DOI:10.19343/j.cnki.11-1302/c.2016.04.007

基于网络结构 Logistic 模型的企业信用风险预警*

方匡南 范新妍 马双鸽

内容提要:随着计算机和互联网的快速发展,特别是在大数据时代,企业积累了大量有关企业经营、财务等相关数据,变量众多且关系纷繁复杂,如果利用传统的 logistic 回归建立企业信用风险预警模型往往效果不好。本文在充分考虑变量间的网络结构(Network)关系基础上,提出了网络结构 Logistic 模型,通过惩罚方法同时实现变量选择和参数估计。蒙特卡洛模拟表明网络结构 Logistic 模型要优于其他方法。最后,我们将其应用到我国企业信用风险预警中,充分考虑财务指标间的网络结构关系,科学地选择评估指标,构建更加适合我国国情的企业信用风险预警方法。

关键词:企业信用风险;网络结构;logistic 模型
中图分类号:C812 文献标识码:A 文章编号:1002-4565(2016)04-0050-06

Forecasting of Enterprise's Credit Risk Based on Network-logistic Model

Fang Kuangnan Fan Xinyan Ma Shuangge

Abstract: With the rapid development of computer and the Internet, especially in the era of big data, some enterprises has accumulated a lot about their operation and finance data. Since the data is numerous and complicated, if we use the traditional logistic regression to build up the enterprise credit risk, the performance usually isn't good. In this paper, we propose network-logistic model based on considering the network relationship among variables, via penalized method to conduct variable selection and parameters estimation simultaneously. Simulation results show that network-logistic model performs better than other compared methods. Finally, we apply it to forecast enterprise's credit risk, under considering the network relationship between financial indicators, select significant variables and build up a suitable credit risk forecasting model for Chinese enterprises.

Key words: Enterprise's Credit Risk; Network; Logistic Model

一、引言

上市公司信用风险预警是通过财务比率数据分析预测企业出现财务危机的可能性。从方法角度来看,信用风险预警方法主要有多元线性判别分析、机器学习、Logistic 回归等,但是这些方法均存在不同程度的缺陷。多元线性判别分析对预测变量有着严格的联合正态分布要求,或者要求协方差矩阵相等,然而大量实证结果表明多数财务比率数据并不满足这一假设条件。机器学习模型除存在过度拟合问题外,需要大量样本数据,而企业的信用风险数据由于其特殊性搜集较为困难。对于传统的 Logistic 模型,随着计算机和互联网的发展,企业的信息纷繁复杂、变量众多,对建模带来较大的难度,此外,各财务指标之间的关系也错综复杂,彼此之间往往呈网络结构关系。本文在充分考虑变量间的网络结构关系基础上,提出了网络结构 Logistic 模型,通过惩罚方法同时实现变量选择和参数估计,并将其应用到我国企业信用风险预警中,充分考虑企业财务指标间的网络结构关系,科学地选择评估指标,以期构建更加

* 本文获国家自然科学基金面上项目"广义线性模型的组变量选择及其在信用评分中的应用"(71471152)、国家社会科学基金重大项目"大数据与统计学理论的发展研究"(13&ZD148)和国家社会科学基金青年项目"大数据的高维变量选择方法及其应用研究"(13CTJ001)的资助。

适合我国国情的企业信用风险预警方法。

二、文献综述

从 1966 年 Beaver 利用单一的财务比率来预测财务状况起,公司信用风险分析已经有近 50 年的历史。Altman(1968) [1] 率先应用多元判别分析的方法对美国企业破产进行预测。但应用线性判别模型的条件为总体服从正态分布且协方差矩阵相等,这些条件在实际中很难得到满足。Ohlson(1980) [2] 以美国的 105 家破产公司和 2058 家正常公司为样本,建立了 Logit 财务困境预测模型并通过实证研究证明 Logit 模型预测效果好于多元线性判别分析。但由于财务指标间的多重共线性,Logit 模型在变量选择时存在限制。为解决指标间共线性问题,Aguilera 等(2006) [3] 将主成分分析与 Logistic 回归相结合来预测企业违约问题。但主成分的实际意义难以解释。此外,机器学习技术被广泛地应用到模型的建立过程中,Franco Varetto(1998) 运用遗传算法研究企业破产风险, Min 和 Lee(2005) 将支持向量机的方法运用到上市公司信用风险预测上。然而这些方法依旧存在弊端,其运算复杂,要求大量的训练样本,且存在过度拟合的危险。

