

文章编号: 1674-8085(2015)02-0006-05

上市公司财务预警的 T 逻辑回归模型

*徐 征¹, 刘遵雄²

(1. 华东交通大学电气学院, 江西, 南昌 330013; 2. 华东交通大学信息工程学院, 江西, 南昌 330013)

摘 要: 针对经典的逻辑回归模型易受到样本类别噪声干扰的问题, 采用 T 逻辑回归算法中的非凸损失函数以弥补这一不足。对 T 逻辑回归模型及求解算法进行了分析, 建立 T 逻辑回归财务预警模型, 并结合沪深上市公司财务数据开展实证分析, 结果表明 T 逻辑回归模型具有较好的分类效果和鲁棒性。

关键词: 财务危机预警; 正则化技术; T -逻辑回归; 过拟合; 非凸损失函数

中图分类号: TP311.13

文献标识码: A

DOI:10.3969/j.issn.1674-8085.2015.02.002

FINANCIAL EARLY-WARNING MODEL OF LISTED COMPANIES USING T -LOGISTIC REGRESSION

* XU Zheng¹, LIU Zun-xiong²

(1. School of Electrical and Electronic Engineering; 2. School of Information Engineering, East China Jiaotong University, Nanchang, Jiangxi 330013, China)

Abstract: The classic logistic regression has the risk of over fitting. It can be solved by the regularization technique of the statistical learning theory. Optimization of convex loss function can ensure that the regularized risk minimization problem converges to the global optimum, but learning algorithm of convex loss function is susceptible to noise. Then T -logistic regression was proposed to amend, introducing T distribution into logistic regression. The non-convex loss function is made up for the deficiency of convex loss functions. Due to the non-convex loss function difficulty to solve, we will be logarithmic the objective function, and convex multiplicative programming is used to solver parameters. Through empirical study, it is found that T -logistic regression model has a good predictability and is tolerant to label noise.

Key words: financial early-warning model; regularization technique; financial T -logistic regression; over fitting; non-convex loss function

0 引言

随着我国证券市场的迅速发展, 上市公司的质量会影响资本市场的健康发展。财务危机预警可以提前揭示风险, 以便采取有效措施, 防范和化解危机^[1-3]。上市公司财务预警是通过研究和分析财务状况指标反映上市公司财务经营的相关数据, 构建金融危机预警模型。由于财务预警指标变量多、相关

性大, 不利于建模分析, 所以通常需要进行变量选择, 然后对独立变量选择模型进行训练, 利用统计学习的稀疏矩阵建模, 并对上市公司的财务预警进行分析, 从而提高模型的解释和预测精度。

经典的逻辑回归对变量的分布没有要求并且预报精度高、解释能力强^[4], 但是与多元线性回归模型一样存在过拟合的问题, 需要通过统计学习理论中的正则化技术来解决^[5]。优化凸性函数能保证

收稿日期: 2014-11-17; 修改日期: 2015-01-20

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项目(13YJC630192); 华东交通大学校立科研课题(09DQ04)

作者简介: *徐 征(1978-), 女, 湖北随州人, 讲师, 硕士, 主要从事数理统计和非线性系统分析与建模研究(E-mail:187391733@qq.com);
刘遵雄(1965-), 男, 江西南昌人, 教授, 博士, 主要从事机器学习和金融统计时序分析研究(E-mail: 153010729@qq.com)。

正则化风险最小化问题收敛到全局最优,但是凸损失函数的学习算法对噪音不具备鲁棒性,因此需要引入非凸损失函数来弥补凸损失函数的不足。

本文将 t 指数族中的学生分布引入逻辑回归,对应的正则化风险最小化问题的损失函数具有非凸性。但是这也导致了正则化优化的风险^[6]。因此将目标函数对数化,采用凸多乘积规划,对目标跟随函数最小化求解参数。 T 逻辑回归能较好处理类别噪音,具有优越的性能。

1 T 逻辑回归理论

给定训练样本集 $(X, Y) = \{(x_1, y_1), \mathbf{L}, (x_m, y_m)\}$, 其中 x_i 取自特定域, $y_i \in \{\pm 1\}$, 样本 x 的特征采用特征映射 f 进行抽取, 假设存在 q 参数的条件分布族, 数据服从标准独立同分布条件, 根据贝叶斯公式^[6-8]有:

$$p(q|X, Y) = p(q) \prod_{i=1}^m p(y_i | x_i, q) / p(Y|X) \propto p(q) \prod_{i=1}^m p(y_i | x_i, q) \quad (1)$$

