

文章编号:1005-0523(2011)06-0042-06

上市公司财务预警的正则化逻辑回归模型

张恒, 秦宾, 许金凤

(华东交通大学信息工程学院, 江西 南昌 330013)

摘要:基于统计学习理论的正则化技术构建L1(一范数约束惩罚)正则化的逻辑回归(Logistic Regression)模型, 同时建立了logistic回归模型和L2(二范数约束惩罚)正则化的logistic回归模型, 结合沪深股市ST公司和正常公司的T-3年和T-2年财务数据进行仿真实验用于上市公司财务预警实证分析。实验结果表明L1正则化的logistic回归模型的有效性, 并且在保证模型预测精度的同时提高模型解释性。

关键词:财务预警; L1正则化; 逻辑回归; 预测精度

中图分类号:F275

文献标志码:A

在全球经济一体化的趋势下, 我国资本市场的快速发展, 上市公司的财务质量受到普遍关注。开展上市公司财务预警研究可以为企业管理者及其利益相关者提供重要的参考依据, 对于企业预防和化解财务危机, 提高危机预警管理水平有着重要的意义。上市公司财务预警是指应用一定的模型方法, 对反映上市公司财务状况的指标和相关数据进行研究和分析, 构建合理有效的预警模型对财务危机进行预报。预警指标变量的选择直接关系到预警模型的建立和预报精度, 其维度高且相关性大, 包含了许多冗余信息, 不利于建模分析, 所以通常需要进行变量选择(或称模型选择)。一般的方法独立于模型的训练过程进行变量选择, 通常使用检验筛选、最优子集选择和逐步回归等方法寻找对响应变量最具有解释性的变量子集。

财务困境建模可以归结于模式识别、数据挖掘等领域的分类或回归问题, 通常使用有监督的学习方法。财务困境预警研究多采用多元统计或人工智能方法^[1-3]。统计模型主要包括多元线性回归模型(multivariate linear regression, MLR)、Fisher线性判别分析(linear discriminant analysis, LDA)和逻辑回归(logistic regression, LR)等^[1, 4-5]。MLR模型假定指标变量独立, 用普通最小二乘方法(ordinary least squares, OLS)求解, 但是直接用于财务困境预警研究很容易陷于过学习。引入统计学习理论中的正则化技术后, 可以使得模型拟合参数值变小, 有助于抑制过学习现象, 进而提高分类器的鲁棒性。带模型系数L2范数与L1范数正则化的MLR分别为岭回归(ridge regression, RR)和最小绝对缩减和变量选择算子(least absolute shrinkage and selection operator, LASSO)回归^[6]。

L1范数正则化技术是获取稀疏模型的具体策略, 具有很好的应用前景。本文提出基于L1范数正则化的逻辑回归财务预警模型, 开展上市公司财务困境预报实证研究, 使用沪深股市A股2007—2009年首次被ST的公司49家的和相同年份的财务状况正常公司54家的财务指标、现金流量等数据, 进行模型学习和性能测试, 并对比一般逻辑回归及L2范数正则化逻辑回归进行实证研究。

1 L1正则化逻辑回归模型

逻辑回归是一种统计分析方法, 能用于因变量二值性的判别, 通过计算样本所属类别的后验概率进行

收稿日期:2011-09-16

基金项目:国家自然科学基金项目(61065003);教育部人文社会科学研究规划基金项目(10YJC630379);江西省自然科学基金项目(2010GZS0034)

作者简介:张恒(1979—), 男, 博士, 研究方向为移动机器人定位、地图构建与导航, 移动无线传感器网络, 机器学习。

分类,能给出判别结果概率意义上的解释。LR模型所用假设简单,不要求误差趋于正态分布,也不要求自变量服从正态分布。该方法已广泛应用于经济,社会科学以及科学研究等诸多领域。但LR对预测变量之间的多重共线性干扰敏感。