国内方面,赵健梅和王春莉(2003) [4] 选取了 40 家 ST 公司和非 ST 公司作为样本,采用 Z-score 方法对上市公司财务危机预警问题进行了实证研究。鲜文铎和向锐(2007) [5] 基于混合 Logit 模型对 A 股上市公司进行预测,放宽了传统标准 Logit 模型的个体选择偏好同质性和不相关备选方案独立性两方面的限制。韩立岩和李蕾(2010) [6] 针对中小上市公司建立了财务危机判别模型。邓晶等(2013) [7] 将因子分析与 Logistic 模型相结合对上市公司信用风险进行预测。王小燕等(2014) [8] 提出了 adSGL-logit 信用评分模型。

Logistic 模型由于其计算简单、系数易解释等特点在实际中使用广泛。但是随着计算机和互联网的快速发展,特别是在大数据时代,企业获取数据越来越方便、快捷,很多企业积累了大量有关企业经营和财务等相关数据,变量众多而且变量间关系纷繁复杂,如果利用传统的 logistic 回归建立企业信用风险预警模型往往效果不好。究其原因主要有: 首先,Logistic 模型中如果包含过多的变量,一方面由于多重共线性等问题可能会降低模型的预测准确性,另一方面模型中选入一些无关变量,会浪费人力物力搜集这些信息; 其次,传统 Logistic 回归以及变量选择方法都忽略了变量间的网络结构关系。因此,如何选择合适的变量是大数据时代下的企业信用风险预警的重点和难点。关于变量选择,目前最常用的是惩罚方法(Penalization) 。国内外学者就利用惩罚函数进行高维回归模型变量筛选问题做了大量的研究。最早的惩罚函数法是由 Hoerl 和 Kennard (1970) 提出的岭回归(ridge regression) 。随后 Frank 和 Fredman (1993) 提出了桥回归(bridge regression) 。Tibshirani(1996) 提出了 LASSO 方法,该方法保留了最优子集的优点,可以同时进行变量选择和参数估计。随后,Fan 和 Li(2001) [9] 提出了 SCAD 法,Zhang(2007) [10] 提出 MCP 变量选择方法,对 LASSO 估计的有偏性进行了改进。然而上述方法在进行变量选择时都将对变量的惩罚与变量之间的相依关系看作是独立的,忽略了变量间的网络结构关系,倾向于在一组高度相关的变量中只选出一个变量,变量之间的高度相关性可能会影响变量选择的效果(Zou 和 Hastie; 2005) 。Huang 等(2011) [11] 在高维的线性回归变量选择中考虑变量间网络结构关系,认为这有助于提高变量选择和预测效果。本文将在原有 MCP 惩罚函数的基础上对变量之间的网络结构关系进行惩罚,提出了网络结构 Logistic 模型,并将其应用到我国企业信用风险的预警中。

三、网络结构 Logistic 模型

(一) 网络结构 Logistic 模型介绍

假设有独立同分布的观测值 (x_i, y_i), $i = 1, 2, \cdots, n$,其中 x_i 是解释变量,y_i 是二元离散被解释变量,即 $y_i \in \{0, 1\}$,则 Logistic 线性回归模型为:

$$\log\left\{\frac{p_\beta(x_i)}{1 - p_\beta(x_i)}\right\} = \eta_\beta(x_i)$$

其中,$\eta_\beta(x_i) = \beta_0 + \sum_{i=1}^{p} x_i^T\beta_i + \varepsilon_i$, $i = 1, 2, \cdots, n$。

采用网络结构 Logistic 模型对 β 进行估计:

$$\hat{\beta} = \arg\min_\beta\left\{\frac{-l(\beta)}{n} + P_{\lambda_1,\lambda_2,\gamma}(\beta)\right\} \quad (1)$$

其中,$l(\beta)$ 是 Logistic 回归的似然函数,即:

$$l(\beta) = (X\beta)^T y - 1_n^T\log[1_n + \exp(X\beta)] \quad (2)$$

$P_{\lambda_1,\lambda_2,\gamma}(\beta)$ 是由 MCP 惩罚和网络结构惩罚两部分构成的惩罚函数,即:

$$P_{\lambda_1,\lambda_2,\gamma}(\beta) = \sum_{j=1}^{p}\rho(|\beta_j|;\lambda_1,\gamma) + \frac{1}{2}\lambda_2\sum_{1\leqslant j\leqslant k\leqslant p}|a_{ij}|(\beta_j - s_{jk}\beta_k)^2 \quad (3)$$