T 逻辑回归规定 $p(y|x, q)$ 服从条件 t 指数族^[9]分布,

$$p(y|x, q) = \exp_t(\langle f(x, y), q \rangle - g_t(q|x)) \quad (2)$$

其中 $1 < t < 2$ 。

学生分布^[10-12]可以看作 t 指数族分布中相对于正态分布先验的形式, 一维学生分布定义为:

$$St(x|m, s, v) = \frac{\Gamma((v+1)/2)}{\sqrt{vp} \Gamma(v/2) s^{1/2}} \left(1 + \frac{(x-m)^2}{vs}\right)^{-(v+1)/2} \quad (3)$$

其中 $\Gamma(\cdot)$ 为伽马函数, $v > 1$, 其均值为有限值。选择 t 使之满足 $-(v+1)/2 = 1/(1-t)$, 令

$$\Psi = \left(\frac{\Gamma((v+1)/2)}{\sqrt{vp} \Gamma(v/2) s^{1/2}}\right)^{-2/(v+1)}$$

则有:

$$St(x|m, s, v) = \exp_t(-\frac{1}{2} \frac{(x-m)^2}{vs} - \frac{1}{t}) \quad (4)$$

其中 $\frac{1}{2} = \frac{2\Psi}{(t-1)vs}$, $\frac{1}{t} = \frac{\Psi-1}{t-1}$ 。

假定 q 先验概率服从学生分布, 则

$$p(q) = \prod_{j=1}^d p(q_j) = \prod_{j=1}^d St(q_j | 0, 2/l, (3-t)/(t-1)) \quad (5)$$

当用学生 t 指数先验, $-\log p(q)$ 不再是凸的, 但这也导致存在优化正则化风险挑战。由于 t 对数函数 $\log_t$ 相对于 $\log$, 是一个单调递增函数, 因此可以利用 $\log_t$ 函数。

$$\log_t(x) = \frac{1}{1-t} (x^{1-t} - 1), \quad u > 0 \quad (6)$$

则最小误差的跟随目标函数 $\hat{J}(q)$ 可以写成:

$$\begin{aligned} \hat{J}(q) &= -\log_t p(q) \prod_{i=1}^m p(y_i | x_i; q) / p(Y|X) = \\ &= \frac{1}{t-1} (p(q) \prod_{i=1}^m p(y_i | x_i; q) / p(Y|X))^{1-t} + \frac{1}{1-t} \end{aligned} \quad (7)$$

其中, $p(Y|X)$ 相对 q 独立。利用式 (2)、(3) 和式 (7), 并令:

$$r_j(q) = (1 + (1-t)(-\frac{1}{2} q_j^2 / 2 - \frac{1}{t})) \quad (8)$$

$$l_i(q) = 1 + (1-t) \left(\left\langle \frac{y_i}{2} f(x_i), q \right\rangle - g_t(q | x_i) \right) \quad (9)$$

可以进一步得到:

$$\hat{J}(q) \propto \prod_{j=1}^d r_j(q) \prod_{i=1}^m l_i(q) + const \quad (10)$$

其中, $const$ 表示常数。当 $t > 1$ 时, 很容易发现 $r_j(q) > 0$ 是关于 q 的凸函数。另一方面, 由于 g_t 具有凸性, 且 $t > 1$ 可以得到 $l_i(q) > 0$ 也是关于 q 的凸函数。所以, $\hat{J}(q)$ 是正凸函数的结果。

2 求解策略及收敛性分析^[6]

根据凸多乘积规划^[14]有:

$$\min_q P((q)) = \prod_{n=1}^N z_n(q) \quad \text{s.t. } q \in R^d \quad (11)$$

其中, $z_n(q)$ 是正凸函数。设定 $N = d + m$, $z_n(q) = r_n(q)$ ($n = 1, \dots, d$) 且 $z_{n+d}(q) = l_n(q)$ ($n = 1, \dots, m$), 对问题 (11) 的优化解可通过求解下列参数问题来获得:

$$\min_x \min_q = \sum_{n=1}^N x_n z_n(q) \quad \text{s.t.} \\ q \in R^d, \quad x > 0, \quad \sum_{n=1}^N x_n \geq 0 \quad (12)$$