考虑二元分类问题,响应变量 $y \in \{0, 1\}$, 每次观测的 p 个预测变量值表示成向量 $x = [x_1, x_2, \dots, x_p]$, 相应的响应变量隶属于类别1的后验概率:

$$p(y=1|x) = \phi(\beta^T x) = \frac{1}{1 + \exp(-\beta^T x)} \quad (1)$$

其中: $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ 为回归模型参数。假定存在 n 个训练样本 $D_n = \{(x^i, y^i)\}_{i=1}^n$, 根据样本数据及隶属类别 D_n 去估计 β , 优化问题的损失函数为对数似然函数(log-likelihood), 表示:

$$\arg \max_{\beta} \{\log p(\beta|D_n)\} = \arg \max_{\beta} \left\{ \sum_{i=1}^n \log(y^i \phi(\beta^T x^i) + (1-y^i)(1-\phi(\beta^T x^i))) \right\} \quad (2)$$

根据LASSO惩罚回归的思想,在损失函数中加对模型系数的L1范数惩罚项,得到L1正则化的logistic回归:

$$\arg \max_{\beta} \left\{ \log p(\beta|D_n) - \lambda \|\beta\|_1 \right\} = \arg \max_{\beta} \left\{ \sum_{i=1}^n \log(y^i \phi(\beta^T x^i) + (1-y^i)(1-\phi(\beta^T x^i))) - \lambda \|\beta\|_1 \right\} \quad (4)$$

对于L1正则化logistic回归模型的求解,Zhang^[7], Shevade等^[8]和Genkin等^[9]都提出了较为有效的算法。这里使用Suhrid等^[10]提出的在线“Shooting”算法求解L1正则化的logistic回归模型。该算法如下

首先,考虑logistic回归的对数似然函数为

$$\sum_{i=1}^n \log(p(y^i|\beta)) = \sum_{i=1}^n \log(y^i \phi(\beta^T x^i) + (1-y^i)(1-\phi(\beta^T x^i))) \quad (5)$$

当 $y^i = 1$ 或 $y^i = 0$ 时,对数似然函数分别为 $\log \phi(\beta^T x^i)$ 和 $\log(1 - \phi(\beta^T x^i))$, 如下

$$\sum_{i=1}^n \log(p(y^i|\beta)) = \begin{cases} \sum_{i=1}^n \log \phi(\beta^T x^i), & y_i = 1 \\ \sum_{i=1}^n \log(1 - \phi(\beta^T x^i)), & y_i = 0 \end{cases} \quad (6)$$

对于上述任意一种情形,可以采用对数似然函数在 $\beta_{i-1}^T x^i$ 邻域内的泰勒展开式去逼近, β_{i-1} 为通过前 $i-1$ 个样本估计得到的模型参数值, c_i 是与 β 无关的函数:

$$\sum_{i=1}^n \log(p(y^i|\beta)) \approx \sum_{i=1}^n (a_i (\beta^T x^i)^2 + b_i (\beta^T x^i) + c_i) = \beta^T \psi_n \beta + \beta^T \theta_n + \sum_{i=1}^n c_i \quad (7)$$

因此原优化问题转化为逼近的优化问题:

$$\arg \max_{\beta} \{\log p(\beta|D_n)\} = \arg \max_{\beta} \left\{ \beta^T \psi_n \beta + \beta^T \theta_n - \lambda \|\beta\|_1 \right\} \quad (8)$$

其中: $\psi_n = \sum_{i=1}^n a_i x^i (x^i)^T$, $\theta_n = \sum_{i=1}^n b_i x^i$ 。

定义 $\Omega = 2\psi_n \beta + \theta_n$, a_i , b_i 由相应的对数似然函数在 $\beta_{i-1}^T x^i$ 处的二阶泰勒展开式的系数获得。相应的“Shooting”算法规则:

- 1) 取迭代初始值 β_0 ;
- 2) 在第 m 步,对所有 $j=1$ 到 p , 做:

$$\beta_j = \begin{cases} 0, & \text{if } |\Omega_j| \leq \lambda \\ \frac{\lambda - \Omega_j}{2\psi_{jj}}, & \text{if } \Omega_j > \lambda \\ \frac{-\lambda - \Omega_j}{2\psi_{jj}}, & \text{if } \Omega_j < -\lambda \end{cases} \quad (9)$$

