lambda_1}(|\beta_j|)+\frac{\lambda_2}{2}\beta^{\mathrm{T}}L\beta\end{aligned} \tag{8}$$

其中, 第 1 项表示加权的线性回归模型, 最后两项表示 SCAD-Net 惩罚函数.

2 理论性质及求解方法

2.1 理论性质

本小节我们给出与 SCAD-Net 正则化方法相关的性质, 包括群组效应以及在p固定且 $n\to\infty$ 情形下的渐近性质.

2.1.1 群组效应

Huang 等[32] 证明了 SCAD-Net 惩罚函数具有群组效应, 如引理 1 与引理 2 所示, 其具体证明过程见文献 [32].

引理 1. 若$\hat{\beta}$是由表达式 (6) 计算得到, 对任意 $\lambda_2>\frac{1}{2(a-1)}$, 若满足$x_i=x_j$, 则有:

$$\hat{\beta}_i=\hat{\beta}_j$$

引理 1 可看作 Zou 等[19] 中引理 2 的进一步结果, 其保证在两个预测变量相等时, 估计参数具有群组效应.

引理 2. 若$\hat{\beta}_i\hat{\beta}_j>0$且$\lambda_2>\frac{1}{2(a-1)}$, 定义:

$$D(i,j)=\frac{|\hat{\beta}_i-\hat{\beta}_j|}{|y|_1}$$

则有:

$$D(i,j)\le\frac{1}{2\lambda_2-\frac{1}{a-1}}\sqrt{2(1-\rho)}$$

其中, $\rho=x_i^{\mathrm{T}}x_j$表示样本相关系数.

引理 2 给出 SCAD-Net 惩罚函数群组效应的量化描述, 即在满足以上条件的前提下, 两个参数的差异具有上界约束. 进一步, 若样本相关系数ρ趋于 1, 则两个估计参数几乎相同.

2.1.2 渐近性

SCAD-Net 惩罚线性回归的目标函数为:

$$\begin{aligned}&\sum_{i=1}^{n}(y_i-x_i{}^{\mathrm{T}}\beta)^2+\sum_{j=1}^{p}P_{\lambda_n^{(1)}}(|\beta_j|)+\frac{\lambda_n^{(2)}}{2}\beta^{\mathrm{T}}L\beta\\&=\sum_{i=1}^{n}(y_i-x_i{}^{\mathrm{T}}\beta)^2+\sum_{j=1}^{p}P_{\lambda_n^{(1)}}(|\beta_j|)\\&\quad+\lambda_n^{(2)}\sum_{u\sim v}\left(\frac{\beta_u}{\sqrt{d_u}}-\frac{\beta_v}{\sqrt{d_v}}\right)^2 w(u,v)\end{aligned}$$

其中, $\lambda_n^{(1)}$与$\lambda_n^{(2)}$ 为样本大小 n的函数. 定理 1 给出了估

计量的渐近性质.

定理 1. 若 $\lambda_n^{(m)}/\sqrt{n}\to\lambda_0^{(m)}\geq 0(m=1,2)$, 并且$C=\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^{\mathrm{T}}x_i\right)$非奇异, 则: $\sqrt{n}(\hat{\beta}_n-\beta)\xrightarrow{d}\arg\min(V)$ 其中,

$$V(u)=-2u^{\mathrm{T}}W+u^{\mathrm{T}}Cu+\lambda_0^{(1)}\sum_{j=1}^{p}\mathrm{sgn}(\beta_j)u_j\left\{I(|\beta_j|)\leq\lambda_0^{(1)}+\frac{(a\lambda_0^{(1)}-|\beta_j|)_+}{(a-1)\lambda_0^{(1)}}I(|\beta_j|)>\lambda_0^{(1)}\right\}$$
$$+2\lambda_0^{(2)}\sum_{i\sim j}\left(\frac{\beta_i}{\sqrt{d_i}}-\frac{\beta_j}{\sqrt{d_j}}\right)\left(\frac{u_i}{\sqrt{d_i}}-\frac{u_j}{\sqrt{d_j}}\right)w(i,j),\ \text{且}W\sim N(0,\sigma^2C)$$

