【统计理论与方法】

基于 AUC 回归的不平衡数据特征选择模型研究

李 扬[1a,1b,1c],李竟翔[2],王园萍[3]

(1. 中国人民大学 a. 应用统计科学研究中心,b. 统计学院,c. 统计咨询研究中心,北京 100872;
2. 美国明尼苏达大学 统计学院,明尼阿波利斯 55455; 3. 日立(中国)研究开发有限公司 顾客创办中心,北京 100190)

摘要:针对不平衡数据的泛化预测和特征选择问题,提出了一种引入 MCP 惩罚函数的 AUC 回归模型(MCP-AUCR)。该模型采用考虑所有阈值信息的优化目标函数,具有处理不平衡数据的能力,并具有较好的特征选择效果;在讨论该模型定义与原理的基础上,提出相应的循环坐标下降训练算法,并通过数值模拟研究验证其优良性质;针对中国股票市场机械、设备、仪表板块中的上市公司,构建了基于 MCP-AUCR 的财务预警模型。研究结果显示:该财务预警模型可以选择出可解释的重要财务指标并进行有效预测,显著优于传统模型。

关键词:AUC 回归;MCP 惩罚;特征选择;财务预警

中图分类号:O212.1;F224.0 **文献标志码**:A **文章编号**:1007-3116(2015)05-0010-07

一、引 言

随着"大数据时代"的来临,在数据采集与存储越来越便捷的同时,也导致了大量信息冗余问题的出现。在预测研究中,研究者为了避免遗漏重要的预测变量,往往向模型中引入尽可能多的预测变量。然而,过多甚至冗余的变量不仅会使训练得到的模型难以解释,还会带来诸如多重共线性、模型统计推断失效等问题,从而降低模型的泛化能力。特征选择方法正是解决这类问题的有效途径:"它通过剔除掉那些不重要的变量,使模型更加简洁且易于解释"[1]。多种基于罚函数的特征选择方法在线性回归分析中已经被提出和使用,如 Lasso、SCAD、MCP、Group Lasso 等方法[2-5]。

从统计角度看,可以将特征选择方法视为寻找最优的预测变量子集合的过程,而评价特征选择效果优劣的标准是模型泛化能力的大小。描述模型泛化能力的指标大多基于单个混淆矩阵计算,例如模型的召回率、准确率等。在样本存在不平衡的情况下(即不同类别样本比例相差悬殊的情况),基于这类指标训练模型的预测精度会受到模型阈值的严重影响。以二分类 Logistic 回归模型为例,训练后 Logistic 回归模型可以输出样本归属于某类(1 类或 0 类)的预测概率,常用的阈值为 50%,即若预测概率大于等于 50%预测为 1 类,否则预测其为 0 类样本,但在两类样本不平衡的情况下,最优的阈值未必为 50%。为了得到较为准确的预测结果,需要对阈值进行优化处理,这就使模型的训练变得相对繁琐[5]。目前,解决不平衡样本问题的方法主要基于抽样理论,如 SMOTE、基于 SMOTE 的改进与基于聚类的欠抽样方法等[6-7]。这类方法的核心思想是利用重抽样扩大少数类样本的样本量,利用欠抽样去除大类样本的噪声,最后构造一个合理的平衡数据集。上述方法对不平衡样本的处理破坏了原始数据的数据结构,并且违背了研究的可重复性原则。所以,本文主要考虑在模型层面上解决不平衡样本下的特征选择,即在不破坏样本结构的前提下构造一个能够有效处理不平衡

收稿日期:2014-10-18;修复日期:2015-03-26

基金项目:国家自然科学基金青年项目《预测模型的结构化变量选择方法研究》(71301162);中国人民大学应用统计科学研究中心自主项目《高维异质性数据的特征选择方法研究》(217614000821)

作者简介:李 扬,男,北京人,副教授,研究方向:相关型数据分析,变量选择模型;
李竟翔,男,北京人,硕士生,研究方向:数据挖掘;
王园萍,女,山东烟台人,研究员,研究方向:金融数据挖掘。

