Statistics and Application 统计学与应用, 2021, 10(3), 556-562
Published Online June 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.103057

Hans汉斯

模型平均对于糖尿病的预测

许　卫

云南财经大学，云南　昆明
Email: 1056111036@qq.com

收稿日期：2021年6月1日；录用日期：2021年6月14日；发布日期：2021年6月28日

摘　要

糖尿病导致高血压、血脂紊乱、心脑血管等疾病的主要原因，本文使用印度皮马女性有关糖尿病的数据集，使用基于逻辑回归的模型平均方法预测皮马女性五年内是否会患糖尿病。在选择模型时，经查阅资料，固定了口服葡萄糖耐量试验2小时后的血浆葡萄糖浓度、餐后2小时的血清胰岛素、糖尿病遗传函数三个指标进行建模，采用Mallows权重选择准则。实验结果表明模型平均相较于简单地逻辑回归方法预测误差率较低，效果较好。

关键词

模型选择，模型平均，逻辑回归，预测

Prediction of Diabetes by the Model Average

Wei Xu

Yunnan University of Finance and Economics, Kunming Yunnan
Email: 1056111036@qq.com

Received: Jun. 1st, 2021; accepted: Jun. 14th, 2021; published: Jun. 28th, 2021

Abstract

Diabetes is the main cause of hypertension, dyslipidemia, cardiovascular and cerebrovascular diseases. In this paper, we use the data set of Pima women in India about diabetes, and use the model averaging method based on logistic regression to predict whether Pima women will have diabetes in five years. In the selection of model, after consulting the data, fixed the oral glucose tolerance test 2 hours after the plasma glucose concentration, postprandial 2 hours of serum insulin, diabetes genetic function three indicators for modeling, using Mallows weight selection criteria. The experimental results show that the prediction error rate of the model average is lower than that of the simple logistic regression method, and the effect is better.

文章引用: 许卫. 模型平均对于糖尿病的预测[J]. 统计学与应用, 2021, 10(3): 556-562.
DOI: 10.12677/sa.2021.103057

Keywords

Model Selection, Model Average, Logistic Regression, Predict

1. 引言

糖尿病是由人体缺乏或相对缺乏胰岛素所导致的一种慢性代谢疾病，也是导致高血压、血脂紊乱、心脑血管等疾病的主要原因。目前，我国糖尿病患者已有 1.3 亿人，慢性肾病患者中不乏同时患有糖尿病的。现代人的生活饮食习惯的改变，使得火锅、甜品、奶茶等高盐、高糖、多油的食物愈加受欢迎，糖尿病前期症状隐匿，容易被忽视。这就使得进行糖尿病高危人群的预测十分有必要。

本文是基于模型平均对 21 岁及 21 岁以上的印度皮马女性糖尿病数据集进行分析建立模型，以此预测皮马女性糖尿病高危人群；在对数据集的分析建模过程中，不论用于模型的选择或是模型参数的估计，在统计学和经济学中模型平均应用广泛。在实际应用中，往往人们并不知道真实的模型，因此模型的选择至关重要，而选择的模型不论是过于简单亦或过于复杂都可能会使得估计或预测的方差偏大，显然当选择的模型不够准确时会使得估计或预测做的不够准确。而在数据集中又难免会出现离群点、破坏分布假定的点、对估计参数影响比例失衡的点我们统称这些点为强影响点。面对这类数据，除了进行数据预处理之外，模型的选择也至关重要。

模型平均是将所有备选模型通过加权的方法，通过组合估计或组合预测来降低模型选择错误带来的估计误差，避免损失原始数据信息，以此来提高估计精度。而根据数据集的不同，Zhang X.等[1]提出广义线性模型的模型平均方法针对处理本文使用的二分类被解释变量的数据集做模型平均。