其中, x_n 为每个样本 x_n 的权重。求解(12)式的具体算法在文献[14]中提出, 但是计算的成本随着 N 的指数倍增加。因此可以分别将 q 和 x 作为变量做最小化分析, 用坐标下降法来求解。

2.1 以 x 为变量求解分析

假设 q 固定, 令 $\tilde{z}_n = z_n(q)$, 则(11)式可以表示为:

$$\min_x \sum_{n=1}^N x_n \tilde{z}_n \quad \text{s.t.} \quad x > 0, \quad \sum_{n=1}^N x_n \geq 1 \quad (13)$$

由于目标函数对 x 而言是线性的, 在可行区域是凸的, 所以(12)式可以用凸优化问题的方法求解。

引入非负拉格朗日乘数 $g \geq 0$, (13)式可以写成

$$L(x, g) = \sum_{n=1}^N x_n \tilde{z}_n + g(1 - \sum_{n=1}^N x_n) \quad (14)$$

相对于 x_n 的导数如下:

$$\frac{\partial}{\partial x_n} L(x, g) = \tilde{z}_n - g \prod_{n=1}^N x_n \quad (15)$$

令梯度为 0, 可以得到 $g = \frac{x_n}{\prod_{n \neq n'} x_n}$ 。当 $x_n > 0$ 时, g 不可能为 0。通过 K.K.T 条件^[15], 可以得到:

$$\prod_{n \neq n'} x_n = 1 \\ (x_1, \dots, x_N) = (g / \tilde{z}_1, \dots, g / \tilde{z}_N), g = \prod_{n=1}^N \tilde{z}_n^{-\frac{1}{N}} \quad (16)$$

而(14)式中的 x_n 是每项 $z_n(q)$ 的权。以上分析表明 $g = \tilde{z}_n x_n$ 相对所有的 n 都是常数。如果 $\tilde{z}_n(q)$ 变得非常大, x_n 的影响就会减少。因此大损失点存在权重上限, 这说明这个算法对异常值的鲁棒性非常好。

2.2 以 q 为变量求解分析

令 $x > 0$, 求解最佳 q 。这一步和回归分析一样, 但是每一成员都有一个权 x , 有:

$$\min_q \sum_{n=1}^N x_n z_n(q) \quad \text{s.t.} \quad q \in R^d \quad (17)$$

这是一个标准的无约束的凸优化问题, 它可以被任何现成的求解器求解, 采用 L-BFGS 高斯-牛顿方法。这样需要求解梯度 $\nabla_q z_n(q)$, 其中 $n=1, \dots, d$:

$$\nabla_q z_n(q) = \nabla_q r_n(q) = (t-1) \tilde{l} q_n \cdot e_n \quad (18)$$

$$\begin{aligned} \nabla_q z_{n+d}(q) &= \nabla_q l_n(q) = \\ (t-1) & \left(\frac{y_n}{2} f(x_n) - \nabla_q g_t(q | x_n) \right) = \\ (1-t) & \left(\frac{y_n}{2} f(x_n) - E_{q_t(y_n, x; q)} \left[\frac{y_n}{2} f(x_n) \right] \right) \end{aligned} \quad (19)$$

其中, e_n 表示 d 维向量包含一个在 n 阶坐标, 别的地方都是零 (即 n 阶单位矢量)。 $q_t(y | x; q)$ 是 $p(y | x; q)$ 的 escort 分布:

$$q_t(y | x; q) = \frac{p(y | x; q)^t}{p(+1 | x; q)^t + p(-1 | x; q)^t} \quad (20)$$

目标函数是单调下降, 保证收敛到一个稳定的点 $p(q)$ 。

3 实证分析

3.1 模型样本及财务指标的选择

首先选取沪深股市 A 股上市公司, 要求规模相当且上市五年左右。从 2005-2007 年选取了 51 家 ST 公司, 并按照大约 1:1 的比例选取了相同时间段的且上市以来从未被 ST 的 51 家正常公司, 其中 2005 年共 32 家、2006 年共 28 家、2007 年共 42 家。

参考国内外相关资料[5-8], 根据盈利能力、债务能力、成长能力、营运能力、现金流量和资本结构, 共选择了 26 个变量。自变量为 Y , 表示该公司是否会出现财务危机, 如果为 ST 公司则表示出现了财务危机, 用 0 表示; 如果没有出现财务危机, 则为非 ST 公司, 用 1 表示。