证明: 定义

$$V_n(u)=\sum_{i=1}^{n}\left\{\left(\varepsilon_i-\frac{u^{\mathrm{T}}x_i}{\sqrt{n}}\right)^2-{\varepsilon_i}^2\right\}+\sum_{j=1}^{p}\left\{P_{\lambda_n^{(1)}}(\beta_j+u_j/\sqrt{n})\right\}$$
$$+\lambda_n^{(2)}\sum_{i\sim j}\left\{\left(\left(\frac{\beta_i}{\sqrt{d_i}}-\frac{\beta_j}{\sqrt{d_j}}\right)+\left(\frac{u_i}{\sqrt{d_i}}-\frac{u_j}{\sqrt{d_j}}\right)\Big/\sqrt{n}\right)^2w(i,j)-\left(\frac{\beta_i}{\sqrt{d_i}}-\frac{\beta_j}{\sqrt{d_j}}\right)^2w(i,j)\right\}$$

其中, $u=(u_1,u_2,\cdots,u_p)^{\mathrm{T}}$, 注意到$V_n(u)$在$\sqrt{n}(\hat{\beta}_n-\beta)$处取得最小值, 且在有限维收敛的情况下有:

$$\sum_{i=1}^{n}\left\{\left(\varepsilon_i-\frac{u^{\mathrm{T}}x_i}{\sqrt{n}}\right)^2-{\varepsilon_i}^2\right\}\to-2u^{\mathrm{T}}W+u^{\mathrm{T}}Cu$$

进一步, 根据式 (2) 和式 (3) 可知:

$$\sum_{j=1}^{p}\left\{P_{\lambda_n^{(1)}}(\beta_j+u_j/\sqrt{n})\text{-}P_{\lambda_n^{(1)}}(\beta_j)\right\}$$
$$=\frac{1}{2}\sum_{j=1}^{p}\frac{P_{\lambda_n^{(1)}}{}'(|z_j|)}{|z_j|}\left\{(\beta_j+u_j/\sqrt{n})^2-{\beta_j}^2\right\}\to\lambda_0^{(1)}\sum_{j=1}^{p}\frac{\beta_ju_j}{|z_j|}\left\{I(|z_j|)\leq\lambda_0^{(1)}+\frac{(a\lambda_0^{(1)}-|z_j|)_+}{(a-1)\lambda_0^{(1)}}I(|z_j|)>\lambda_0^{(1)}\right\}$$

对任意$\beta_j\approx z_j$, 有:

$$\sum_{j=1}^{p}\left\{P_{\lambda_n^{(1)}}(\beta_j+u_j/\sqrt{n})-P_{\lambda_n^{(1)}}(\beta_j)\right\}\to\lambda_0^{(1)}\sum_{j=1}^{p}\mathrm{sgn}(\beta_j)u_j\left\{I(|\beta_j|)\leq\lambda_0^{(1)}+\frac{(a\lambda_0^{(1)}-|\beta_j|)_+}{(a-1)\lambda_0^{(1)}}I(|\beta_j|)>\lambda_0^{(1)}\right\}$$

同样地, 关于第三项有:

$$\lambda_n^{(2)}\sum_{i\sim j}\left\{\left(\left(\frac{\beta_i}{\sqrt{d_i}}-\frac{\beta_j}{\sqrt{d_j}}\right)+\left(\frac{u_i}{\sqrt{d_i}}-\frac{u_j}{\sqrt{d_j}}\right)/\sqrt{n}\right)^2w(i,j)-\left(\frac{\beta_i}{\sqrt{d_i}}-\frac{\beta_j}{\sqrt{d_j}}\right)^2w(i,j)\right\}$$
$$\to2\lambda_0^{(2)}\sum_{i\sim j}\left(\frac{\beta_i}{\sqrt{d_i}}-\frac{\beta_j}{\sqrt{d_j}}\right)\left(\frac{u_i}{\sqrt{d_i}}-\frac{u_j}{\sqrt{d_j}}\right)w(i,j)$$

因此, 在有限维收敛的情况下有:

$$V_n(u)\xrightarrow{d}V(u)$$

又V_n 为凸函数且V 有最小值, 可得:

$$\arg\min(V_n)=\sqrt{n}(\hat{\beta}_n-\beta)\xrightarrow{d}\arg\min(V).$$

证毕.