式(3)中，$\rho(t;\lambda_1,\gamma) = \lambda_1\int_0^{|t|}[1 - x/(\gamma\lambda_1)]_+ dx$ 为 MCP 惩罚项。关于 MCP 方法的详细介绍见 Zhang(2010)[12]。式(3)的第 2 项非负二次型为网络结构惩罚项，其中 a_{ij} 为自变量之间网络结构关系的一种度量，即相邻矩阵(Adjacency Matrix)的元素，$s_{ij} = \mathrm{sgn}(a_{ij})$。MCP 惩罚项是对回归系数稀疏性的惩罚，通过控制 λ_1 和 γ，对回归系数 $\hat{\beta}$ 进行压缩。随着 λ_1 的增大，$\hat{\beta}$ 逐渐被压缩至 0。网络结构惩罚项的主要作用是对回归系数进行平滑。根据 Huang 等(2011)[11]的研究结论，可用自变量协方差矩阵的 3 次幂表示自变量之间的网络结构关系，即 $(a_{ij})_{ij=1,\cdots p} = (corX)^3$。网络结构惩罚项使正相关的自变量的回归系数趋同，而使负相关变量的回归系数符号存在相异趋势。

(二)回归系数 $\hat{\beta}$ 的估计

本文采用坐标下降法(Coordinate Descent，简称 CD)对参数进行估计。该算法每次变化 $\hat{\beta}$ 中的一个系数 β_i 而固定其他系数 $\beta_j(j\neq i)$，$j = 1,2,\cdots,p$ 不变，寻找 β_i 的最优值使目标函数达到最小。遍历每一回归系数寻找最优 $\hat{\beta}$。重复上述过程直到 $\hat{\beta}$ 收敛。为了与 CD 算法对应，本文对目标函数做如下整理：

$$R(\beta_k) = -l(\beta)/n + \rho(|\beta_k|;\lambda_1,\gamma) + \frac{1}{2}\lambda_2\sum_{1\leqslant j<k\leqslant p}|a_{jk}|(\beta_k - s_{jk}\beta_j)^2 \quad (4)$$

其中：

$$-l(\beta)/n = \frac{1}{n}(1_n^T\log[1_n + \exp(X_{-k}\beta_{-k} + X_k\beta_k)]) - \frac{1}{n}\beta_k X_k^T y + c$$

$$\rho(|\beta_k|;\lambda_1,\gamma) = \left(\lambda_1|\beta_k| - \frac{\beta_k^2}{2\gamma}\right)I(|\beta_k| < \lambda_1\gamma) + \frac{\lambda_1^2\gamma}{2}I(|\beta_k| > \lambda_1\gamma)$$

$$\frac{1}{2}\lambda_2\sum_{1\leqslant j<k\leqslant p}|a_{jk}|(\beta_k - s_{jk}\beta_j)^2 = \frac{1}{2}\lambda_2\left[\sum_{j\neq k}|a_{jk}|\beta_k^2 - 2\beta_k\sum_{j\neq k}a_{jk}\beta_j\right]$$

则 CD 算法可表示为：

(1)初始化 β，令 $\beta = (0,\cdots,0)$。

(2)对于 $k = 0,1,\cdots,p$

若 $\left|x_k^T(y - p_{\beta\_k}) + \lambda_2\sum_{j\neq k}a_{kj}\beta_j\right| < \lambda_1$，则 $\beta_k = 0$。否则，$\beta_k = \mathrm{argmin}[R(\beta_k)]$。

(3)重复步骤(2)直到该过程收敛。

其中，p_{β_k} 为 $\beta_k = 0, y = 1$ 的概率值，$\left|x_k^T(y - p_{\beta\_k}) + \lambda_2\sum_{j\neq k}a_{kj}\beta_j\right| < \lambda_1, \beta_k = 0$ 为 KKT 条件。