由于有些财务指标没有显著差异, 因此需要对样本数据进行显著性检验。本文采用 SPSS 对样本进行单样本 K-S 检验, 判断是否服从正态分布。通

过分析，发现大多数指标不服从正态分布。于是，采用非参数检验。非参数检验方法很多，有 Mann-Whitney U 检验、Kolmogorov-Smirnov Z 检验、Moses 极限反应和 Wald-Wolfowitz 游程检验^[16]。这里选用最常见的 Mann-Whitney U 检验。利用 SPSS 软件分析，将不能通过 Mann-Whitney U 检验的指标剔除，将剩下指标组成新的样本数据。

3.2 模型实证分析结果

实验平台为 MATLAB7.1。利用最终保留下来的 16 个财务预警指标数据作为模型的输入，训练样本与测试样本的比为 2:1，训练样本共 68 家，测试样本 34 家。

根据文献[6]可知非凸函数不容易受初值的影响，因此初值的选择可以有一定的随意性，选 $q = 0.015$ ， $x = 0.2$ 。经过 MATLAB 仿真，没有加噪声的训练样本与测试样本的误差数据如表 1，加入 10%噪声后的仿真数据如表 2。

表 1 噪声为 0，仿真误差表

Table 1 The error of the simulation when noise equals zero

模 型	训练样本 (%)		测试样本 (%)	
	第一类错误	第二类错误	第一类错误	第二类错误
L ₁ 正则化	0	5.88	22.22	16.67
L ₂ 正则化	2.94	2.94	16.67	16.67
$t = 1.2$	0	2.94	16.67	11.11
$t = 1.4$	2.94	2.94	11.11	16.67
$t = 1.6$	0	2.94	11.11	11.11
$t = 1.8$	0	5.88	16.67	11.11

由表 1 可知，在噪声为 0 的时候，用正则化模型和 T 逻辑回归模型的训练样本的正确率都非常高，但是对测试样本而言，T-logstic 模型的正确率要高于正则化逻辑回归模型 3~5 个百分点。可见 T 逻辑回归模型具有较好的预见性。

表 2 噪声为 10%，仿真误差表

Table 2 The error of the simulation when noise equals 10%

模 型	训练样本 (%)		测试样本 (%)	
	第一类错误	第二类错误	第一类错误	第二类错误
L ₁ 正则化	2.94	14.71	27.78	22.22
L ₂ 正则化	5.88	14.71	27.78	27.78
$t = 1.2$	0	5.88	16.67	16.67
$t = 1.4$	0	8.82	22.22	11.11
$t = 1.6$	2.94	5.88	16.67	11.11
$t = 1.8$	2.94	2.94	22.22	11.11

加入 10%的噪声后，整体的正确性都有所下降。L₁ 正则化模型的训练样本的正确率由 97.06% 下降到 91.18%，L₂ 正则化模型的训练样本的正确率由 97.06%下降到 89.71%，而 T 逻辑回归模型的正确率下降与 t 的选择有关，但总体下降幅度不大，仅下降了 0~3 个百分点，特别是在 $t = 1.8$ 的时候，正确率没有受到噪声的干扰。测试样本的波动相对而言更大，L₁ 正则化模型的训练样本的正确率由 82.35%下降到 76.46%，L₂ 正则化模型的训练样本的正确率由 85.29%下降到 73.53%，T 逻辑回归模型的训练样本的正确率也有所下降，但是幅度也不大，都维持在 85%以上。

加入噪声至 30%，训练样本与测试样本的误差变化趋势见图 1，通过对比可以观察到 T 逻辑回归模型对噪声的干扰有较强的鲁棒性。虽然 L₁，L₂ 正则化模型和 T 逻辑回归模型都随着噪声干扰的增加，误差率会提高，但是 T 逻辑回归的误差率要远低于正则化模型。随着噪声干扰的增加，测试样本的误差率增加幅度降低，T 逻辑回归的误差率与 t 的取值相关性降低。