2.2 求解方法

本节给出模型 SSNL 的求解算法, 具体如下:

(1) 固定 v 更新β时, 相当于解决 SCAD-Net 正则化回归问题, 即 SNL. 本文利用坐标下降法进行求解, 具体来说, 式 (8) 关于$\beta_j(j=1,2,\cdots,p)$求导, 可得:

$$\frac{\partial\ell}{\partial\beta_j}=-\sum_{i=1}^{n}v_ix_{ij}(y_i-x_i^{\mathrm{T}}\beta)+P_{\lambda_1}{}'(|\beta_j|)+\lambda_2\beta^{\mathrm{T}}L_j$$
$$=-\sum_{i=1}^{n}v_ix_{ij}(y_i-x_{i,-j}{}^{\mathrm{T}}\beta_{-j}-x_{ij}\beta_j)+\frac{P_{\lambda_1}{}'(|z_j|)}{|z_j|}\beta_j$$
$$+\lambda_2\beta_{-j}{}^{\mathrm{T}}L_{-j,j}+\lambda_2L_{jj}\beta_j$$

令其等于 0, 有:

$$\hat{\beta}_j=\frac{\sum_{i=1}^{n}v_ix_{ij}(y_i-x_{i,-j}{}^{\mathrm{T}}\beta_{-j})-\lambda_2\beta_{-j}{}^{\mathrm{T}}L_{-j,j}}{\sum_{i=1}^{n}v_ix_{ij}{}^2+\dfrac{P_{\lambda_1}{}'(|z_j|)}{|z_j|}+\lambda_2L_{jj}} \tag{9}$$

具体更新算法如算法 1 所示.

算法 1. SNL

1. 令t=0,$\beta_j(t)=\beta_j{}^0$, 其中β^0为 OLS 估计量, 各调优参数可利用交叉验证方法得到;
2. 通过式 (9) 依次更新$\beta_j(t)$,j=1,2,$\cdots$,p;
3. 令$t\leftarrow t+1$, 重复步骤 2 直至:

$$\sum_{j=1}^{p}|\beta_j(t)-\beta_j(t-1)|<10^{-6}$$

(2) 固定β更新 v 时, 式 (8) 关于v_i求导, 可得:

$$\frac{\partial\ell}{\partial v_i}=l_i(\beta^t)-\tau$$

进而有:

$$\hat{v}_i=\begin{cases}1, & l_i(\beta^t)\le\tau\\ 0, & l_i(\beta^t)>\tau\end{cases} \tag{10}$$

对于样本 i, 若其损失小于超参数 τ, 则可将其视为高质量样本, 相对应的 v_i设为 1, 否则设为 0. 显然, 对于样本损失小于 τ的样本会被纳入模型中. 一旦得到 v, 我们进一步增大 τ的值, 这样具有更大损失的样本将会进入模型当中, 重复上述步骤直至收敛, 完整算法如算法 2 所示.

算法 2. SSNL

输入: 训练集$\{X_{n\times p},y_{n\times 1}\}$, 超参数 τ, μ>1
输出: 模型参数 β

1. 初始化v_i^0=1 (i=1,2,$\cdots$,n)及 τ
2. 基于算法 1 更新 β^m;
3. 基于式 (10) 更新 v^m
4. $\tau\leftarrow\mu\tau$;
5. 令$m\leftarrow m+1$, 重复步骤 2–4, 直至:

$$\sum_{j=1}^{p}|\beta_j(m)-\beta_j(m-1)|<10^{-6}$$

3 数值结果

3.1 模拟数据分析

为检验本文所提出 SSNL 模型的预测表现, 我们首先按照以下方式模拟出一个简单的基因调控网络: 假设有 200 个转录因子 (TFs), 每个转录因子调控 10 个基因,由此产生由 2200 个基因 (节点) 组成的生物基因调控网络, 转录因子之间以及与其调控的基因之间形成网络的边. 为了简单起见, 我们进一步假设模型中只有 4 个转录因子以及其调控的基因与响应变量 y 有关. 对于第一个模型, 我们按照以下方式来生成相关数据:

(1) $y=X\beta+\varepsilon$

(2) $\beta=\Big(5,\underbrace{\frac{5}{\sqrt{10}},\ldots,\frac{5}{\sqrt{10}}}_{10},-5,\underbrace{\frac{-5}{\sqrt{10}},\ldots,\frac{-5}{\sqrt{10}}}_{10},3,\underbrace{\frac{3}{\sqrt{10}},\ldots,\frac{3}{\sqrt{10}}}_{10},-3,\underbrace{\frac{-3}{\sqrt{10}},\ldots,\frac{-3}{\sqrt{10}}}_{10}\Big)^{\mathrm{T}}\in\mathbb{R}^{2200}$

(3) ε为误差项, 且$\varepsilon_i\sim 5\times N(0,1)$.

(4) 200 个转录因子服从标准正态分布, 即$x_{TF_j}\sim N(0,1)$, $j=1,2,\cdots,200$.

(5) 每个 TF 与其调控的单个基因均服从二元正态分布, 且相关系数为ρ.

对于模型 2, 我们假设

$$\beta=(5,\underbrace{\frac{-5}{\sqrt{10}},\ldots,\frac{-5}{\sqrt{10}}}_{3},\underbrace{\frac{5}{\sqrt{10}},\ldots,\frac{5}{\sqrt{10}}}_{7},-5,\underbrace{\frac{5}{\sqrt{10}},\ldots,\frac{5}{\sqrt{10}}}_{3},\underbrace{\frac{-5}{\sqrt{10}},\ldots,\frac{-5}{\sqrt{10}}}_{7},3,\underbrace{\frac{-3}{\sqrt{10}},\ldots,\frac{-3}{\sqrt{10}}}_{3},\underbrace{\frac{3}{\sqrt{10}},\ldots,\frac{3}{\sqrt{10}}}_{7},-3,\underbrace{\frac{3}{\sqrt{10}},\ldots,\frac{3}{\sqrt{10}}}_{3},\underbrace{\frac{-3}{\sqrt{10}},\ldots,\frac{-3}{\sqrt{10}}}_{7})^{\mathrm{T}}\in\mathbb{R}^{2200}$$

其他设置与模型 1 完全一致. 该模型假设同一转录因子所调控的基因既可以对响应变量 y 产生正影响, 也可以对其产生负影响.

我们将模拟数据分为训练集和测试集, 其中训练集所占比例为 70%. 在实际应用中, 人们通常使用 k (k=3, 5 或 10) 折交叉验证的方法来选择调优参数, 然而, 不同的 k 折交叉验证的方法很可能产生非常相似的预测结果[33,34]. 此外, 研究结果表明将交叉验证的折数从 10 减少到 3, 可以使算法的计算时间减少一半以上. 鉴于本文所提出的模型存在多个调优参数, 因此, 最终我们采用 3 折交叉验证的方法来选择最佳调优参数.

将基因相关系数ρ分别设为 0.2、0.5 以及 0.7. 每

种情况都独立重复模拟 50 次, 并计算得到相应的均方误差 (PMSE). 我们还进一步给出表征模型特征选择效果的两个指标, 分别是 P 和 TP. 其中 P 表示模型中非零系数的个数, TP 表示实际模型中非零系数的个数. 表 1 给出了各模型在不同情形下的模拟结果, 其中 Lasso-Net 表示 Lasso 和网络正则化; SCAD-Net 表示不使用自主学习方法的 SCAD 和网络正则化.