样本中特征选择问题的模型。

研究者提出利用接收者操作特征曲线（ROC，Receiver Operating Characteristic Curve）及其下围面积（AUC，Area Under the ROC Curve）评估模型的泛化能力[8]。ROC 曲线将同一模型每个阈值对应的假阳比率（FPR，False Positive Rate）、真阳比率（TPR，True Positive Rate）都描绘在坐标空间中，而 AUC 即 ROC 曲线的下围面积。如果以 Y_D 作为模型输出的 1 类样本的响应值，$Y_{\bar{D}}$ 作为模型输出的 0 类样本的响应值，则 AUC 可以被理解为 $Y_D > Y_{\bar{D}}$。相对于基于单个混淆矩阵计算的指标，AUC 不受阈值变化的影响，因此可简化针对阈值优化的过程。另一方面，与 Song 和 Ma 提出的 U 统计量类似，AUC 可作为特征选择方法中的目标函数进行优化[9]。这类方法可以同时解决不平衡数据预测与特征选择问题，具有很好的研究价值。

本文针对不平衡数据的特征选择问题，提出一种基于 AUC 的回归模型，通过引入惩罚函数使之具备特征选择的能力。利用数值模拟研究验证该方法的模型泛化能力与特征选择能力，并将该方法用于上市公司财务预警模型的实证研究中，以讨论其应用价值。

二、AUC 回归特征选择模型

（一）AUC 与 AUC 回归模型

设样本量为 n，响应变量 $y_i \in \{0,1\}$，$i = 1,2,\cdots,n$，$X_i = (X_{i,1}, X_{i,2}, \cdots, X_{i,p})$ 为第 i 个样本的 p 维预测向量，则响应变量 y_i 与预测变量 X_i 之间的模型关系可表示为：

$$P(y_i = 1 \mid X_i) = G(X_i\beta) \tag{1}$$

其中 G 为未知的单调递增连接函数，而 $\beta = (\beta_1, \beta_2, \cdots, \beta_P)^T$ 为 p 维列向量，即回归系数向量。基于 G 的单调增假设，则式(1) 所构造的分类规则即为 $X_i\beta$。例如，对于某个阈值 c，如果 $X_i\beta > c$，则可认为 $y_i = 1$，反之则认为 $y_i = 0$。为了构造 ROC 曲线，引入真阳比率（TPR）与假阳比率（FPR），它们均为阈值 c 的函数，其定义为 $\mathrm{TPR}(c) = P(X\beta > c \mid y = 1)$，$\mathrm{FPR}(c) = P(X\beta > c \mid y = 0)$，其值可由样本的频率进行估计。以 FPR 为横轴，TPR 为纵轴，在每一个阈值 c 下计算出相应的坐标作图即可得到 ROC 曲线。模型的整体泛化能力则能够被 ROC 曲线的下围面积 AUC 所表示。AUC 不受阈值 c 的影响，为系数向量 β 的函数，如式(2)：

$$\mathrm{AUC}(\beta) = \frac{1}{n_D n_H} \sum_{i \in \mathbb{D}, j \in H} I(X_i\beta - X_j\beta > 0) \tag{2}$$

其中 $I(\cdot)$ 为示性函数，$\mathbb{D} = \{i \mid y_i = 1\}$ 与 $H = \{i \mid y_i = 0\}$，所以对于 β 的估计可以通过最大化 AUC(β) 来得到。这等价于构造了一个基于 AUC 的分类回归模型（为保证模型的可识别性，本文将 $\|\beta\|$ 设定为 1）。

式(2) 所示的目标函数并非连续，导致建模过程中针对 AUC(β) 的优化相对困难。为解决此问题，Ma 和 Huang 引入了平滑 AUC 函数，如式(3)[8]：

$$S(\beta) = \frac{1}{n_D n_H} \sum_{i \in \mathbb{D}, j \in H} f(X_i\beta - X_j\beta) \tag{3}$$

其中 $f(\cdot)$ 是一个一阶可导的分布函数，满足 $\lim_{x \to -\infty} f(x)$ 且 $\lim_{x \to \infty} f(x) = 1$。$f(\cdot)$ 是对示性函数的一个近似，$f(\cdot)$ 的选择影响着参数估计，但其对参数估计的影响并不显著[10]。本文将应用 $f(x) = \frac{1}{(1 + \exp(-x))}$ 对 AUC 进行平滑处理，并通过最大化 $S(\beta)$ 训练基于 AUC 的回归模型。