(三)调和参数 λ_1、λ_2 的选择

考虑到传统的 CV(Cross Validation)参数选择方法计算量太大，以及基于 AIC/BIC 准则将忽略掉网络结构惩罚项使 λ_2 趋于 0 的情况，本文提出了双层参数选择法。具体方法如下：

(1)设定 λ_2 取值范围，在每一个 λ_2 值下采用 AIC/BIC 准则选择最优 λ_1 的值，形成参数对(λ_1，λ_2)；

(2)采用 CV 参数选择法，选择最优参数对。

双层参数选择方法既避免了 AIC/BIC 准则对参数 λ_2 处理上的偏误，相比 CV 参数选择法又减少了计算复杂度。本文以 k-fold 为例说明计算复杂度的减少，采用 CD 算法重复次数作为计算复杂度的度量。设备选参数 λ_1 的个数为 L_1，备选参数 λ_2 的个数为 L_2，对于单纯的 k-fold 方式选择参数，计算复杂度为 kL_1L_2，而采用双层变量选择法计算复杂度为 $L_2(L_1 + k)$。

对于正则化参数 γ，就 MCP 模型而言 Zhang(2010)[12]建议采用 $\gamma = 2/(1 - \max_{j\neq k}|x_j'x_k|/n)$，而在 Breheny 和 Huang(2011)[13]的模拟中建议 $\gamma = 3$，并且该文还试了几个不同的值，得出的结论基本是一样的。本文取 $\gamma = 5$，同时本文尝试取几个不同的值，结果基本是一致的。

四、模拟实验

本文通过蒙特卡罗模拟方法比较网络结构 Logistic 模型、MCP Logistic 模型、SCAD Logistic 模型、LASSO Logistic 模型的优劣。数值分析模型为：

$$\log\left\{\frac{p(Y = 1\mid X)}{1 - p(Y = 1\mid X)}\right\} = \eta_\beta(X) = X^T\beta \quad (5)$$

本文共进行了两组模拟，分别设置了两组不同的真实系数结构，详见例 1 和例 2。本文利用双层

参数选择法选择调和参数,其备选集合为 $\lambda_1 \in \{k \cdot 10^l: k = 1,3,5,\cdots,9; l = \cdots, -1,1,\cdots\}$,$\lambda_2 \in \{k \cdot 10^l: k = 1,2,3,\cdots 9; l = \cdots, -1,1,\cdots\}$ 。由于模型结果对正则化参数不敏感,考虑计算的简便性取正则参数 $\gamma = 5$。取样本容量 $n = 100$,每种情况重复 100 次试验。为了比较本文提出的网络结构 Logistic 模型、传统的 MCP Logistic 模型、SCAD Logistic 模型和 LASSO Logistic 模型的优劣,本文所选取的评判标准有: 所选显著变量个数(num) 1000 个样本外的错误识别率(ER) ,显著变量的错误发现率(FDR) 、假阴性率(FNR) 。

例 1: 设除常数项外自变量个数为 p,$p \in \{50, 100, 200\}$,真实回归系数取 $\beta = (0,1,1,\cdots,1,0,\cdots, 0)$,真实模型的显著变量个数为 25 个,且取值都为 1。除常数项外,自变量服从标准正态分布,且每 5 个自变量为一组,组内自变量 x_i 与 x_j 之间的相关系数为 $\rho^{|i-j|}$,$\rho \in \{0.5, 0.9\}$,组间变量相互独立①。

例 2: 设除常数项外自变量个数为 p,$p \in \{50, 100, 200\}$,真实回归系数取 $\beta = (0,1,\cdots,1, -3, \cdots, -3,1,\cdots,1, -3,\cdots, -3,1,\cdots,1,0,\cdots,0)$ 。除常数项外,共 25 个显著变量,每 5 个为 1 组,第 1 组、第 3 组、第 5 组变量的系数均为 1,第 2 组、第 4 组变量的系数均为 -3,其余变量系数均为 0。自变量服从标准正态分布,组内自变量 x_i 与 x_j 之间的相关系数为 $\rho^{|i-j|}$,$\rho \in \{0.5, 0.9\}$,组间变量相互独立②。