如表 1 所示, 在不同模型不同相关系数下, 本文提出的 SSNL 方法相比于 Lasso-Net 和 SCAD-Net 方法均给出最小的 PMSE. 此外, 在识别真正相关基因时, SSNL 相比于其他两种方法具有最高的准确性 (TP). 例如, 在 $\rho = 0.7$ 的情形下, 利用 SSNL 模型计算得到的 TP 值均超过 40, 几乎达到模型的真实值 44. 上述结果表明 SSNL 方法在处理高维度低样本、高噪声、高相关性的复杂数据集时具有良好的表现.

3.2 实际数据分析

为进一步论证 SSNL 模型的预测效果, 我们收集得到了乳腺癌细胞系数据集. 该数据集共有 56 个样本, 其中每个样本都隶属于一个确定的细胞亚型, 通过对其进行编码可以得到一个数值型响应变量. 此外, 每个乳腺细胞样本包含 39 653 个基因, 并且这些基因之间存在交互关系. 通过加权基因共表达网络分析, 我们可以得到相应的加权网络. 然后将基因表达数据与该调控网络相结合, 得到最终的研究数据集. 我们旨在探索基因网络与关注的表型之间的关联关系以及网络中的核心基因.

表 1 各模型在不同情形下的模拟结果

	ρ	Lasso-Net			SCAD-Net			SSNL		
		PMSE	P	TP	PMSE	P	TP	PMSE	P	TP
模型1	0.2	424.5	124.7	22.9	363.8	123.9	35.1	353.3	109.9	36.3
	0.5	646.1	126.5	37.3	638.0	126.3	37.4	636.2	123.7	38.5
	0.7	1165.9	126.7	38.1	1187.8	125.6	38.8	1135.8	123.8	40.8
模型2	0.2	596.1	123.9	33.6	596.8	124.6	37.8	582.2	113.2	38.3
	0.5	412.6	132.3	36.6	414.0	128.9	35.1	408. 9	127.3	37.8
	0.7	932.8	126.1	38.9	928.7	125.4	39.2	916.6	124.9	40.4

我们将数据集随机打乱, 使约 70% 的样本成为训练样本, 剩余 30% 的样本作为测试样本. 类似于上文模拟中的情形, 我们采用 3 折交叉验证来估计得到最佳的调优参数. λ_1 与 λ_2的候选值均来自于{0.01: 0.1: 5}(起始值: 步长: 终值), μ来自于{1.1: 0.1: 3}以及 τ来自于{0.1: 0.05: 0.5}. 独立重复 10 次, 计算得到相应的均方误差 (PMSE) 以及模型中非零系数的个数 P, 具体结果如表 2 所示.

表 2 各模型在乳腺癌细胞系数据集上的结果

模型	PMSE	P
SSNL	281.8	124.2
SCAD-Net	301.9	146.3
Lasso-Net	364.1	138.8

从表 2 可以看出, 本文提出的 SSNL 方法给出了最小的 PMSE, 其表现显著优于 Lasso-Net 方法, 且优于不使用自主学习的 SCAD-Net 方法. 此外, 在特征选择方面, 尽管 3 种方法的数值表现效果相当, 但 SSNL 方法仍优于其他两种对比方法. 上述结果再次说明本文所提出的 SSNL 模型在处理高维复杂网络数据集时具有良好的表现.

4 结论与展望

融合分析为基因组研究提供了一种有效的分析角度. 传统的融合分析方法是将多个数据集组合成一个集成的数据集, 然后直接对数据进行分析. 然而, 这种集成方法非但不能消除内部偏差, 甚至可能给融合数据集增加新的随机噪声和估计误差, 从而降低融合分析的统计功效. 本文提出了一种新的融合分析模型 SSNL, 该模型融合了自主学习 (SPL) 和 SCAD-Net 正则化方法. 一方面, SPL 方法能够先从低噪声样本中学习出一个基本模型, 然后通过高噪声样本学习使得模型更加稳健. 另一方面, 特征选择是 SSNL 模型的重要组成部分. SCAD 罚函数是一种常见的特征选择方法, 但 SCAD 罚函数仅是从计算的角度出发, 没有利用任何先验信息. 故在已有研究的基础上, 本文给出了结合网络结构信息的 SCAD-Net 惩罚, 并对这一问题进行了一些理论探究, 包括群组效应和渐近性质. 不同情形下的模拟分析结果以及在乳腺癌细胞系数据集上的分析结果均表明, SSNL 方法在处理高维复杂网络数据集时具有良好的预测表现.