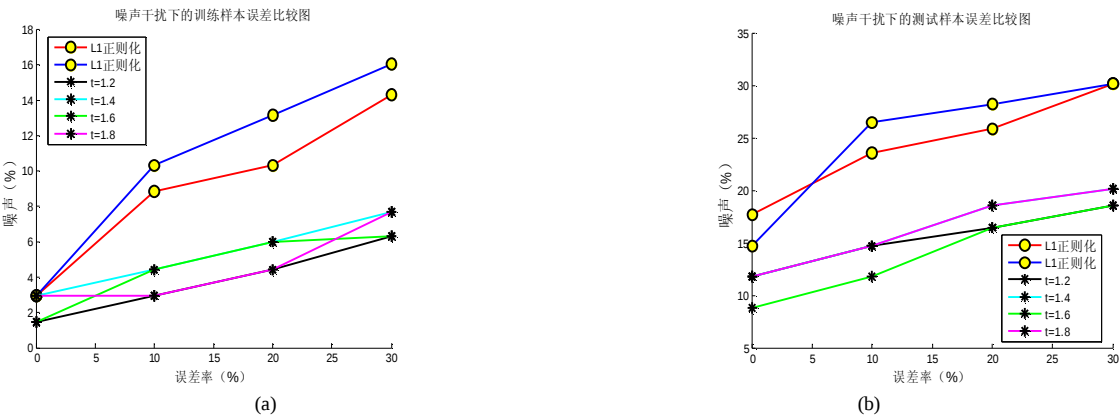


图 1 噪声干扰下的训练样本 (a 图) 与测试样本 (b 图) 的误差变化

Fig.1 The error rates of the training samples (a) and the testing samples (b) when noise equals 30%

4 结语

本文将 t 指数族中的学生分布引入逻辑回归,建立了上市公司的财务预警模型。经过仿真,将 T 逻辑回归模型与 L_1 和 L_2 正则化逻辑回归模型比较,证实该模型具有较好的预测功能,并有较强的抗噪声能力。在以后的进一步研究中,将考虑深入讨论 T 逻辑回归与初值选择的相关性,以及 t 的选择对 T 逻辑回归模型精度的影响。

参考文献:

- [1] 陈艳,张海君.上市公司财务预警模型的研究[J].财经问题研究,2007,(6):92-96.
- [2] 许珂,卢海.基于 Logistic 回归模型的房地产上市公司财务危机预警研究[J].常州工学院学报,2012(4):61-66.
- [3] 吴冬梅,朱俊,庄新田,等.基于支持向量机的财务危机预警模型[J].东北大学学报:自然科学版,2010,31(4):601-604.
- [4] 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,2012,27(12):21-26.
- [5] 刘遵雄,郑淑娟,秦宾,等. L_1 正则化 Logistic 回归在财务预警中的应用[J].经济数学,2012,29(2):106-110.
- [6] Ding N, Vishwanathan S V N. t -Logistic regression[C]. Advances in Neural Information Processing Systems, 2008:294.
- [7] Buscema M, Breda M, Lodwick W. Training with Input Selection and Testing (TWIST) Algorithm: A Significant Advance in Pattern Recognition Performance of Machine Learning[J]. Journal of Intelligent Learning Systems and Applications, 2013, 5: 29-38.
- [8] Hastie T, Tibshirani R, Friedman J, et al. The elements of statistical learning: data mining, inference and prediction[J]. The Mathematical Intelligencer, 2005, 27(2): 83-85.
- [9] De Souza A M C, Tsallis C. Student's t -and r -distributions: Unified derivation from an entropic variational principle[J]. Physica A: Statistical Mechanics and its Applications, 1997, 236(1): 52-57.
- [10] Naudts J. Generalized thermostatistics based on deformed exponential and logarithmic functions[J]. Physica A: Statistical Mechanics and its Applications, 2004, 340(1): 32-40.
- [11] Naudts J. Generalized thermostatistics and mean-field theory[J]. Physica A: Statistical Mechanics and its Applications, 2004, 332: 279-300.
- [12] Naudts J. The q -exponential family in statistical physics[J]. Central European Journal of Physics, 2009, 7(3): 405-413.
- [13] Dawson R. Wittgenstein on pure and applied mathematics[J]. Synthese, 2014, 191(17): 4131-4148.
- [14] Kuno T, Yajima Y, Konno H. An outer approximation method for minimizing the product of several convex functions on a convex set[J]. Journal of Global optimization, 1993, 3(3): 325-335.
- [15] Zhou W J, Chen X L. On the convergence of a modified regularized Newton method for convex optimization with singular solutions[J]. Journal of Computational and Applied Mathematics, 2013, 239: 179-188.
- [16] 张文彤.世界优秀统计工具 SPSS 11 统计分析教程高级篇[M]. 北京:北京希望电子出版社,2002:S91-92.