例 1、例 2 的模拟结果反映出网络结构 Logistic 模型识别的显著变量个数较其他模型更接近真实情况(真实显著变量数是 25 个) ,LASSO 选择的变量数往往过多,而 MCP 和 SCAD 选择的变量数往往偏少。与其他模型相比,网络结构 Logistic 模型的因变量错误识别率(ER) 、显著变量的错误发现率(FDR) 、假阴性率(FNR) 都是最低的,尤其是当变量间存在高度相关性时,网络结构 Logistic 模型的优越性更加突出,相应的 ER、FDR、FNR 都远远低于其他方法,即说明网络结构 Logistic 模型在变量选择上远远好于其他方法。随着变量数量 p 的增大,网络结构 Logistic 模型比其他方法表现得更稳定。

五、企业信用风险预警分析

以往的研究多以“特殊处理公司/正常公司”作为财务困境/非财务困境的代表样本[4,5]。除经特殊处理公司外,目前还难以从公开的报告中获得较多的其他类型的财务困境公司样本(如破产公司) ,所以本文仍沿用以特殊处理的公司作为财务困境公司样本的方式。考虑到样本收集的难度,本文仅以被* ST 的上市公司作为财务困境公司的样本。

(一) 财务指标的选取

不同的财务指标从不同侧面反应企业的财务状况和经营业绩。在结合已有文献的基础上[4,5],本文共选择了涵盖每股指标、盈利能力、偿债能力、成长能力、营运能力、资产结构等方面的 48 个指标。变量的分类、名称和符号详见表 1。因变量定义为 0 -1 分类变量: 1 为* ST 公司,0 为非* ST 公司。本文研究的财务指标数据与* ST 公司相关资料均来自于 RESSET 数据库。

表 1　　财务指标表

指标分类	指标名称	变量符号	指标分类	指标名称	变量符号
每股	每股收益	X_1	营运	营业周期(天/次)	X_{26}
指标	每股净资产	X_2	能力	存货周转率(次)	X_{27}
	每股营业收入	X_3		存货周转天数(天/次)	X_{28}
	每股营业利润	X_4		应收账款周转率(次)	X_{29}
	每股资本公积金	X_5		应收账款周转天数(天/次)	X_{30}
	每股盈余公积金	X_6		应付账款周转率(次)	X_{31}
	每股公积金	X_7		应付账款周转天数(天/次)	X_{32}
	每股未分配利润	X_8		流动资产周转率(次)	X_{33}
	每股留存收益	X_9		固定资产周转率(次)	X_{34}
	每股经营活动现金流量	X_{10}		股东权益周转率(次)	X_{35}
	每股净现金流量	X_{11}		总资产周转率	X_{36}
资产	资产负债率	X_{12}	成长	每股收益增长率	X_{37}
结构	非流动资产/总资产	X_{13}	能力	营业收入增长率	X_{38}
	固定资产比率	X_{14}		净利润增长率	X_{39}
	股东权益/全部投入资本	X_{15}		净资产增长率	X_{40}
	权益乘数	X_{16}		资产总计相对年初增长率	X_{41}
	营运资金	X_{17}			
盈利	净资产收益率	X_{18}	偿债	流动比率	X_{42}
能力	资产报酬率	X_{19}	能力	速动比率	X_{43}
	资产净利率	X_{20}		产权比率	X_{44}
	投入资本回报率	X_{21}		股东权益/负债合计	X_{45}
	销售净利率	X_{22}		有形净值债务率	X_{46}
	销售成本率	X_{23}		经营净现金流量/负债合计	X_{47}
	营业总成本/营业总收入	X_{24}		经营净现金流量/流动负债	X_{48}
	净利润	X_{25}			