经过 p 次计算,用新的估计值替换上一步的结果。

3) 重复第2步,直到 β_m 收敛。

对比一般逻辑回归和L2正则化逻辑回归的迭代解,上述结果可以看出:L1范数约束下的损失函数在求解模型参数的过程中,在系数的绝对值得到缩减的同时可以使得部分对模型响应变量影响较小的变量系数置为0,具有很好的稀疏作用,从而达到小的计算量和变量选择的效果。

2 数据选择及检验

2.1 样本和财务指标选择

本文以沪深两市A股上市公司为研究对象,沪深股市一些上市公司因“状况异常”而被实行“特别处理”,称为ST公司,这里把ST公司定义为财务危机公司。删除非正常且数据缺失的公司,从2007—2009年间首次被ST公司中,选择了49家财务危机的公司样本。在选择非ST公司时,考虑到行业相等、规模近似等因素,选择了54家财务正常上市公司作为配对样本。

合理的财务指标选取是构建一个成熟的财务预警模型的关键,在借鉴国内外相关文献选取指标的基础上,考虑综合反映上市公司的偿债能力、盈利能力、营运能力、成长能力、现金流量等变量因素,初步筛选出27个财务指标见表1。

表1 财务预警指标

Tab.1 Financial early-warning index

指标代号	指标名称	指标代号	指标名称
x1	流动比率	x15	流动资产对总资产的比率
x2	速动比率	x16	主营业务增长率
x3	现金比率	x17	总资产扩张率
x4	流动资产对负债总额比率	x18	净资产增长率
x5	存货流动负债比率	x19	每股经营的现金流量
x6	债务偿付比率	x20	资产的经营现金流量
x7	总资产收益率	x21	主营业务现金比率
x8	资本金收益率	x22	现金流量结构比率
x9	扣除非经常损益后的利润率	x23	每股净资产
x10	营业利润率	x24	每股收益
x11	成本费用利润率	x25	扣除非经常损益的每股收益
x12	流动资产周转率	x26	每股主营收入
x13	总资产周转率	x27	每股经营活动产生的现金流量净额
x14	净资产周转率		

ST公司和非ST公司的财务指标应该是能够区别的,因此进入预警模型的指标至少能有效显著地区的区别ST公司和非ST公司,也就是说选择显著性变量。

2.2 实证检验

考虑我国上市公司被“特别处理”的具体情况,选取上市公司被ST年份的前的3年(即T-3年)财务数据和对应正常公司财务数据为训练样本数据建立模型。为了全面考察模型的预测能力,本文同时采用T-2年

财务数据建立各自预测模型,比较模型的预测准确度。

使用MATLAB进行仿真实验,求解出T-3年数据L1正则化logistic回归模型对应预测指标系数(见表2)。

表2 T-3年L1正则化logistic回归模型系数

Tab.2 Coefficients of L1 regularized logistic regression model in the year of T-3

变量	系数值	变量	系数值	变量	系数值	变量	系数值
x0	0.599 5	x9	0.000 0	x20	0.000 0	x26	-0.079 1
x3	0.000 0	x12	-0.249 9	x22	-0.035 9	x27	-0.000 1
x6	0.000 0	x13	0.000 0	x23	-0.053 1		
x7	0.000 0	x14	0.000 0	x24	0.000 0		
x8	0.000 0	x19	0.000 0	x25	0.000 0		

所以最终模型的表达式为

$$p(y=1|x)=\frac{1}{1+\exp(-0.599\ 5+0.244\ 9x_{12}+0.035\ 9x_{22}+0.053\ 1x_{23}+0.079\ 1x_{26}+0.000\ 1x_{27})}$$

模型的训练结果使得部分对模型响应变量 y 影响较小的变量系数变为0,而对模型预测值产生较大影响的变量保留下来,实现了变量的选择。上述实验结果可以看出,模型最后保留的指标有 x_{12} , x_{22} , x_{23} , x_{26} , x_{27} ,这5个指标分别表示流动资产周转率,现金流量结构比率,每股净资产,每股主营收入,每股经营活动产生的现金流量净额。这5个指标分别来自现金流量和单股指标这两个方面,这表明在T-3年,ST公司虽然账面上可能保持盈利,但是这些财务指标已经出现了异常状况。

