DOI:10.15923/j.cnki.cn22-1382/t.2017.5.01

基于 SIS 的基因表达数据分析

王福友， 白 冰， 徐平峰*

（长春工业大学 基础科学学院，吉林 长春 130012）

摘 要： 用 SIS 方法对 36 位白血病患者中 7 126 个基因的高维数据进行降维，结合 Lasso 变量选择方法选出可能的致病基因。根据响应变量的数据类型建立了广义线性模型（Logistic 模型）。通过比较 AIC & BIC 准则以及 CV 交叉验证方法下的拟合概率图得出最优模型。

关键词： 高维数据；变量选择；SIS 方法；Lasso

中图分类号： O 212.4 **文献标志码：** A **文章编号：** 1674-1374(2017)05-0417-04

Analysis of gene expression data based on SIS method

WANG Fuyou， BAI Bing， XU Pingfeng*

(School of Basic Sciences, Changchun University of Technology, Changchun 130012, China)

Abstract: With SIS method, the dimension of 7 126 genes data from 36 leukemiapatients is decreased, and then the possible pathogenic genes are selected by means of Lasso variables. Based on data type of the variables, a generalized linear model (Logistic model) is established. The optimal model for fitting probability graph is obtained, by comparing the AIC & BIC criterion with Cross Validation (CV) verfification.

Key words: high dimensional data; variable selection; SIS method; Lasso.

0 引 言

现代技术不断发展，很多领域都产生海量复杂的数据，尤其是在医学和生物信息学等方面，寻找癌症的致病基因或影响因素一直是一个非常重要的问题，因为基因的数目非常多，而医学实验的观测样本却非常少，这种典型的高维数据导致计算量迅速上升；高维数据导致空间的样本数变少，使得某些统计上的渐近性难以实现；传统的数据处理方法在处理这类数据时不能满足稳健性要求[1]，确定致病基因比较困难。这些新现象产生了许多挑战性的工作。

事实上，许多高维统计学习问题都可以抽象为如下问题：从实际中可以得到一个或多个输出变量 y，以及与它们有关的特征或协变量 $x_1, x_2, \cdots, x_p$ 的 n 次观测，我们需要基于这些观测建立 y 与 $x_1, x_2, \cdots, x_p$ 的数学模型。与传统统计方法不同的是，此处一般情况协变量的维数 p 大于

收稿日期： 2017-06-11

基金项目： 国家自然科学基金资助项目(11401047,11571050)；吉林省科技厅发展计划基金资助项目(20140520059JH)

作者简介： 王福友(1992—)，男，河北石家庄人，长春工业大学硕士研究生，主要从事图模型方向研究，E-mail:994742613@qq.com.

*通讯作者：徐平峰(1979—)，男，汉族，吉林长春人，长春工业大学副教授，博士，主要从事图模型方向研究，E-mail:xupingfeng@ccut.edu.cn.

n，有时甚至是远大于 n（$p \gg n$）。这种情况下通常认为真实模型位于一个低维空间（至少协变量维数 p 要比样本容量 n 低），也就是常说的稀疏性（sparsity）假定[2]，否则，建立的模型根据所观测的样本是不可识别的。因此，在维数较高时采取的方法一般是变量降维，即变量选择。

那么，如何在大量的基因中对变量进行选择，SIS 方法就是处理高维情况下降维问题的，这是一种截断式的选择方法，在某些约束条件下，SIS 可以把高维线性模型从 p 维降到 $[n\gamma]<n$ 维下，这样就可以利用一些传统的方法进行变量选择。所以，文中在对基因数据应用 SIS 方法初步降维后，又结合 Lasso 变量选择方法进一步降维，在建立合适的广义线性模型[3]（Logistic 模型）后，通过比较 AIC 和 BIC 准则、CV 交叉验证方法下的拟合概率图得出最优模型。