本文使用 3 折交叉验证 (CV) 方法来选择 SSNL

模型中出现的惩罚参数. 然而, 当遇到多个超参数时, 使用 CV 方法进行网格搜索需要消耗大量的时间与内存. 最近, 一种进化计算 (EC) 方法被用来调整惩罚参数, 并且表现良好[35]. 针对本文情形, EC 方法可能是一个更好的选择. 此外, 我们还考虑将 SPL+SCAD-Net 方法拓展到其他回归模型中, 如广义线性回归等.

参考文献

1 Dang EL, Yang SY, Song CJ, *et al*. BAP31, a newly defined cancer/testis antigen, regulates proliferation, migration, and invasion to promote cervical cancer progression. Cell Death & Disease, 2018, 9(8): 791.

2 Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: Classification, prognostication, and prediction. The Lancet, 2011, 378(9805): 1812–1823. [doi: 10.1016/S0140-6736(11)61539-0]

3 Ali HR, Rueda OM, Chin SF, *et al*. Genome-driven integrated classification of breast cancer validated in over 7500 samples. Genome Biology, 2014, 15(8): 431. [doi: 10.1186/s13059-014-0431-1]

4 Hay M, Thomas DW, Craighead JL, *et al*. Clinical development success rates for investigational drugs. Nature Biotechnology, 2014, 32(1): 40–51. [doi: 10.1038/nbt.2786]

5 Ivanov S, Liao SY, Ivanova A, *et al*. Expression of hypoxia-inducible cell-surface transmembrane carbonic anhydrases in human cancer. The American Journal of Pathology, 2001, 158(3): 905–919. [doi: 10.1016/S0002-9440(10)64038-2]

6 Walker E, Hernandez AV, Kattan MW. Meta-analysis: Its strengths and limitations. Cleveland Clinic Journal of Medicine, 2008, 75(6): 431–439. [doi: 10.3949/ccjm.75.6.431]

7 Benito M, Parker J, Du Q, *et al*. Adjustment of systematic microarray data biases. Bioinformatics, 2004, 20(1): 105–114. [doi: 10.1093/bioinformatics/btg385]

8 Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics, 2007, 8(1): 118–127. [doi: 10.1093/biostatistics/kxj037]

9 Shabalin AA, Tjelmeland H, Fan C, *et al*. Merging two gene-expression studies via cross-platform normalization. Bioinformatics, 2008, 24(9): 1154–1160. [doi: 10.1093/bioinformatics/btn083]

10 Deshwar AG, Morris Q. PLIDA: Cross-platform gene expression normalization using perturbed topic models. Bioinformatics, 2014, 30(7): 956–961. [doi: 10.1093/bioinformatics/btt574]

11 Deng K, Zhang F, Tan QL, *et al*. WaveICA: A novel algorithm to remove batch effects for large-scale untargeted metabolomics data based on wavelet analysis. Analytica Chimica Acta, 2019, 1061: 60–69. [doi: 10.1016/j.aca.2019.02.010]

12 Lazar C, Meganck S, Taminau J, *et al*. Batch effect removal methods for microarray gene expression data integration: A survey. Briefings in Bioinformatics, 2013, 14(4): 469–490. [doi: 10.1093/bib/bbs037]

13 Qi LS, Chen LB, Li Y, *et al*. Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: A case study for resected stage I non-small-cell lung cancer. Briefings in Bioinformatics, 2016, 17(2): 233–242. [doi: 10.1093/bib/bbv064]

14 Kumar MP, Packer B, Koller D. Self-paced learning for latent variable models. Proceedings of the 23rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2010. 1189–1197.