利用了相同的财务数据建立了logistic回归模型和L2正则化logistic回归模型,实验结果见表3。

表3 T-3年数据各模型性能比较

Tab.3 Performance comparision of various models in the year of T-3

模型	模型输入变量数	训练准确率/%	测试准确率/%
logistic	27	77.94	45.71
L2正则化logistic	27	75.00	65.71
L1正则化logistic	5	69.12	71.43

根据上述模型的预测准确率,可以看出logistic回归模型在训练样本上的准确率最高,而在测试集上的准确率却很糟糕,这也印证了logistic回归模型过拟合现象,即logistic回归模型泛化能力不好。测试集结果表明,L1正则化logistic回归模型预测结果最好,L2正则化其次,普通logistic回归模型预警效果最坏。求得3种不同logistic回归预测模型系数如图1所示。

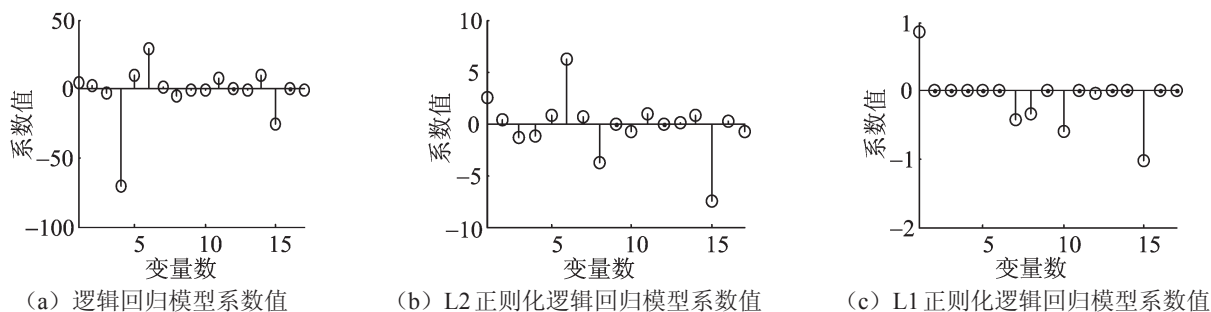


图1 T-3年数据3种模型系数值

Fig.1 Coefficient value of the three models in the year of T-3

可见,采用规则化技术的logistic回归模型系数明显得到了压缩,原始的logistic回归模型求解出的参数是没有被压缩的,所以数据的波动对模型的预测效果影响很大,模型性能不稳定。比较L2正则化和L1

正则化 logistic 回归模型,虽然二者的模型系数都被压缩了,但是 L1 正则化能将对模型相应变量影响较小的系数压缩为 0,实现了变量的选择;而 L2 正则化却仅仅能够实现对模型系数的压缩,因此也就不具备变量选择功能;L1 正则化 logistic 回归模型解释性更好。

使用 SPSS 11.5 for windows 对模型指标进行相关性分析,求取反映变量间相互线性关系的相关系数,可以判断变量密切程度。相关系数一般在 $-1\sim+1$ 范围内波动,其值愈接近 1,两个变量间的直线相关愈密切,愈接近 0,相关愈不密切。相关系数为 $0\sim0.3$ 表示相关程度低普通,在 $0.3\sim0.5$ 表示相关程度普通,相关系数 $0.5\sim0.7$ 表示相关程度显著,相关系数在 $0.7\sim0.9$ 表示相关程度高,相关系数在 $0.9\sim1.0$ 表示相关程度极高。使用 T-3 年计算所有财务指标间的相关系数,其数值多高于 0.7,说明这些指标变量间相关性较大。对 L1 正则化 logistic 回归模型 5 个系数非 0 的财务指标进行相关性分析,相关系数矩阵,见表 4。

表 4 模型变量的相关系数矩阵

Tab.4 Correlation coefficient matrix of the model variables

	x12	x22	x23	x26	x27
x12	1.000	-0.128	0.195	0.546	0.524
x22	-0.128	1.000	-0.310	-0.636	-0.392
x23	0.195	-0.310	1.000	0.445	0.563
x26	0.546	-0.636	0.445	1.000	0.689
x27	0.524	-0.392	0.563	0.689	1.000