1 高维线性模型的变量选择

1.1 SIS 方法介绍

Fan 和 Lv[4] 提出了一种新的较简单降维方法——安全独立筛选（SIS）方法。

令 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$ 是 n 维独立响应变量，n 是样本容量。考虑线性回归模型

$$\boldsymbol{Y}=\boldsymbol{X}\beta+\boldsymbol{\varepsilon}$$

其中 $\beta=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$ 是一个 p 维参数 $\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)^{\mathrm{T}}$；$\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$ 是一个 $n\times p$ 的设计矩阵，为方便讨论，假定 $\boldsymbol{X}$ 为列标准化的矩阵，$\boldsymbol{Y}$ 为中心化向量。即 $\boldsymbol{X}$ 中每一列所代表的变量的样本均值为 0，样本标准差为 1，$\boldsymbol{Y}$ 的样本均值为 0。

令 $M_*=\{1\leqslant i\leqslant p;\beta_i\neq 0\}$ 为我们感兴趣的真实稀疏模型的指标集，$s=|M_*|$ 代表 M_* 中元素的个数，也就是真实模型中回归系数不为 0 的个数。令 $\boldsymbol{\omega}=(\omega_1,\omega_2,\cdots,\omega_p)^{\mathrm{T}}$ 对于任何给定的 $\gamma\in(0,1)$，定义子模型 $M_Y=\{1\leqslant i\leqslant p;|\omega_i|$ 为前最大的 $[\gamma_n]$ 个$\}$，其中 $[n\gamma]$ 表示 $n\gamma$ 整数部分。

这样就可以把全模型指标集 $\{1,2,\cdots,p\}$ 降到一个子模型指标集 M_Y，其中，元素的个数 $n\gamma<n$，这就是文献[4]的 SIS 方法，其计算复杂性为 $O(np)$。

1.2 SIS 方法过程

1）$\sigma=10^{-3}$（初定），$m=n/\log n$；

2）计算 $\omega_i=\boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{Y}(i=1,2,\cdots,n)$ 或写成向量形式 $\boldsymbol{\omega}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$；

3）把 $|\omega_i|$ 按照从大到小排序，并选取其中 m 个最大的 $|\omega_i|$，不妨记为 $|\omega|_{(1)},|\omega|_{(2)},\cdots,|\omega|_{(m)}$；

4）如果 $|\omega|_{(m)}>\sigma\sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}y_i^2}$，则继续下一步；否则 $m=m-1$，继续判断，不妨记最后所选取的 ω_i 为 $|\omega|_{(1)},|\omega|_{(2)},\cdots,|\omega|_{(m1)}$，$m_1\leqslant m$；

5）选取 $|\omega|_{(1)},|\omega|_{(2)},\cdots,|\omega|_{(m1)}$ 所对应的自变量，不妨记其对应的观测分量为 $z_1,z_2,\cdots,z_{m1}$，注意 $z_1,z_2,\cdots,z_{m1}$ 为 $x_1,x_2,\cdots,x_p$ 的一个子集，其变量个数为 m_1。

2 基因表达数据实例分析

2.1 数据描述

文中引用数据为白血病基因表达数据集[5]中的部分数据，包含 20 个急性淋巴细胞白血病（$y=0$）和 14 个急性骨髓性白血病（$y=1$）患者的 $p=7\,126$ 个基因表达数据。其中 y 表示分类因变量（$y=0$ 或 1）。$\{x_1,x_2,\cdots,x_p\}$ 表示白血病基因自变量。

2.2 方法应用及分析

利用 SIS 结合 Tibshirani 提出的 Lasso 惩罚似然方法[6]讨论数据中 34 名观测样本的基因筛选问题，并给出相应结果。

首先在 R 软件中，应用 SIS 程序包中惩罚似然函数把 7 126 个治病基因经过自变量筛选，将维度降低，然后再结合传统的模型选择方法如 AIC 准则、BIC 准则[7]、10 折交叉验证法[8]（CV）等给出最终模型的解释变量及相应参数向量。

经研究表明，在 R 软件的 SIS 程序包中，SIS 过程选择的最终模型类型为 cv.ncvreg、cv.glmnet 的拟合模型。对于惩罚函数的选项，如果惩罚函数为 SCAD、MCP，则返回的拟合对象的类型为 ncvreg（适用于建立普通线性回归模型）；否则，当惩罚函数为 Lasso 时，返回的拟合对象的类型为 glmnet（适用于建立广义线性模型或 Cox 比例风险模型[9]）。在本研究实例中，因变量是分类的离散变量，建立的是 Logistic 回归模型。所以只给出了 Lasso 惩罚函数下的结果，见表 1。

