Statistics and Application 统计学与应用, 2019, 8(3), 553-559
Published Online June 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.83062

Hans汉斯

A Comparative Study of Variable Selection Methods for High Dimensional Data Based on Logistic Regression Model

Dan Liao

College of Science, North China University of Technology, Beijing
Email: 2423078281@qq.com

Received: Jun. 6th, 2019; accepted: Jun. 21st, 2019; published: Jun. 28th, 2019

Abstract

High-dimensional data has become a hot research field in modern large data analysis. Variable selection is a widely-used method for high-dimensional data analysis. A large number of high-dimensional variable selection methods have appeared in the literatures. In order to compare the scope of application, advantages and disadvantages of several influential methods, in this paper, we consider the variable selection methods such as lasso and adaptive lasso to study the variable selection problem in logistic regression model. Firstly, by random simulation experiments, we compare the prediction and selection effects of different variable selection methods in low and high dimensions respectively. Then, we do further empirical analysis in the real data. The results show that under the same conditions, adaptive lasso has more advantages than lasso in model prediction and interpretability.

Keywords

High Dimensional Data, Variable Selection, Logistic Regression Model

基于Logistic回归模型的高维数据变量选择方法比较研究

廖　丹

北方工业大学理学院，北京
Email: 2423078281@qq.com

收稿日期：2019年6月6日；录用日期：2019年6月21日；发布日期：2019年6月28日

文章引用：廖丹. 基于 Logistic 回归模型的高维数据变量选择方法比较研究[J]. 统计学与应用, 2019, 8(3): 553-559.
DOI: 10.12677/sa.2019.83062

摘　要

高维数据已成为现代大数据分析中的热点研究领域。变量选择是一种被广泛用于高维数据分析问题的方法。文献中已出现大量高维变量选择方法，为研究其中有影响的几种方法的适用范围和利弊，本文考虑了lasso、自适应lasso等变量选择方法来研究logistic回归模型中的变量选择问题。首先，通过随机模拟实验研究，分别在低维和高维的情况下比较不同变量选择方法的预测和变量选择效果。然后，在实际数据集中做进一步地实证比较研究。研究结果表明：在同等条件下，自适应lasso在模型预测和可解释性方面均比lasso更具优势。

关键词

高维数据，变量选择，Logistic回归模型

1. 引言

随着科学技术的发展，很多研究领域中的大规模的高维数据越来越多。高维数据分析逐渐成为现代大数据分析中的一个重要的研究课题。对大规模高维数据分析的现实需求推动了统计思维、数据分析方法及理论研究的发展。

在机器学习领域，降维或变量选择技术被广泛运用于高维数据的分析问题中。降维或变量选择技术的一个关键思想是“稀疏原理”，即假设真实模型仅由预测变量的一个稀疏子集来决定。在统计学习过程中，当真实模型存在稀疏表示时，变量选择不仅能确保拟合模型的预测精度，也能提高模型的可解释性。现有文献中提出了很多变量选择的标准。传统的变量选择方法可追溯到C_p、AIC、BIC 等，但它们都存在计算开销会随着数据维度的增加而增加的问题。目前，常用变量选择方法主要有子集选择、系数压缩(或正则化)等方法。子集选择同样存在计算开销很大的问题，而且子集选择是一个离散的过程，这导致模型拟合变异性较高。系数压缩或正则化方法是一个连续的变量选择过程，即可同时进行系数估计和变量选择。它对系数施加一定程度的惩罚，使系数被压缩甚至压缩至 0。虽然正则化方法得到的估计是有偏差的，但却能在很大程度上改进模型的预测效果，即牺牲较小的偏差而在很大程度上降低方差，在提高预测精度的同时到达变量选择的目的，实现模型预测精度和可解释性的平衡[1]。在正则化方法中，具有里程碑意义 lasso (Tibshirani, 1996) [2]是一种基于“ℓ_1”惩罚的正则化方法。在此基础上，又发展出了 adaptive lasso (Hui ZOU, 2006) [3]等不同形式的变量选择方法。

本文首先介绍 logistic 回归模型、lasso、自适应 lasso 的理论研究，然后通过随机模拟实验，在低维数据和高维数据的情况下，从系数估计效果、模型的可解释性、分类正确率等方面对不同的变量选择方法进行比较，并进一步在实际数据集进行比较。