在皮马女性糖尿病数据集中因被解释变量仅有两类取值，0 表示未患糖尿病，1 表示患糖尿病，面对这样的二分类被解释变量，本文选取的模型是基于逻辑回归算法。

2. 基于逻辑回归的模型平均

2.1. 逻辑回归

逻辑回归是用以处理被解释变量为二分类数据的情况，即 $Y \sim b(1,p)$。如果用线性函数 $y = x^{\mathrm{T}}\beta$ 做拟合容易受强影响点的影响，逻辑回归采用了 Sigmod 函数，即：

$$g(z) = \frac{1}{1+\mathrm{e}^{-z}}$$

减弱强影响点的影响。Sigmod 函数不仅将原先值域为 $(-\infty,+\infty)$ 映射到 $(0,1)$ 区间，并且输出的结果还具有统计学意义，如果认为被解释变量取 1 (即事件发生)的概率为 p，则用 Sigmod 函数输出的即为事件发生的概率，即 $p = \frac{1}{1+\mathrm{e}^{-x^{\mathrm{T}}\beta}}$。

2.2. 广义线性模型的模型平均

模型平均是将所有的备选模型通过权重平均起来，避免将“所有的鸡蛋放在同一个篮子里”，以此

来规避模型选择错误的风险[1]。

广义线性模型的指数分布族有如下形式：

$$f\left(y_i|\theta_i,\varphi\right)=\exp\left\{\frac{y_i\theta_i-b\left(\theta_i\right)}{\varphi}+c\left(y_i,\varphi\right)\right\}$$

其中，θ_i，φ 均为参数，$b(\cdot)$，$c(\cdot)$ 均为已知的函数，θ_i 具有形式 $\theta_i=x_i^T\beta$，通过取分别 x_i 的不同维的元素估计 β。记 $X=\left(x_1,x_2,\cdots,x_n\right)^{\mathrm{T}}$，$Y=\left(y_1,y_2,\cdots,y_n\right)^{\mathrm{T}}$，$\theta=\left(\theta_1,\theta_2,\cdots,\theta_n\right)^{\mathrm{T}}$。假设我们总共有 S 个备选模型，在第 s 个模型中，通过极大似然估计获得 β 的估计 $\hat{\beta}_{(s)}$，$\hat{\beta}_{(s)}$ 与获取的 x_i 的维度相对应，对应未获取 x_i 的维度数据的其它维值均为 0。权重向量 $\omega=\left(\omega_1,\omega_2,\cdots,\omega_S\right)^{\mathrm{T}}$，$\omega\in[0,1]^S:\sum_{i=1}^{S}\omega_i=1$。

模型平均 β 的估计 $\hat{\beta}_{(\omega)}$：

$$\hat{\beta}_{(\omega)}=\sum_{s=1}^{S}\omega_s\hat{\beta}_{(\omega)}$$

θ 的真值 θ_0，模型平均 θ_0 的估计：

$$\theta\left(\hat{\beta}_{(\omega)}\right)=\left[\theta_1\left(\hat{\beta}_{(\omega)}\right),\cdots,\theta_n\left(\hat{\beta}_{(\omega)}\right)\right]^{\mathrm{T}}=X\hat{\beta}_{(\omega)}$$

目前有许多权重选择准则，例如：张新雨等[2]罗列出基于信息准则权重选择方法的 AIC 和 BIC、Mallows 准则，本文使用的是 Mallows 准则。

Mallows 权重选择标准为：

$$G(\omega)=2\varphi^{-1}\mathrm{B}\left(\hat{\beta}_{(\omega)}\right)-2\varphi^{-1}y^{\mathrm{T}}\theta\left(\hat{\beta}_{(\omega)}\right)+\lambda_n\omega^{\mathrm{T}}k$$

其中，$\mathrm{B}\left(\hat{\beta}_{(\omega)}\right)=\sum_{i=1}^{n}b\left[\theta_i\left(\hat{\beta}_{(\omega)}\right)\right]$，$\lambda_n\omega^{\mathrm{T}}k$ 为惩罚项，$k=\left(k_1,k_2,\cdots,k_S\right)^{\mathrm{T}}$，$k_s$ 为第 s 个模型中参数的个数，λ_n 为调整参数，常用的取值为 2 或者 $\log(n)$，本文 $\lambda_n=2$。

权重向量的求解：

$$\omega=\arg\min_w G(\omega)$$

2.3. 基于逻辑回归的模型平均

伯努利分布的指数分布族：

$$f(y;p)=\exp\left\{ya-\log\left(1+\mathrm{e}^a\right)\right\}$$

其中 $a=\log\left(\frac{p}{1-p}\right)$，$\varphi=1$，$b(a)=\log\left(1+\mathrm{e}^a\right)$，$c\left(y_i,\varphi\right)=0$。

