

基于 ROSE 和 SCAD 方法的广义可加 Logistic 回归模型

摘要：本文提出了 ROSE-SCAD-广义半参数可加 Logistic 回归模型。首先，对不平衡数据，采用 ROSE 算法人工合成新的数据；然后，对连续型变量采用 B 样条进行展开，并将分类变量转换为哑变量，也称虚拟变量；再通过 SCAD 方式进行变量选择用于最终的预测。

一、改进措施

在现实生活中，变量的作用通常不是线性的，Hastie 和 Tibshirani 对传统线性模型进行拓展，提出了广义可加模型，其一般形式如下：

$$g(\mu) = \beta_0 + f_1(X_1) + \dots + f_p(X_p) \tag{1}$$

其中， $g(\cdot)$ 为连接函数，单调可微， $f_j(\cdot)$ ， $j=1,2,\dots,p$ ，为未知非参数光滑函数， $E(f_j(X_j))=0, j=1,2,\dots,p$ ，本文将连接函数取为 logit 函数。

当有些变量无法判定其在模型中的形式时，故将(1)推广到广义半参数可加模型，即

$$g(\mu) = \alpha + X^T\beta + f_1(Z_1) + \dots + f_q(Z_q) \tag{2}$$

其中， $X = (X_1, X_2, \dots, X_p) \in R^p$ 为线性部分变量，通常为分类变量， $Z = (Z_1, Z_2, \dots, Z_q) \in R^q$ 非线性部分变量，通常为连续变量。

为了对高维海量数据进行变量选择，张等[1]提出了 Group LASSO 与广义半参数可加模型结合的形式：

$$\hat{\beta} = argmin(-l(\beta) + \sum_{g=1}^G \lambda_g \|\beta_{I_g}\|_2 + \sum_{k=1}^q \delta_k \|\gamma_{J_k}\|_2) \tag{3}$$

考虑不平衡数据的情况，本文使用 ROSE 算法对样本数据进行类别不均衡处理。进而本文提出了 ROSE-SCAD-广义半参数可加 Logistic 回归模型。

二、数据来源与采集时间

1、来自西南金点子的 2021 年 4 月 16 日 27 个多元金融板块和 35 个石油板块的股票数据的共计 62 个数据，均保留后 5 个数据作为测试集，其余作为训练集；

2、来自 <https://archive.ics.uci.edu/ml/index.php> 的 Bank Marketing Data Set，本文选取 bank-full.csv，共计 4521 个数据，排除日期和月份两个变量后共计 15 个变量进行实验。

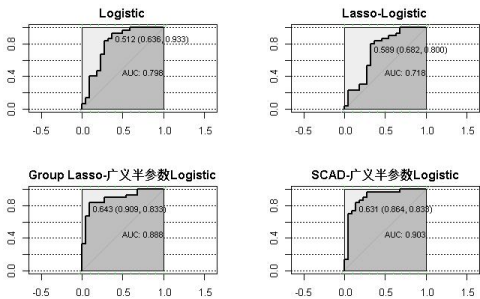
三、R 程序分析结果

1、多元金融板块和石油板块

表(1) 不同方法的内外预测正确率

	内预测正确率	外预测正确率
Logistic	0.462	0.8
Lasso-Logistic	0.423	0.8
Group Lasso-广义半参数 Logistic	0.481	0.8
SCAD-广义半参数 Logistic	0.481	0.7

观察表(1)发现 SCAD-广义半参数 Logistic 表现得并不是很好，但是观察图(1)其 AUC 值是最大的，可能只是样本划分的原因。



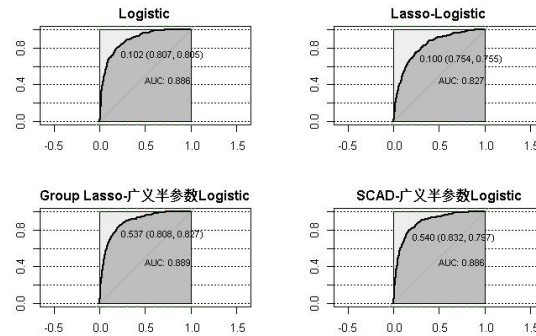
图(1) 不同方法的 AUC

2、Bank Marketing Data Set

表(2) 不处理数据时不同方法的内外预测正确率

	内预测正确率	外预测正确率
Logistic	0.90518	0.90273
Lasso-Logistic	0.88496	0.89020
Group Lasso-广义半参数 Logistic	0.89602	0.89462
SCAD-广义半参数 Logistic	0.90139	0.90420

当不对数据做处理时，SCAD-广义半参数 Logistic 的预测效果是最好的。其 AUC 值也达到了 0.886。



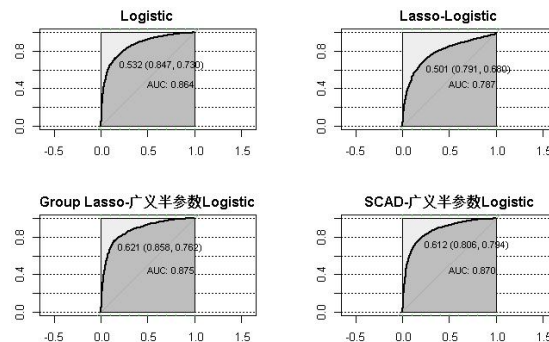
图(2) 不处理数据时不同方法的 AUC

采用 ROSE 算法对数进行处理后的效果如下。

表(3) ROSE 算法处理后不同方法的内外预测正确率

	内预测正确率	外预测正确率
Logistic	0.77971	0.78629
Lasso-Logistic	0.73388	0.72292
Group Lasso-广义半参数 Logistic	0.80373	0.79735
SCAD-广义半参数 Logistic	0.80310	0.79145

采用 ROSE 算法后内预测效果变差了。



图(3) ROSE 算法处理后不同方法的 AUC

四、结论与展望

本文对于非平衡数据采用了 ROSE 算法对数据进行人工合成进而改进，这是基于数据特征做的一种变换，虽然效果欠佳，但提供了一种思路，即在做改进时可以基于数据特征进行，又比如，方^[3]等基于数据中含有有标签数据和无标签数据将张等^[1]提出的模型改进成为半监督形式的模型。

参考文献：

- [1] 张娟,张贝贝.基于 Group-LASSO 方法的广义半参数可加信用评分模型应用研究[J].数理统计与管理,2016,35(03):517-524.
- [2] 韦勇凤,向一波.基于 Group-Lasso 方法的非均衡数据信用评分模型[J].中国科学院大学学报,2021,38(02):181-188.
- [3] 方匡南,陈子岚.基于半监督广义可加 Logistic 回归的信用评分方法[J].系统工程理论与实践,2020,40(02):392-402.