2. 理论模型介绍

2.1. Logistic 回归模型

本文主要考虑分类问题中的一个重要的模型：logistic 回归模型。给定数据$(\boldsymbol{x}_i, y_i), i=1,2,\cdots,n$，其中，

$\boldsymbol{x}_i \in \mathbb{R}^p$ 是预测变量， $y_i \in \{0,1\}$ 是服从二项分布的响应变量，logistic 回归模型的定义为

$$\mathrm{P}(y_i = 1 \mid \boldsymbol{x}_i) = p_i(\beta) = \frac{\exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)} \tag{1}$$

其中，$\boldsymbol{\beta} \in \mathbb{R}^p$ 是 $p\times 1$的系数向量。通常用极大似然估计法求解未知系数向量 $\boldsymbol{\beta}$，即求解(2)式的优化问题。

$$\hat{\boldsymbol{\beta}}_{MLE} = \arg\max_{\beta} \ell(\boldsymbol{\beta}) = \arg\max_{\beta} \sum_{i=1}^{n}\left[y_i \log p_i(\boldsymbol{\beta}) + (1-y_i)\log\left\{1-p_i(\boldsymbol{\beta})\right\}\right] \tag{2}$$

从分析上来说，由于(2)不存在闭合解，因此用牛顿迭代法来求解(2)式。牛顿迭代法的迭代过程公式如下：

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} - \left\{\sum_{i=1}^{n} w_i\left(\hat{\boldsymbol{\beta}}^{(t)}\right)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right\}^{-1} \frac{\partial \ell\left(\hat{\boldsymbol{\beta}}^{(t)}\right)}{\partial \boldsymbol{\beta}} \tag{3}$$

其中，$w_i(\boldsymbol{\beta}) = p_i(\boldsymbol{\beta})\left\{1-p_i(\boldsymbol{\beta})\right\}$，$\ell(\boldsymbol{\beta})$为(2)式中的损失函数。若(3)收敛，迭代过程终止。

2.2. 变量选择方法

Lasso 是一种可以同时进行变量选择和系数估计的正则化方法(Tibshirani, 1996) [3]。Logistic 回归模型的 lasso 估计可定义为

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\beta} \sum_{i=1}^{n}\left[-y_i\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) + \log\left(1+\mathrm{e}^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)\right] + \lambda\sum_{j=1}^{p}\left|\beta_j\right| \tag{4}$$

其中，λ 是非负的正则参数。(4)式中的第二项也称为“ℓ_1”惩罚项。当 λ 逐渐增大时，lasso 方法可连续地将系数向 0 压缩，当 λ 足够大时，一些系数可以被压缩为 0，从而达到变量选择的目的。从(4)显然可见，lasso 方法对所有系数所施加的惩罚都是相同的，因此在 lasso 方法下，系数的值越大，估计结果的偏差越大，而且 lasso 估计不具有神谕性(oracle properties)，神谕性指正确选择真实模型中的变量的概率收敛到 1，且非零系数的估计是渐近正态的。ZOU (2006)提出了自适应 lasso (adaptive lasso)，其关键思想是在“ℓ_1”惩罚项中，对不同的系数 β_j 使用不同的自适应权重。Adaptive lasso (以下简称 alasso)估计定义为

$$\hat{\beta}_{alasso} = \arg\min_{\beta} \sum_{i=1}^{n}\left[-y_i\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) + \log\left(1+\mathrm{e}^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)\right] + \lambda\sum_{j=1}^{p}\omega_j\left|\beta_j\right| \tag{5}$$