各财务指标之间的相关关系错综复杂,设 ρ 为判定变量之间存在相依关系的临界值,当相关系数小于 ρ 时,则认为变量之间关系不重要。

① 限于篇幅,模拟结果略去,有需要的读者可向作者索取。
② 限于篇幅,模拟结果略去,有需要的读者可向作者索取。

(二) 准确率对比与变量选择

将财务指标进行标准化之后,利用传统的全变量 Logistic 模型、LASSO Logistic 模型、MCP Logistic 模型、SCAD Logistic 模型和本文提出的网络结构 Logistic 模型进行建模与预测分析。首先,将样本数据按 3:1 比例进行划分,每次随机抽取用 75 个样本作为训练样本,剩余 25 个样本作为测试样本进行样本外预测检验。将该过程重复 100 次,计算模型的平均预测准确率,结果显示,传统的全变量 Logistic 回归的平均预测准确率远远低于基于惩罚项的 MCP、SCAD、LASSO 和网络结构 Logistic 回归,而且标准差也是最大的,这说明将传统的全变量 Logistic 回归直接应用到企业信用分析预警上往往效果欠佳。网络结构 Logistic 回归的预测准确率是所有方法中最高的,说明考虑了变量间网络结构关系可以大大提高 Logistic 回归的预测准确率。然后利用网络结构 Logistic 回归对全部 100 家上市公司的信用情况进行预警分析,变量的筛选结果和对应的系数估计结果见表 2。从所选指标所属分类上看,每股指标、盈利能力指标、成长能力指标对财务困境的预测是显著的。每股收益、每股净资产、每股营业利润、每股资本公积金、每股公积金、每股未分配利润、每股留存收益越小,则企业遇到财务危机的可能性越大。赵健梅、王春莉(2003)[14] 通过对 40 家 ST 企业的分析得到每股净资产是反应企业财务危机状况的显著性指标,佐证了本文所得指标的显著性。净资产收益率、资产报酬率、资产净利率、投资回报率、销售净利率越高,销售成本率、营业成本/营业收入越低,说明企业的盈利能力越强,则企业面临财务危机的可能性越小。陈静(1999)[15] 提出净资产收益率显著反应企业财务状况,说明盈利能力对企业的财务状况影响较大,与本文结论类似。此外,本文认为反映企业成长能力的每股收益增长率、净资产增长率、资产总计相对于年初增长率都是评价企业财务状况的重要指标。其中净资产增长率同样被杨海军和太雷(2009)[16] 选为预测上市公司财务困境的重要指标。

(三) 稳健性分析

现实中,大企业和中小企业在信用风险特征上可能有所不同。比如,大企业和中小企业的资本结构不同,以及中小企业相对于大企业借贷较难等。因此,本文参考工信部联合企业(2011)300 号文件确定的中小企业的认定标准来划分大型企业与中小企业。我们发现只有城城股份、新都酒店、华东数控三家被划分为中小型企业,因此,我们把剩下的 22 家大型上市公司按照采用 1:3的比例作为财务困境公司与正常公司的配对比,共收集了 88 家大型企业的财务比率数据。然后,每次随机从中抽取 66 个样本作为训练样本,剩余 22 个样本作为测试样本进行样本外预测检验。将该过程重复 100 次,计算模型的平均预测准确率。表 3 是大型企业的显著变量个数及模型预测准确率,表 4 是利用网络结构 Logistic 回归对全部 88 家大型上市公司的变量的筛选结果和对应的系数。从表 3 和表 4 可以看出,网络结构 Logistic 回归的预测效果也是最好的。说明本文提出的网络结构 Logistic 回归的预测具有很好的稳健性。由于本文收集到的数据里只有 3 家中小企业上市公司,样本量过小,没法针对中小企业的财务困境进行分析。但我们相信本文提出的网络结构 Logistic 回归方法同样适合于中小企业的财务困境预警。

表 2 网络结构 logistics 模型系数估计结果

变量	系数	变量	系数	变量	系数
X0	-2.7439	X8	-0.2548	X22	-0.3462
X1	-0.3307	X9	-0.2354	X23	0.3265
X2	-0.3226	X18	-0.3583	X24	0.3339
X4	-0.3248	X19	-0.3491	X37	-0.3252
X5	-0.3948	X20	-0.3459	X40	-0.3838
X7	-0.3784	X21	-0.3481	X41	-0.5139

表 3 大型上市公司显著变量个数及准确率

	全变量	LASSO	SCAD	MCP	网络结构
显著变量个数	49	13.90	5.94	5.48	12.71
	-	(2.15)	(1.15)	(1.04)	(6.90)
准确率	66.40	88.50	90.46	90.17	90.87
	(10)	(6.4)	(5.1)	(4.7)	(5.3)

表 4 大型上市公司网络结构 logistics 模型系数估计结果

变量	系数	变量	系数	变量	系数
X0	-2.4554	X9	-0.2138	X24	0.2716
X1	-0.2704	X18	-0.2967	X25	-0.2926
X2	-0.2488	X19	-0.2903	X37	-0.2939
X4	-0.2670	X20	-0.2824	X38	-0.2687
X5	-0.2754	X21	-0.2876	X39	-0.3240
X7	-0.2700	X22	-0.2782	X40	-0.2952
X8	-0.2252	X23	0.2628	X41	-0.2845