L1 正则化 logistic 回归选取的变量之间的相关系数均低于 0.7,可见,L1 正则化 logistic 回归模型具有稀疏性,指标间不具备较高的相关性。使用 T-2 年财务数据建立 3 种 logistic 回归财务预警模型,实验结果见表 5。

表 5 T-2 年数据三种模型性能比较

Tab.5 Performance comparison of the three models in the year of T-2

模型	模型输入变量数	训练准确率/%	测试准确率/%
logistic	27	98.53	77.14
L2 正则化 logistic	27	89.71	80.00
L1 正则化 logistic	8	89.71	82.86

表 3 和表 5 实验结果显示,普通 logistic 回归模型性能在训练数据集上的表现好于正则化的 logistic 回归模型,但是在测试集上的表现明显差于正则化 logistic 回归模型。可以说明引入正则化技术在损失函数中加入惩罚项后,模型系数得到了显著压缩,有效的避免的过拟合现象,提高了模型的稳定性,并获得了较好的预测精度(使用 T-3 年,T-2 年数据的 L1 正则化 logistic 回归模型预报精度分别为 71.43%,82.86%)。L1 正则化 logistic 回归与 L2 正则化 logistic 回归模型的差别在于,前者可以使得部分对模型响应影响较小的指标变量系数变为 0,缩减了模型的变量集,实现了模型的变量选择功能,使模型更具解释性。

3 结论

逻辑回归模型已广泛用于公司财务预警研究方面,但是它对财务指标之间的多重共线性干扰敏感,模型泛化能力差。本文基于统计学习理论中的正则化技术提出 L1 正则化 logistic 回归模型,用于上市公司财务预警实证研究,有效地处理财务预警中指标变量相关性大的问题,获得的模型不仅具有很好的预报精度,而且解释性好。使用沪深股市 ST 公司和正常公司财务数据 T-3 年,T-2 年数据进行实验,验证了 L1 正则化 logistic 回归模型的有效性。今后将结合稀疏惩罚问题,深入开展稀疏统计模型研究,并将之应用于生产实践和科学研究中。

参考文献:

- [1] 吴世农,卢贤义. 我国上市公司财务困境的预测模型研究[J]. 经济研究,2001(6): 46-55.
- [2] 杨海军,太雷. 基于模糊支持向量机的上市公司财务困境预测[J]. 管理科学学报,2009,12(3):102-110.
- [3] 杨淑娥,黄礼. 基于BP神经网络的上市公司财务预警模型[J]. 系统工程理论与实践,2005,25(1):12-18.
- [4] 王宝富,李南. 财务困境的预测研究[J]. 南京航空航天大学学报:社会科学版,2007,9(3):61-64.
- [5] TIBSHIRANI R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society, 1996, Series B (Methodological), 58(1):267-288.
- [6] FU W J. Penalized Regressions: The Bridge Versus the Lasso[J]. Journal of Computational and Graphical Statistics, 1998, 7(3):397-416.
- [7] ZHANG T. On the dual formulation of regularized linear systems[J]. Machine Learning, 2002, 46(1):91-129.
- [8] SHEVADE K, KEERTHI S S. A Simple and efficient algorithm for gene selection using sparse logistic regression[J]. Bioinformatics, 2003, 19(17):2246-2253.
- [9] GENKIN D, LEWIS D, MADIGAN D. Large-scale bayesian logistic regression for text categorization[J]. Technometrics, 2007, 49(3):291-304.
- [10] BALAKRISHNAN S, MADIGAN D. Algorithms for sparse linear classifiers in the massive data setting[J]. Journal of Machine Learning Research, 2008, 9(6):313-337.

Regularized Logistic Regression Model in Financial Early-warning System with Listed Companies

Zhang Heng, Qin Bin, Xu Jinfeng

(School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: L1 norm penalized logistic regression model is proposed based on regularization technique of statistical theory. Logistic regression model and L2 regularized logistic regression model are established. Combining two years' financial data of Shanghai-Shenzhen stock ST companies and normal counterparts, simulating experiment is conducted to analyze the financial early-warning system of listed companies. The results demonstrate the good performance and predicting accuracy of L1 logistic regression model

Key words: financial early-warning system; L1 regularization; logistic regression; accuracy prediction