其中，$\boldsymbol{\omega} = (\omega_1, \omega_2, \cdots, \omega_P)^{\mathrm{T}}$ 为已知的基于数据集得到的权重向量。不同于 lasso，自适应 lasso 估计具有神谕性。Logistic 回归模型中，通常用极大似然估计来设计自适应权重，即 $\omega_j = \left(\left|\left(\hat{\beta}_{MLE}\right)_j\right|\right)^{-\gamma}$，$\gamma$ 是一个大于 0 的常数，该方法记为 APLR。然而，在高维数据中，由(2)式定义的极大似然估计可能无法求解，此时基于极大似然估计的自适应权重的变量选择方法就不再适用于高维数据的情况。为解决 alasso 在高维数据中面临的问题，有学者用 lasso 估计来设计自适应权重(Bielza, C., Robles, V., & Larrañaga, P. (2011)) [4]，即 $\omega_j = \left(\left|\left(\hat{\beta}_{lasso}\right)_j\right|\right)^{-\gamma}$，该方法记为 LAPLR。又有学者提出了基于预测变量的相关系数来设计自适应权重的方法[5]，即 $\omega_j = \left(\left|\left(\hat{\beta}_{cb}\right)_j\right|\right)^{-\gamma}$，其中 $\hat{\beta}_{cb}$ 可通过求解(6)式的优化问题得到，该方法记为 CAPLR。

$$\hat{\beta}_{cb} = \arg\min_{\beta} \sum_{i=1}^{n}\left[-y_i\left(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) + \log\left(1+\mathrm{e}^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)\right] + \lambda\sum_{i=1}^{p-1}\sum_{j>i}\left\{\frac{\left(\beta_i-\beta_j\right)^2}{1-\rho_{ij}} + \frac{\left(\beta_i+\beta_j\right)^2}{1+\rho_{ij}}\right\} \tag{6}$$

ρ_{ij} 为第 i，j 个预测变量之间的相关系数，注意 $\rho_{ij} \neq 1$。

3. 随机模拟研究

本文通过随机模拟实验来比较 2.2 中提到的 lasso 及基于不同权重设计方案的自适应 lasso 在 logistic 回归模型中的变量选择和预测效果。我们主要通过系数的均方误差 $MSE\left(\hat{\boldsymbol{\beta}}\right)$，变量选择的特异性 $Specificity\left(\hat{\beta},\beta\right)$ 和敏感性 $Sensitivity\left(\hat{\beta},\beta\right)$、分类正确率率这几个指标来进行比较[6]。它们的定义如(7)，(8)，(9)所示。

$$MSE\left(\hat{\boldsymbol{\beta}}\right)=\frac{1}{B}\sum_{b=1}^{B}\left\|\hat{\boldsymbol{\beta}}_b-\boldsymbol{\beta}\right\|^2 \tag{7}$$

$$Sensitivity\left(\hat{\beta},\beta\right)=\frac{\#\left\{(b,j):\hat{\beta}_{bj}\neq 0,\beta_{bj}\neq 0\right\}}{\#\left\{(b,j):\beta_{bj}\neq 0\right\}} \tag{8}$$

$$Specificity\left(\hat{\beta},\beta\right)=\frac{\#\left\{(b,j):\hat{\beta}_{bj}=0,\beta_{bj}=0\right\}}{\#\left\{(b,j):\beta_{bj}=0\right\}} \tag{9}$$

其中，B 表示随机模拟的次数，“#”表示计数。从(8)式和(9)式可以看出，敏感性指标反映的是真实模型中系数非零的变量的选择情况，敏感性越接近于 1，说明变量选择方法越能识别模型的非零系数，若等于 1，则说明真实模型的非零参数能完全被识别出来。特异性指标的值越接近于 1，说明变量选择的方法能正确识别真实模型中的零系数变量，模型的可解释性越好，若其值等于 1，则说明真实模型的零参数能完全被识别出来。两个指标的值越接近于 1，变量选择的效果越好。

3.1. 低维数据情形

首先，我们先研究低维情况下的变量选择问题。本文模拟实验的参数设置如下：$n=500, p=20$，预测变量的稀疏水平 $s=5$，系数真值为 $\beta_0=(1,1,1,1,1,0,0,0,\cdots,0)^{\mathrm{T}}$，$\beta_0$ 的前 s 个系数为 1，其余为 0。预测变量之间的自相关系数 $\rho=0.5$，均值向量 $\boldsymbol{\mu}=\mathbf{1}_p^{\mathrm{T}}$。设计矩阵 $X_{n\times p}$ 中每一个观测分别从以下几个多元分布中独立生成：1) 多元正态分布：$X\sim N\left(\boldsymbol{\mu},\Sigma\right)$，$\Sigma_{ij}=\rho^{|i-j|}$ (下同)，记为 G；2) 多元 t 分布：$X\sim t_d\left(\boldsymbol{\mu},\Sigma\right)$，$d=1,3,10$ 为自由度，对应数据集分别记为 T1，T3，T10。然后按照模型(1)生成响应变量，随机模拟实验次数为 $B=100$。