特别地，若将 $f(x) = x$ 作为对示性函数的一种近似，则模型退化为：

$$S(\beta) = \sum_{i=1} (n_D(1 - y_i) + n_H y_i) y_i x_i \beta \tag{4}$$

显而易见，最大化式(4) 是一个简单的加权分类模型，这也是基于 AUC 的回归模型能够处理不平衡样本分类问题的直观解释。

（二）MCP 正则化的 AUC 回归特征选择模型

为了使基于 AUC 的回归模型具有特征选择能力，可在其优化函数上加上惩罚项[9]，如式(5)：

$$S(\beta) - J(\beta;\lambda,\alpha) \tag{5}$$

其中 $J(\beta;\lambda,\alpha)$ 即为惩罚项，它是回归系数 β 的一个函数。本文考虑引入 MCP（The Minimax Concave Penalty）罚函数[11]，而构造 MCP 罚函数的目的是让模型在剔除不重要变量的同时保护重要的变量不被惩罚，这使模型就像是只对重要的变量做训练了一样，即 MCP 罚函数具备了哲人（oracle）性质。MCP 罚如式(6)：

$$J(\beta;\lambda,\alpha) = \sum_{i=1}^{P} g(|\beta_i|;\lambda,\alpha) \tag{6}$$

其中

$$g(\theta;\lambda,\alpha) = \begin{cases} \lambda\theta - \dfrac{\theta^2}{2\alpha} & \theta \leqslant \alpha\lambda \\ \dfrac{1}{2}\alpha\lambda^2 & \theta > \alpha\lambda \end{cases} \tag{7}$$

且

$$g'(\theta;\lambda,\alpha) = \begin{cases} \lambda - \dfrac{\theta}{\alpha} & \theta \leqslant \alpha\lambda \\ 0 & \theta > \alpha\lambda \end{cases} \tag{8}$$

可以看到，如果将 $g(\theta;\lambda,\alpha)$ 在 $\theta = 0$ 处作一阶泰勒

展开的话，则有：

$$g(\theta;\lambda,\alpha) \cong g(\theta;\lambda,\alpha)+g'(\theta;\lambda,\alpha)\theta = \begin{Bmatrix}\lambda\theta & \theta \leqslant \alpha\lambda \\ 0 & \theta > \alpha\lambda\end{Bmatrix} \tag{9}$$

这就表示对某个预测变量对应的回归系数当 $|\beta_j| \leqslant \alpha\lambda$ 时，对这个系数的处理 MCP 罚接近 LASSO 罚，即进行压缩；当 $|\beta_j| > \alpha\lambda$ 时，MCP 罚函数对该系数的处理等价于不加罚的模型，这就是 MCP 罚函数具备哲人性质的一个通俗解释。对于调谐参数 λ，可以通过交叉验证的 AUC 进行选择[12]。

三、循环坐标下降算法

针对基于 AUC 回归的特征选择模型，本文提出一个循环坐标下降算法进行求解。实际上，优化引入 MCP 惩罚函数的 AUC 回归模型（MCP－AUCR）等价于优化式(10) 的目标函数，即：

$$\min_{\beta} -S(\beta)+J(\beta;\lambda,\alpha) \tag{10}$$

其中

$$S(\beta)=\frac{1}{n_D n_H}\sum_{i\in\mathbb{D},j\in H} f(X_i\beta - X_j\beta)$$

$$f(x)=\frac{1}{1+\exp(-x)}$$

$$J(\beta;\lambda,\alpha)=\sum_{i=1}^{P} g(|\beta_i|;\lambda,\alpha)$$

$$g(\theta;\lambda,\alpha)=\begin{Bmatrix}\lambda\theta-\frac{\theta^2}{2\alpha} & \theta\leqslant\alpha\lambda \\ \frac{1}{2}\alpha\lambda^2 & \theta>\alpha\lambda\end{Bmatrix}$$

循环坐标下降法是指通过沿 β 的各个分量的方向轮流搜索求得最优解的过程，其优势有两点：其一，将单个高维优化问题转化为多个一维优化问题，使训练模型的难度大大降低；其二，循环坐标下降法总能够收敛到某个稳定点。为了进一步简化优化步骤，本文应用泰勒展开分别处理式(10) 中的 $S(\beta)$ 与 $J(\beta;\lambda,\alpha)$。对于 β 的某一个分量 β_j，其优化过程等价于最小化式(11) 所示关于 z 的函数 $h(z)$：

$$\begin{aligned}h(z) &= -S(\beta+ze_j)+J(\beta+ze_j;\lambda,\alpha)+S(\beta)-J(\beta;\lambda,\alpha)\\ &= -S(\beta+ze_j)+J(\beta+ze_j;\lambda,\alpha)+\text{constant}\\ &= L_j(z;\beta)+g(|\beta_j+z|;\lambda,\alpha)+\text{constant}\\ &\approx L'_j(0;\beta)z+\frac{1}{2}L''_j(0;\beta)z^2+g'(0;\lambda,\alpha)(|\beta_j+z|)+\text{constant}\end{aligned} \tag{11}$$