15 Jiang L, Meng DY, Mitamura T, *et al*. Easy samples first: Self-paced reranking for zero-example multimedia search. Proceedings of the 22nd ACM International Conference on Multimedia. New York: ACM, 2014. 547–556.

16 Ma ZL, Liu SQ, Meng DY, *et al*. On convergence properties of implicit self-paced objective. Information Sciences, 2018, 462: 132–140. [doi: 10.1016/j.ins.2018.06.014]

17 Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267–288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]

18 Fan JQ, Li RZ. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 2001, 96(456): 1348–1360. [doi: 10.1198/016214501753382273]

19 Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(2): 301–320. [doi: 10.1111/j.1467-9868.2005.00503.x]

20 Tibshirani R, Saunders M, Rosset S, *et al*. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(1): 91–108. [doi: 10.1111/j.1467-9868.2005.00490.x]

21 Efron B, Hastie T, Johnstone I, *et al*. Least angle regression. The Annals of Statistics, 2004, 32(2): 407–499.

22 Zou H. The adaptive lasso and its oracle properties. Journal of the American Statistical Association, 2006, 101(476):

1418–1429. [doi: 10.1198/016214506000000735]

23 Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68(1): 49–67. [doi: 10.1111/j.1467-9868.2005.00532.x]

24 Huang HH, Liang Y. Hybrid $L_{1/2+2}$ method for gene selection in the Cox proportional hazards model. Computer Methods and Programs in Biomedicine, 2018, 164: 65–73. [doi: 10.1016/j.cmpb.2018.06.004]

25 Huang HH, Liu XY, Liang Y. Feature selection and cancer classification via sparse logistic regression with the hybrid $L_{1/2+2}$ regularization. PLoS One, 2016, 11(5): e0149675. [doi: 10.1371/journal.pone.0149675]

26 Liang Y, Liu C, Luan XZ, *et al*. Sparse logistic regression with a $L_{1/2}$ penalty for gene selection in cancer classification. BMC Bioinformatics, 2013, 14(1): 198. [doi: 10.1186/1471-2105-14-198]

27 Zeng LM, Xie J. Group variable selection via SCAD-L_2. Statistics, 2014, 48(1): 49–66. [doi: 10.1080/02331888.2012.719513]

28 Li CY, Li HZ. Network-constrained regularization and variable selection for analysis of genomic data. Bioinformatics, 2008, 24(9): 1175–1182. [doi: 10.1093/bioinformatics/btn081]

29 Chen JY, Zhang SH. Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data. Bioinformatics, 2016, 32(11): 1724–1732. [doi: 10.1093/bioinformatics/btw059]

30 Wang RX, Su C, Wang XT, *et al*. Global gene expression analysis combined with a genomics approach for the identification of signal transduction networks involved in postnatal mouse myocardial proliferation and development. International Journal of Molecular Medicine, 2018, 41(1): 311–321.

31 Chung F. Laplacians and the Cheeger inequality for directed graphs. Annals of Combinatorics, 2005, 9(1): 1–19. [doi: 10.1007/s00026-005-0237-z]

32 Huang HH, Liang Y. An integrative analysis system of gene expression using self-paced learning and SCAD-Net. Expert Systems with Applications, 2019, 135: 102–112. [doi: 10.1016/j.eswa.2019.06.016]

33 Singh-Blom UM, Natarajan N, Tewari A, *et al*. Prediction and validation of gene-disease associations using methods inspired by social network analyses. PLoS One, 2013, 8(5): e58977. [doi: 10.1371/journal.pone.0058977]

34 Zeng XX, Liao YL, Liu YS, *et al*. Prediction and validation of disease genes using HeteSim scores. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, 14(3): 687–695. [doi: 10.1109/TCBB.2016.2520947]

35 Wang S, Shen HW, Chai H, *et al*. Complex harmonic regularization with differential evolution in a memetic framework for biomarker selection. PLoS One, 2019, 14(2): e0210786. [doi: 10.1371/journal.pone.0210786]