权重选择标准为：

$$G(\omega)=2\sum_{i=1}^{n}\left\{y_i\log\left(1+\exp\left(x_i^{\mathrm{T}}\hat{\beta}_{(\omega)}\right)\right)\right\}-\sum_{i=1}^{n}\left\{2y_ix_i^{\mathrm{T}}\hat{\beta}_{(\omega)}\right\}+2\omega^{\mathrm{T}}k$$

3. 糖尿病高危人群的预测

3.1. 数据来源

本文所用的数据集是来自 kaggle 网站 www.kaggle.com/uciml/pima-indians-diabetes-database 上的有关 21 岁及以上的皮马印度女性的有关身体相关指标数据，根据数据预测皮马女性 5 年内是否会患糖尿病。该数据集指标分别为：口服葡萄糖耐量试验 2 小时后的血浆葡萄糖浓度 x_1、用餐 2 小时后的血清胰岛素 x_2 (单位：μu/ml)、糖尿病遗传函数 x_3、怀孕的次数 x_4、舒张压 x_5 (单位：mmHg)、三头肌皮褶厚度 x_6 (单位：mm)、体重指数 x_7 (体重(kg)/(身高(m)^2)、年龄 x_8、类变量(0 或 1)。

3.2. 数据预处理

数据集中存在诸如血糖、舒张压、三头肌皮褶厚度等数据都存在数据为 0 的情况，但显然这些数据为异常值，本文删除了这些带有异常值的样本，实际有效样本量为 392 个。口服葡萄糖耐量实验主要是检测机体对血糖的调节功能，经过检测得到的血浆葡萄糖浓度检测是判定是否为糖尿病的重要指标；用餐后 2 小时后的血清胰岛素指标用于检测胰岛 β 细胞功能，胰岛 β 细胞用于分泌胰岛素，因此该指标是可以用来预测胰岛功能的重要指标；糖尿病不论是一型还是二型都具有一定的遗传性，糖尿病遗传函数用于判定对每个家庭而言患糖尿病的家庭遗传情况。这三个指标在糖尿病判定中有着至关重要的作用，因此在进行模型平均的备选模型选择时，本文选择了包含这三个指标的所有模型。

3.3. 基于模型平均建模

因解释变量在选择时已经确定选择了口服葡萄糖耐量试验 2 小时后的血浆葡萄糖浓度 x_1、用餐 2 小时后的血清胰岛素 x_2、糖尿病遗传函数 x_3 这三个解释变量，剩余的 5 个解释变量不确定包含哪些解释变量预测效果会更好，所以这样我们共有 32 个组合模型。又因为被解释变量为二分类变量，根据逻辑回归的变换得到的模型形如：$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\cdots\right)\right)\right)^{-1}$，所有模型如表 1 所示。

Table 1. List of alternative models
表 1. 备选模型列表

	备选模型
1	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3\right)\right)\right)^{-1}$
2	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_8x_8\right)\right)\right)^{-1}$
3	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_7x_7\right)\right)\right)^{-1}$
4	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_7x_7+\beta_8x_8\right)\right)\right)^{-1}$
5	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_6x_6\right)\right)\right)^{-1}$
6	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_6x_6++\beta_8x_8\right)\right)\right)^{-1}$
7	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_6x_6+\beta_7x_7\right)\right)\right)^{-1}$
8	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_6x_6+\beta_7x_7+\beta_8x_8\right)\right)\right)^{-1}$
9	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_5x_5\right)\right)\right)^{-1}$
10	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_5x_5+\beta_8x_8\right)\right)\right)^{-1}$
11	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_5x_5+\beta_7x_7\right)\right)\right)^{-1}$