其中 $L_j(z;\beta) = -S(\beta+ze_j)$

最小化式(11) 得到解 d，如式(12) 所示：

$$d=\begin{Bmatrix}-\dfrac{L'_j(0;\beta)+g'(0;\lambda,\alpha)}{L''_j(0;\beta)} \\ \text{if } L'_j(0;\beta)+g'(0;\lambda,\alpha)\leqslant L''_j(0;\beta)\beta_j \\ -\dfrac{L'_j(0;\beta)-g'(0;\lambda,\alpha)}{L''_j(0;\beta)} \\ \text{if } L'_j(0;\beta)+g'(0;\lambda,\alpha)\geqslant L''_j(0;\beta)\beta_j \\ -\beta_j \qquad \text{其他}\end{Bmatrix} \tag{12}$$

解得 d 之后，设定 $\gamma=0.9$，测试 $\gamma^t d$，$t=0,1,2$ 直至找到 t 使 $h(\gamma^t d)<0$，则更新 $\beta_j \leftarrow \beta_j+\gamma^t d$。循环上述步骤直至收敛或超过预设的循环次数，即可完成对引入 MCP 惩罚函数的 AUC 回归模型的训练。具体流程如下所示。

引入 MCP 惩罚函数的 AUC 回归模型训练算法：

输入：

训练集 $\{(x_j,y_i)\}_{i=1}^{n}$，其中 $x_i\in R^p$，$y_i\in\{-1,+1)$

模型参数 λ，$a=30$，最大迭代次数 iter. max = 10

当前迭代次数 iter = 0

设定 $\gamma=0.9$，设定初始值 $\beta\in R^p$

输出：

1. while 算法不收敛或 iter <= iter. max do
2. for j = 1, 2, …, p do
3. 根据式 12 计算方向 d
4. 测试 $d,\gamma d,\gamma^2 d,\cdots$，直到找到 t 使目标函数式 10 下降
5. 更新 $\beta_j \leftarrow \beta_j+\gamma^t d$
6. end for
7. iter = iter + 1
8. end while

四、模拟研究

本文通过数值模拟研究，验证了引入 MCP 惩罚函数的 AUC 回归模型的不平衡数据处理能力和特征选择能力。为了体现该方法的优势，本文针对模拟数据同时构建基于 L1－SVM，L1 Logistic 回归的模型，比较其在泛化能力和特征选择能力两方面的表现。

在模拟研究中，首先确定数据的样本量 n，两类样本比例 $n_+ : n_-$，总变量数 p 和有效变量数 p_0；然后随机生成预测变量 $x_1,x_2,\cdots,x_p \sim i.i.d.\ N(0,1)$，从中抽取 p_0 个预测变量，利用其构造分类超平面 $\{x: f(x)=\beta_0+\beta_{(1)}x_{(1)}+\cdots+\beta_{(p_0)}x_{(p_0)}\}$。通过调整 β_0 使数据满足预先设定的样本不平衡性；最后标记每一个样本的响应值 $y_i=\text{sign}(f(x)+\varepsilon)$，其中 ε 是服从标准正态分布的随机误差项。将生成数据按

1∶1 的比例随机分为训练集和测试集，再在训练集中利用循环坐标下降算法（表 1）确定最优参数组合并构建模型。之后，在测试集中计算模型的预测 AUC，召回率（Recall），命中率（Precision），F 得分、$F2$ 得分，得到特征选择的结果，进而评价模型的分类预测与特征选择效果。

在模拟研究中，设定 $n = 3\ 000$，总特征数 $p = 50$，有效特征数 $p_0 = 10$。利用上述方法分别生成样本比例为 1∶5、1∶10、1∶20 的不平衡数据，再基于这些数据构建模型，计算模型评价指标。为了避免数据的随机性对数值模拟造成的影响，将在每个样本比例下重复模拟过程 100 次，然后记录各个模型的评价指标均值，所得结果如表 1 所示。