由表 1 可以看出，在不同的数据分布中，基于不同权重的自适应 lasso 的预测效果及变量选择结果均优于 lasso。从模型预测的角度来说，自适应 lasso 的系数估计的 MSE 均小于 lasso 的 MSE，且自适应 lasso 方法的分类正确率均大于 lasso 方法。从变量选择的角度来说，四种方法的敏感性均为 1，表明四种变量选择方法均能正确识别出真实模型中的非零系数(或重要变量)，但在四个数据集中，lasso 的特异性均低于自适应 lasso 方法，说明 lasso 方法更容易将系数为零 0 的变量选入模型中，模型的可解释性没有自适应 lasso 的好，这也反映了自适应 lasso 方法的神谕性。值得注意的是，基于不同权重的自适应 lasso 方法适用于不同数据分布。正态分布中，APLR 效果最好，在 T1 和 T3 中，LAPLR 最优，而在 T10 中则是 CAPLR 最好。

3.2. 高维数据情形

本节研究高维数据变量选择的问题，即 $p>n$。由于极大似然估计不适用于高维数据的情形，因此在一部分我们只比较 lasso、LAPLR、CAPLR 的变量选择和预测效果。

Table 1. The variable selection and prediction results in low dimension setting
表 1. 低维情形下的变量选择及预测结果

数据	方法	MSE	分类正确率	敏感性	特异性
G	lasso	0.5828	0.9445	1.00	0.8480
	LAPLR	0.3284	0.9424	1.00	0.9400
	APLR	0.2380	0.9479	1.00	0.9940
	CAPLR	0.2741	0.9444	1.00	0.9380
T1	lasso	0.9292	0.9651	1.00	0.4600
	LAPLR	0.6200	0.9696	1.00	0.9360
	APLR	0.8197	0.9695	1.00	0.9307
	CAPLR	0.3544	0.9691	1.00	0.9393
T3	lasso	0.4263	0.9463	1.00	0.6707
	LAPLR	0.2170	0.9465	1.00	0.8953
	APLR	0.2560	0.9476	1.00	0.8587
	CAPLR	0.2960	0.9476	1.00	0.8507
T10	lasso	4.3922	0.9240	1.00	0.2813
	LAPLR	3.7008	0.9271	1.00	0.4760
	APLR	3.3473	0.9289	1.00	0.4247
	CAPLR	2.9012	0.9324	1.00	0.5027

随机模型试验的参数设置如下：$n=500, p=1000, s=5$，β_0 的前五个元素等于 1，其余为 0。其余参数与低维情形相同。

由表 2 可知，在高维数据中，两种自适应 lasso 方法的预测及变量选择效果均优于 lasso 方法。从模型预测的角度来说，两种自适应 lasso 系数估计的均方误差 MSE 更小，分类正确率更高。从变量选择的角度来说，四种方法的敏感性都等于 1，表明这两种方法均均能正确识别真实模型中的重要变量，而自 LAPLR、CAPLR 的特异性在四个数据分布中都等于 1，表明自适应 lasso 能完全识别出真实模型，具有很好的神谕性，模型的可解释性要优于 lasso。

Table 2. The variable selection and prediction results in high dimension setting
表 2. 高维情形下的变量选择及预测结果

数据	方法	MSE	分类正确率	敏感性	特异性
G	lasso	2.1328	0.8431	1.00	0.9589
	LAPLR	1.0352	0.8765	1.00	1.00
	CAPLR	1.6061	0.8637	1.00	0.9878
T1	lasso	1.5561	0.8908	1.00	0.9756
	LAPLR	0.7094	0.9013	1.00	1.00
	CAPLR	0.8462	0.9020	1.00	1.00
T3	lasso	1.1132	0.8057	1.00	0.9664
	LAPLR	0.6816	0.8058	1.00	1.00
	CAPLR	0.4923	0.8100	1.00	1.00
T10	lasso	0.8466	0.8531	1.00	0.9687
	LAPLR	0.2975	0.8627	1.00	1.00
	CAPLR	0.2101	0.8600	1.00	1.00