Continued

12	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_5x_5+\beta_7x_7+\beta_8x_8\right)\right)\right)^{-1}$
13	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_5x_5+\beta_6x_6\right)\right)\right)^{-1}$
14	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_5x_5+\beta_6x_6+\beta_8x_8\right)\right)\right)^{-1}$
15	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_5x_5+\beta_6x_6+\beta_7x_7\right)\right)\right)^{-1}$
16	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8\right)\right)\right)^{-1}$
17	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4\right)\right)\right)^{-1}$
18	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_8x_8\right)\right)\right)^{-1}$
19	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_7x_7\right)\right)\right)^{-1}$
20	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_7x_7+\beta_8x_8\right)\right)\right)^{-1}$
21	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_6x_6\right)\right)\right)^{-1}$
22	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_6x_6+\beta_8x_8\right)\right)\right)^{-1}$
23	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_6x_6+\beta_7x_7\right)\right)\right)^{-1}$
24	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_6x_6+\beta_7x_7+\beta_8x_8\right)\right)\right)^{-1}$
25	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5\right)\right)\right)^{-1}$
26	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_8x_8\right)\right)\right)^{-1}$
27	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_7x_7\right)\right)\right)^{-1}$
28	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_7x_7+\beta_8x_8\right)\right)\right)^{-1}$
29	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6\right)\right)\right)^{-1}$
30	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_8x_8\right)\right)\right)^{-1}$
31	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7\right)\right)\right)^{-1}$
32	$p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8\right)\right)\right)^{-1}$

对于第 s 个模型通过极大似然估计得到 β 的估计值 $\hat{\beta}_s$，再通过模型平均的最小化目标函数 $G(\omega)$ 获取权重 ω 的估计，结果为表 2。

Table 2. Weight ω
表 2. 权重 ω

数值	数值	数值	数值
5.4353e−07	6.1118e−07	0.4150	5.2252e−07
2.5956e−06	3.3889e−07	1.4697e−06	7.6433e−07
5.5971e−07	8.0781e−07	0.0009	4.6422e−07
0.4250	4.3690e−07	1.1934e−06	4.6151e−07
9.0619e−07	7.5702e−07	1.1224e−06	3.4192e−07
0.1591	2.7979e−07	5.9069e−07	
4.2150e−07	4.4774e−07	5.6776e−07	
1.0032e−06	1.6059e−06	3.9658e−07	
3.6995e−07	7.7663e−07	7.8445e−07	

上表为由模型平均方法计算出对于各个模型的权重，权重依次按列排序，即对一个模型的权重即为第一列第一个值，第二个模型的权重即为第一列第二个值，依次类推。从表中可以看出相较于逻辑回归模型即备选模型的最后一个模型 $p=\left(1+\exp\left(-\left(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8\right)\right)\right)^{-1}$，模型平均选择的模型主要集中在第 4 个模型、第 6 个模型、第 19 个模型以及第 21 个模型。

3.4. 模型平均的预测

为了比较模型平均和逻辑回归的预测效果，将 392 个有效样本随机分为训练集和测试集，本文分别抽了总样本的 50%、55%、60%、65%、70%、75%、80%、85%、90%作为训练集，对应剩下的样本作为测试集，对于相同比例的训练集，进行重复实验 1000 次，比较平均误差率。实验结果详见表 3。

Table 3. Comparison of error rates between model average and logistic regression average
表 3. 模型平均与逻辑回归平均误差率比较

模型平均平均误差率	逻辑回归平均误差率
0.2393	0.2422
0.2381	0.2407
0.2388	0.2414
0.2359	0.2386
0.2367	0.2401
0.2359	0.2380
0.2328	0.2351
0.2361	0.2383
0.2318	0.2345

4. 结论

使用基于逻辑回归的模型平均方法对印度皮马女性糖尿病数据集进行建模，对比逻辑回归，模型平均的预测平均误差率一致比逻辑回归的预测平均误差率小，可以看出模型平均的估计效果要比逻辑回归的估计效果好，说明在预测皮马女性糖尿病时使用基于逻辑回归的模型平均方法相较于逻辑回归更为准确。

参考文献

[1] Zhang, X., Yu, D., Zou, G., *et al*. (2016) Optimal Model Averaging Estimation for Generalized Linear Models and Generalized Linear Mixed-Effects Models. *Publications of the American Statistical Association*, **111**, 1775-1790. https://doi.org/10.1080/01621459.2015.1115762

[2] 张新雨，邹国华. 模型平均方法及其在预测中的应用[J]. 统计研究, 2011(6): 97-102.