表 1　数值模拟结果表

不平衡率	模型	总变量数	有效变量数	召回率	命中率	F 得分	$F2$ 得分	AUC
1∶1	MCP－AUCR	25.150	10.000	0.686	0.954	0.793	0.724	0.950
	L1－SVM	24.250	10.000	0.950	0.678	0.772	0.862	0.949
	L1－Logistic	16.850	10.000	0.958	0.631	0.399	0.439	0.892
1∶5	MCP－AUCR	29.150	10.000	0.410	0.999	0.576	0.463	0.941
	L1－SVM	27.300	9.950	0.824	0.572	0.634	0.705	0.939
	L1－Logistic	17.000	9.950	0.838	0.648	0.692	0.747	0.943
1∶10	MCP－AUCR	27.850	9.700	0.363	0.997	0.524	0.413	0.930
	L1－SVM	28.050	9.350	0.755	0.532	0.559	0.612	0.927
	L1－Logistic	14.300	9.150	0.763	0.520	0.566	0.630	0.912
1∶20	MCP－AUCR	29.400	9.400	0.359	0.998	0.516	0.408	0.895
	L1－SVM	10.500	7.100	0.595	0.409	0.399	0.439	0.892
	L1－Logistic	9.100	6.400	0.510	0.419	0.419	0.453	0.839

首先，考虑三个模型在不平衡数据分类预测（泛化能力）方面的表现。最重要的描述模型预测能力的指标是预测集 AUC。从表 2 可以看到：MCP－AUCR 训练出的模型的 AUC 在不同不平衡比例下都略微高于其他两个模型，这说明 MCP－AUCR 的预测能力至少不亚于常用的 L1－SVM 与 L1－Logistic 模型；从 F 得分与 $F2$ 得分的角度上看，MCP－AUCR 与 L1－SVM 和 L1－Logsitic 模型也是不分上下的；考虑模型的特征选择能力，从表 2 中可以看到这三个模型都能有效降低模型维度，选择出有效的预测变量。通过对比可进一步发现，在有效预测变量的选择上，MCP－AUCR 模型略微优于其他两个模型，尤其对于高不平衡率（1∶20）的情况。综上，可以得出结论，即在不平衡样本分类预测问题中，本文提出的 MCP－AUCR 在泛化能力方面和特征选择方面，都有着良好的表现。

五、实证研究：企业财务预警模型的特征选择

企业财务预警模型是指：借助企业提供的财务报表、经营计划及其他相关会计资料对企业的经营活动、财务活动等进行分析预测，以发现企业在经营管理活动中存在的潜在经营风险和财务风险，并在危机发生之前向企业经营者发出警告，督促企业管理当局采取有效措施，避免潜在的风险演变成损失，起到未雨绸缪作用的一类模型[13]。本文关注的财务预警模型是针对中国股票市场中的上市公司，利用其当期的财务指标预测其未来可能出现的财务困境（被评为“ST”）。从统计学的研究角度看，这是一个典型的二分类问题，其重点在于选出能有效预测财务困境并易于解释的指标，利用财务指标构造分类器，使其能够有效地预测样本在未来是否会被评为“ST”。

由于每年被评为“ST”的上市公司属于极少数，所以文献中将年内所有被评为“ST”的上市公司作为阳性样本纳入模型，然后以一定比例（如 1∶1、1∶2等）随机或对应地选取正常的上市公司作为阴性样本纳入模型[3,14]，这种处理的优点在于保证了模型中两类样本的样本量都足够大，回避了两类样本的不平衡所带来的预测精度问题（这里的不平衡是指“ST”样本量与非“ST”样本量显著不相等）。但是，这种取样方式忽略了上市公司类型间的差异以及现实中“ST”与非“ST”两类上市公司数量极端不平衡的情况，由于模型中的样本不能代表整个股票市场或某个分类板块，这种取样方式下训练出的模型甚至难以应用在样本以外的上市公司中。所以，将一个板块内的全部上市公司视作输入样本是一个更为合理的选择，但如果不加处理地直接将某个分类板块下的股票全部纳入模型，则样本的不平衡性会非常严重（数据中“ST”样本显著少于非