4. 实证分析

我们用数据集“Pima Indians Diabetes”对上述几种方法的预测和变量选择效果进行比较。Pima Indians Diabetes 数据集包含 392 个医疗记录，每个记录包含 8 个医学检测指标以及糖尿病检测结果。数据集的变量描述如表 3 所示，“diabetes”为响应变量，取值为$\{0,1\}$，其余变量为预测变量。

Table 3. The information of Pima Indians Diabetes
表 3. 皮马印第安人糖尿病数据集信息

变量	含义
pregnant	怀孕次数
glucose	血糖浓度
pressure	舒张压
triceps	三头肌皮肤皱褶厚度
insulin	血清胰岛素浓度
mass	体重
pedigree	糖尿病谱系功能
age	年龄
diabetes	糖尿病检测结果

*注：Pima Indians Diabetes 来源于 UCI 机器学习数据库。

在实证分析中，本文按 80%、20%的比例将 Pima Indians Diabetes 数据集划分为训练集，测试集。在训练集中，我们使用 10 重交叉验证，选择出最优的模型，然后在测试集中用该模型进行预测。由表 4 可知，lasso、APLR、CAPLR 选出的变量个数均为 6 个，而 APLR 选出了 5 个变量。四种方法共同选出的变量有“pregnant”、“glucose”、“triceps”、“mass”、“pedigree”，从预测角度来说，自适应 lasso 方法的预测准确率均优于 lasso 方法，其中 LAPLR 的预测准确率最高，为 79.21%。

Table 4. Prediction and variable selection of the four methods
表 4. 四种方法预测及变量选择结果

		lasso	LAPLR	APLR	CAPLR
系数估计结果	截距项	−8.5568	−9.6380	−9.0407	−9.9288
	pregnant	0.0706	0.1053	0.1593	0.1031
	glucose	0.0339	0.0370	0.0369	0.03780
	pressure	0	0	0	0
	triceps	0.0223	0.0224	0.0252	0.0261
	insulin	0	0	0	0
	mass	0.0385	0.0514	0.0492	0.0513
	pedigree	0.7819	1.1265	1.2209	1.1167
	age	0.0285	0.0261	0	0.0290
选择的变量个数		6	6	5	6
分类正确率		74.36%	78.21%	79.49%	76.92%

*选择的变量个数不包含截距项。

5. 研究结论

本文通过随机模拟实验及实际数据分别比较了 lasso、APLR、LAPLR、CAPLR 这四种变量选择方法的预测及变量选择效果。随机模拟实验的结果表明：在低维和高维的数据中，APLR、LAPLR、CAPLR 的预测及变量选择能力均优于 lasso，但是 APLR、LAPLR、CAPLR 方法之间并不存在绝对最优的方法，在不同的数据集中，这三种方法各有优势。实证分析结果与随机模拟的结果基本一致，即：相较与 lasso 而言，三种自适应 lasso 方法的预测准确率更高，APLR 方法得到的模型更稀疏，可解释性更好。值得强调的是，不同的变量选择方法适用于不同的数据集，在实际分析中，要对多种方法进行综合比较，从而选择出最适合的方法。

基金项目

北京市属高校基本科研业务费 NO.110052971921/103 资助项目。

参考文献

[1] Hastie, T., Tibshirani, R. and Friedman, J. (2009) The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Edition, Springer, Berlin.

[2] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[3] Zou, H. (2006) The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, **101**, 1418-1429. https://doi.org/10.1198/016214506000000735

[4] Bielza, C., Robles, V. and Larrañaga, P. (2011) Regularized Logistic Regression without Apenalty Term: An Application to Cancer Classification with Microarray Data. *Expert Systems with Applications*, **38**, 5110-5118. https://doi.org/10.1016/j.eswa.2010.09.140

[5] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Statistical Learning. 2nd Edition, Springer, Berlin, 204-219.

[6] 宋瑞琪, 朱永忠, 王新军. 高维数据中变量选择研究[J]. 统计与决策, 2019, 3(2): 13-16