表 1　不同准则下的应用惩罚函数 Lssso 变量选择结果(glmnet)

准则	筛选变量个数	变量对应索引	Lambda 值(3 位有效数字)
CV	7	2 020、3 252、3 320、4 847、5 817、6 041、6 373	0.001 600
AIC	6	1 779、2 020、3 252、3 320、4 847、5 817	0.000 329
BIC	5	1 779、2 020、3 320、4 847、5 817	0.026 100

在表 1 中，SIS 过程从试验组 7 126 个基因中通过 Lasso 筛选出自变量，以此达到降维的目的，当然也给出了相应的参数向量：

1)结合 CV(10 折交叉验证)得到最终模型的参数估计值为 x_{2020}，x_{3252}，x_{3320}，x_{4847}，x_{5817}，x_{6041}，x_{6373}，分别对应模型中 $x_1,x_2,\cdots,x_7$。

$$\beta_0=-2.256$$
$$\beta_1=2.589$$
$$\beta_2=0.989$$
$$\beta_3=1.962$$
$$\beta_4=1.593$$
$$\beta_5=1.498$$
$$\beta_6=0.173$$
$$\beta_7=0.257$$

2)结合 AIC 准则得到最终模型的参数估计值为 x_{1779}，x_{2020}，x_{3252}，x_{3320}，x_{4847}，x_{5817}，分别对应模型中 $x_1,x_2,\cdots,x_6$。

$$\beta_0=-3.005$$
$$\beta_1=0.382$$
$$\beta_2=3.71$$
$$\beta_3=1.21$$
$$\beta_4=2.45$$
$$\beta_5=2.408$$
$$\beta_6=0.1\ 724$$

3)结合 BIC 准则得到最终模型的参数估计值为 x_{1779}，x_{2020}，x_{3320}，x_{4847}，x_{5817}，分别对应模型中 $x_1,x_2,\cdots,x_5$。

$$\beta_0=-1.385$$
$$\beta_1=0.236$$
$$\beta_2=1.281$$
$$\beta_3=1.042$$
$$\beta_4=1.078$$
$$\beta_5=0.705$$