“ST”样本），此时常规的模型（如 Logistic 回归）难以识别“ST”样本。因此，一个能够处理不平衡样本分类问题的模型是财务预警所必需的。

（一）样本和财务指标选择

本文针对证监会分类制造业中的机械、设备、仪表板块中上市公司的构建，引入 MCP 惩罚函数的 AUC 回归财务预警模型。在样本选择上，选取板块中的全部上市公司作为样本纳入模型。在预测变量的选择上，通过文献研究，纳入 60 个与财务预警相关的指标（如表 2 所示，数据源于国泰安经济金融研究数据库）[15－16]。模型中的响应变量为上市公司是否被评为“ST”。考虑预测变量与响应变量之间的时间滞后效应，本文选择利用 $t-3$ 年的财务指标预测 t 年时上市公司是否会被评为 ST[17]。经过剔除缺失值、异常值等处理后，数据中共包含 312 个上市公司，其中“ST”公司为 14 个，被标记为“＋1”，非“ST”公司为 298 个，被标记为“－1”。

表 2　财务指标名称列表

指标名称	符号	指标名称	符号
现金流量比率	CF_1	营运资金比率	SD_3
营业收入现金比率	CF_2	营运资金对资产总额比率	SD_4
销售收入现金比率	CF_3	营运资金对净资产总额比率	SD_5
盈余现金保障倍数	CF_4	营运资金	SD_6
每股经营活动现金净流量	CF_5	流动资产比率	LD_1
每股投资活动现金净流量	CF_6	固定资产比率	LD_2
每股筹资活动现金净流量	CF_7	股东权益对固定资产比率	LD_3
每股现金净流量	CF_8	流动负债比率	LD_4
每股企业自由现金流	CF_9	长期负债比率	LD_5
每股股权自由现金流	CF_10	权益对负债比率	LD_6
应收账款周转率	OP_1	有形净值债务率	LD_7
存货周转率	OP_2	负债与权益市价比率	LD_8
应付账款周转率	OP_3	每股净资产	SPR_1
营运资金（资本）周转率	OP_4	每股公积金	SPR_2
流动资产周转率	OP_5	每股未分配利润	SPR_3
固定资产周转率	OP_6	市净率	SPR_4
长期资产周转率	OP_7	市盈率	SPR_5
总资产周转率	OP_8	市现率	SPR_6
股东权益周转率	OP_9	市销率	SPR_7
每股营业收入	OP_10	资本积累率	DE_1
营业毛利率	PR_1	固定资产增长率	DE_2
资产报酬率	PR_2	总资产增长率	DE_3
总资产净利润率（ROA）	PR_3	营业收入增长率	DE_4
流动资产净利润率	PR_4	前五大股东持股比率	SS_1
固定资产净利润率	PR_5	第一大第二大比值	SS_2
净资产收益率（ROE）	PR_6	前五大股东持股平方和	SS_3
息税前利润	PR_7	管理者持股比例	SS_4
每股收益	PR_8	董事总人数	BS_1
流动比率	SD_1	独立董事比例	BS_2
速动比率	SD_2	综合杠杆	RK_1

（二）研究结果与分析

为了构建引入 MCP 惩罚函数的 AUC 回归财务预警模型，一方面，首先需要确定模型中 λ。本文采用五折交叉验证并计算五次平均预测 AUC 值方法确定最优的模型参数 λ，模型通过交叉验证确定了最优的 $\lambda=0.9$。另一方面，为体现 MCP－AUCR 相对于传统模型的优越性，将用同样方法训练 L1－Logistic 与 L1－SVM 模型，并与 MCP－AUCR 训练出的模型做比较。结果显示：MCP－AUCR 的五折交叉验证的平均值为 0.932，说明基于 MCP－AUCR 的财务预警模型具备较强的泛化能力，而 L1－Logistic 回归模型与 L1－SVM 模型的平均预测 AUC 分别为 0.890 和 0.862，其泛化能力不如 MCP－AUCR，也体现了 MCP－AUCR 在财务预警模型构建中的适用性及优越性。

考虑为财务预警模型选择重要的特征预测变量，这里将全部样本作为训练集，并训练 MCP－AUCR 模型，输出模型中系数不为 0 的财务指标（如表 3）。由表 3 可知，虽然共选择了 60 个财务指标作为预测变量，但模型中实际包含的有效特征财务指标仅有 4 个，这可大大降低实证研究模型的维数。

表 3　特征选择结果

指标名称	符号	标准化系　数	计算公式
营运资金比率	SD_3	－0.496	营运资金/流动资产
营运资金对资产总额比率	SD_4	－0.605	营运资金/资产总额
每股净资产	SPR_1	－0.173	股东权益总额/普通股股数
每股未分配利润	SPR_3	－0.598	未分配利润/总股数