类似于通常的预测方法，不同方法下预测的拟合概率图分别如图 1～图 3 所示。

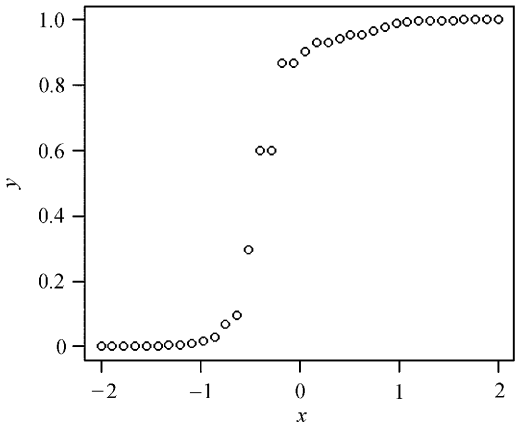

图 1　CV 法下预测的拟合概率图

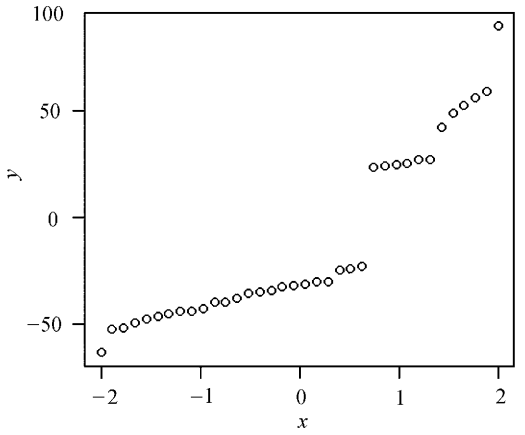

图 2　AIC 准则下预测的拟合概率图

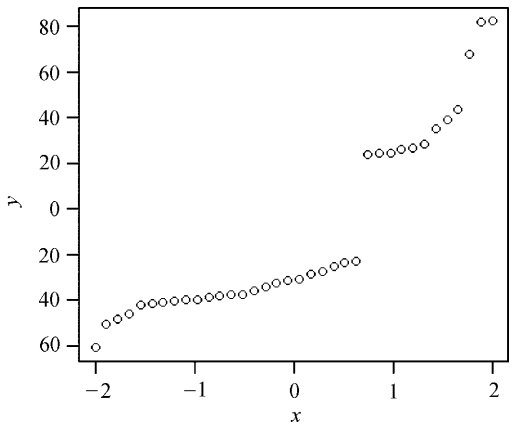

图 3　BIC 准则下预测的拟合概率图

理论上，Logistic[10]模型最佳的预测拟合图应是一条S曲线，在3种最终模型的参数估计都通过检验的情况下，显然CV(10折交叉验证)下的拟合概率图更加趋近于一条S曲线，也就得到了符合本例数据的最佳模型，即应用惩罚函数Lasso+CV方法给出的广义线性模型：

$$p=\frac{e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7}}{1+e^{\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7}}=\frac{e^{-2.256+2.589x_1+0.989x_2+1.962x_3+1.593x_4+1.498x_5+0.173x_6+0.257x_7}}{1+e^{-2.256+2.589x_1+0.989x_2+1.962x_3+1.593x_4+1.498x_5+0.173x_6+0.257x_7}}$$

3 结 语

对医学上高维数据基于SIS方法进行了分析。变量选择是一种特殊的模型选择方法，文中给出了SIS方法与经验似然有机结合SIS+CV方法以及SIS+AIC等方法。这个算法既保留了原有方法的渐近性质，又降低了实际中对误差项的分布要求，取长补短、计算简单、想法直观。研究结果表明，文中方法在对高维线性模型作变量选择时，其结果整体上可信度很高。

总之，近年来，对于各种研究领域中有关高维数据的研究一直在进行，尤其是在医学方面，在大量的基因组中寻找治病基因，并逐步走向成熟，对理论的探讨以及对实例的处理也都有很多成果。而且关于对高维数据处理和变量选择的问题应用面也越来越广泛。随着对高维数据问题的研究发现，现今对高维数据的处理方法越来越多元化。文中所考虑的高维数据变量选择方法只是处理高维数据方法中的一部分，随着科学技术的迅猛发展和理论研究的进一步探究，更多新的方法逐渐被提出，高维数据的变量选择研究领域也将得到更进一步发展。

参考文献：

[1] 刘卓.高维数据分析中的降维方法研究[D].长沙：中国人民解放军国防科学技术大学，2002.

[2] 李玲玲.高维线性模型的变量选择[D].南宁：广西师范大学，2007.

[3] 乔治·H.邓特曼.广义线性模型[M].上海：上海人民出版社，2011.

[4] Fan J, Lv J. Sure independence screening for ultra-high dimensional feature space [J]. J. R. Stat. Soc. Ser. B, 2008, 70: 849-911.

[5] Golub T R, Slonim D K, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring [J]. Science, 1999, 5439(286): 531-537.

[6] Tibshirani R. Regression shrinkage and selection via the Lasso [J]. Journal of the Royal Statistical Society, 2011, 73(3): 267-288.

[7] 崔静.广义线性模型下罚估计量的性质[D].西安：西北大学，2011.

[8] Feng Y, Yu Y. Consistent cross-validation for tuning parameter selection in high-dimensional variable selection [EB/OL]. [2017-06-11]. http://www.statslab.cam.ac.uk/~yy366/index_files/1308.5390v1.pdf.

[9] Saldana D, Feng Y. SIS: An R rackage for sure independence screening in ultrahigh dimensional statistical models [EB/OL]. [2017-06-11]. http://www.stat.columbia.edu/~yangfeng/pubs/jss1375.pdf.

[10] 陈胜利，覃家君.基于logistic增长模型的企业集团生存关系分析[J].长春工业大学学报：自然科学版，2005，26(1)：54-58.