同时，本模型在特征预测变量的解释性上也有很好的表现。首先，两个关于营运资金的指标（营运资金比率和营运资金对资产总额比率）被模型选择出来，这表明营运资金对于企业财务预警有着重要的作用，其系数是负的，说明营运资金越多企业越不容易陷入财务危机。从公司运营角度考虑，营运资金是企业流动资产和流动负债的总称，流动资产减去流动负债的余额称为净营运资金，营运资金越多说明不能偿还的风险越小。因此，营运资金的多少可以反映偿还短期债务的能力，所以营运资金被选择是非常符合预期的。其次，模型选出的第二大显著指标是每股未分配利润，其系数依然是负的，这说明每股未分配利润越高，公司的财务状况就会越健康，发生财务危机的风险越小。未分配利润是企业留待以后年度进行分配的结存利润，未分配利润有两个方面的含义：一是留待以后年度分配的利润；二是尚未指定特定用途的利润。资产负债表中的未分配利润项目反映了企业期末在历年结存的尚未分配

的利润数额，若为负数则为尚未弥补的亏损。每股未分配利润直接体现了上市公司的盈利能力，而上市公司的盈利能力越强财务风险越小，所以每股未分配利润的选择也是符合预期的。第三，模型选择出了每股净资产指标，其值是负的，说明每股净资产越大上市公司越不容易发生财务危机。

六、讨　论

本文提出的引入 MCP 惩罚函数的 AUC 回归模型将 AUC 作为目标函数直接优化，有效处理了数据中的不平衡问题，且具有较好的特征选择能力，但本研究只考虑了单个变量的特征选择问题，没有考虑解释变量间的群组相关结构。因此，当解释变量间存在较强群组结构时使用本模型可能会因忽略信息带来估计的偏差，在后续研究中可以考虑引入群组化的惩罚函数（如群组化 MCP 惩罚等）。另一方面，本文仅对线性模型展开了讨论。如果实际数据中特征变量与响应变量间为非线性关系，模型会产生有偏的参数估计，不利于模型泛化能力的最大化及其特征选择。可以考虑应用半参数技术 B 样条曲线（B－Spline）处理模型中可能存在的非线性预测变量[18]。最后，考虑实证研究中不同类型样本其显著的预测变量可能不同，譬如在企业财务预警分析研究中，对于不同规模（或不同行业）的企业（样本），其财务风险影响因素可能不同。因此，在变量选择研究中要考虑异质性因素对不同样本的有效预测变量的影响，因此在后续研究中可以考虑一种基于整合分析的异质性数据特征选择方法，即采用内外两层惩罚函数形式实现对异质性样本的变量差异化选择，帮助经济管理及其他领域实证研究者提高量化预测与科学决策的水平。

参考文献：

[1] 李扬，曾宪斌. 面板数据模型的惩罚似然变量选择方法研究 [J]. 统计研究，2014(3).

[2] Tibshirani R. Regression Shrinkage and Selection Via the Lasso [J]. Journal of the Royal Statistical Society (Series B), 1996(1).

[3] 刘遵雄，黄志强，孙清，等. SCAD 惩罚逻辑回归的财务预警模型 [J]. 统计与信息论坛，2012(12).

[4] 李淞淋，李扬，易丹辉. 有监督 Group MCP 方法的稳健性研究 [J]. 统计与信息论坛，2014(6).

[5] Li Y, Qin Y, Xie Y, Tian F. Grouped Penalization Estimation of Osteoporosis Data in Traditional Chinese Medicine [J]. Journal of Applied Statistics, 2013(4).

[6] Chawla N V, Bowyer K W, Hall L O, Kegelmeyer W P. SMOTE: Synthetic Minority Over-Sampling Technique [J]. Journal of Artificial Intelligence Research, 2002(16).

[7] 林舒杨，李翠华，江弋，林琛，邹权. 不平衡数据的降采样方法研究 [J]. 计算机研究与发展，2011(3).

[8] Ma S, Huang J. Regularized ROC Method for Disease Classification and Biomarker Selection with Microarray Data [J]. Bioinformatics, 2005(24).

[9] Song X, Ma S. Penalized Variable Selection with U-Estimates [J]. Journal of Nonparametric Statistics, 2010(4).

[10] Ma S, Huang J. Combining Multiple Markers for Classification Using ROC [J]. Biometrics, 2007(3).

[11] Zhang C. Nearly Unbiased Variable Selection Under Minimax Concave Penalty [J]. The Annals of Statistics, 2010(2).

[12] Jiang D, Huang J, Zhang Y. The Cross-Validated AUC for MCP Logistic Regression With High-Dimensional Data [J]. Statistical Methods in Medical Research, 2013(5).

[13] 王丹. 国内外财务风险预警研究文献综述 [J]. 合作经济与科技，2012(1).

[14] 刘遵雄，郑淑娟，秦宾，等. L1 正则化 Logistic 回归在财务预警中的应用 [J]. 经济数学，2012(2).

[15] 刘开瑞. 财务预警分析指标 [J]. 生产力研究，2007(4).

[16] 邱南南. 上市公司财务综合评价模型研究 [D]. 中国人民大学博士学位论文，2008.

[17] 邹学兵，魏秋萍，张景肖. 基于随机森林的财务困境预测研究 [J]. 统计学评论，2012(6).

[18] Zeng X, Ma S, Qin Y, Li Y. Variable Selection in Semiparametric Models for the Strong Hierarchical Longitudinal Data [J]. Statistics and Its Interface, 2015(8).

Study on the Feature Selection Method with the Penalized AUC Regression for the Imbalanced Data

LI Yang[1a,1b,1c], LI Jing-xiang[2], WANG Yuan-ping[3]

(a. Center for Applied Statistics, b. School of Statistics,

【统计理论与方法】

基于 LAD-LASSO 方法的逐段常数序列中的变点估计

李 强[1,2],王黎明[1]

(1. 上海财经大学 统计与管理学院,上海 200433;2. 泰山学院 数学与统计学院,山东 泰安 271021)

摘要:结构突变(变点)问题是统计学、经济学和信号处理等领域中的热点问题之一。当误差分布服从重尾分布或数据集含异常值时,LAD 估计比 OLS 估计更加稳健;LASSO 是一种流行的压缩估计和变量选择方法,将这两种经典的方法结合起来,提出基于 LAD-LASSO 的逐段常数时间序列变点估计的一种新的研究方法,其基本思想是把变点估计问题转化成变量选择问题来处理,在转化过程中对相应优化问题的约束条件仅做一次松弛。随机模拟表明:所提出的估计方法是切实可行的,算法更加简单易行,且估计结果具有很好的稳健性。

关键词:LAD;LASSO;变点;稳健性;变量选择

中图分类号:O212.1;F064.1 **文献标志码**:A **文章编号**:1007-3116(2015)05-0016-06

一、引 言

结构突变(变点)问题是统计学和经济学中的热点问题之一。宽泛的变点问题可以描述如下:观察一个按时间顺序发生的随机过程,探讨在其随机元素的分布或分布参数中是否有某个变化发生,或者说确认所观察到的随机过程是同质还是异质。"变点"就是指某个时点,在此时点上样本的分布或数字

收稿日期:2014-10-10;**修复日期**:2015-01-05

基金项目:全国统计科研计划重点项目《基于结构突变理论的通货膨胀持久性研究》(2011LZ035);山东省自然科学基金项目《基于 LASSO 与现代非参数方法的变点检测及其应用研究》(ZR2014AL006);上海财经大学研究生创新基金项目《基于 LASSO 与现代非参数统计方法的变点检测及其应用研究》(CXJJ-2014-445)

作者简介:李 强,男,山东泰安人,博士生,讲师,研究方向:应用数理统计与经济管理统计;
王黎明,男,山东青州人,理学博士,应用经济学博士后,教授,博士生导师,研究方向:应用数理统计与数量金融。

c. Statistical Consulting Center, Renmin University of China, Beijing 100872, China;
2. School of Statistics, University of Minnesota, Minneapolis 55455, USA;
3. Customer Co-creation Project, Hitachi (China) Research & Development Corporation, Beijing 100190, China)

Abstract: In this study, we propose an AUC (area under the ROC curve) regression with the MCP (the Minimax Concave Penalty) regularization (MCP-AUCR) to deal with the forecasting and feature selection issues for the imbalanced data. The proposed method can solve the imbalanced issues for the optimization of AUC based target and has a good performance on the feature selection. We discuss the idea of the MCP-AUCR and an iterative coordinate descent algorithm. Numerical studies are conducted to show the good property of the proposed method. And an applied study of the financial early warning system for Chinese listed corporations is analyzed as an illustrative example.

Key words: AUC regression; MCP penalty; feature selection; financial early warning system

(责任编辑:郭诗梦)