六、结论

本文在考虑变量间网络结构关系基础上提出了

网络结构 Logistic 回归,具有同时实现变量选择和系数估计的特点,并将该方法应用到我国上市公司信用风险预警中,充分考虑各财务比率指标之间的网络结构关系,对企业信用风险进行预测。本文的主要结论有: 第一,本文构建了针对二元离散变量的网络结构 Logistic 模型,该方法在进行系数压缩时充分考虑了变量之间的网络结构,使变量筛选更具科学性,并且为降低该方法的计算复杂性,本文提出了双层变量选择法,降低了计算难度。第二,根据蒙特卡罗模拟结果,当变量之间存在紧密相依关系时,网络结构模型比 MCP、SCAD、LASSO 模型的变量选择和预测效果更好,尤其是当变量间相关系数很大时,网络结构 Logistic 模型的变量选择和预测结果表现更为突出。第三,本文对我国企业财务危机预测的实证分析中,发现财务指标之间存在显著的网络结构关系,传统的全变量 Logistic 模型表现差于其他方法,说明现在企业最常用的全变量 Logistic 企业信用预警方法是有问题的,而网络结构 Logistic 模型预测准确率是最高的,通过该模型的分析发现每股指标、盈利能力指标、成长能力指标是影响企业信用风险的主要因素。第四,我们将上市公司进一步细分为大型上市公司和中小型上市公司,发现网络结构 Logistic 模型对大型上市公司的信用风险预测准确率也是最高的,说明该方法在信用风险预测方面具有较好的稳健性。

此外,本文虽然主要研究基于 MCP 惩罚下的网络结构 Logistic 模型,但是该方法同样可以扩展到 Poisson 回归、有序 Logistic 回归、条件 Logistic 回归等其他广义线性模型中,同时也可以在广义线性模型中考虑其他惩罚方法以及研究这些方法在不同经济管理领域中的应用,这将是我们下一步的研究方向。

参考文献

[1] Edwad I Altman. Financial Ratios, Discriminate analysis and the prediction of corporate bankruptcy [J]. Journal of Finace, 1968, 23(4): 589 – 609.

[2] James A Ohlson. Financial ratios and the probabilistic prediction of bankruptcy [J]. Journal of Accounting Research, 1980, 18(1): 109 – 131.

[3] Ana M Aguilera, Manuel Escabias, Mariano J Valderrama. Using principal components for estimating logistic regression with high-dimensional multicollinear data [J]. Computational Statistics & Data Analysis, 2006, 50(8): 1905 – 1924.

[4] 赵健梅,王春莉. 财务危机预警在我国上市公司的实证研究[J]. 数量经济技术经济研究,2003(7): 134 – 138.

[5] 鲜文铎,向锐. 基于混合 Logit 模型的财务困境预测研究[J]. 数量经济技术经济研究,2007(9): 68 – 76.

[6] 韩立岩, 李蕾. 中小上市公司财务危机判别模型研究[J]. 数量经济技术经济研究,2010(8): 102 – 115.

[7] 邓晶, 秦涛, 黄珊. 基于 Logistic 模型的我国上市公司信用风险预警研究[J]. 金融理论与实践,2013 (2): 22 – 26.

[8] 王小燕, 方匡南, 谢邦昌. 基于 adSGL-logit 的信用卡信用评分模型研究[J]. 统计研究, 2014(09): 107 – 112.

[9] Jianqing Fan, Runze Li, Variable selection via nonconcave penalized likelihood and its oracle properties [J]. Journal of the American Statistical Association, 2001, 96(456): 1348 – 1360.

[10] Cun-Hui Zhang. Penalized linear unbiased selection [J]. Department of Statistics, 2007, Technical Report(2007 – 003): 1 – 22.

[11] Jian Huang, Shuangge Ma, Hongzhe Li, et al. The sparse Laplacian shrinkage estimator for high-dimensional regression [J]. The Annals of Statistics, 2011, 39(4): 2021 – 2046.

[12] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty [J] The Annals of Statistics, 2010, 38(2): 894 – 942.

[13] Patrick Breheny, Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection [J] Ann Appl Stat, 2011, 5(1): 232 – 253.

[14] 石晓军,肖文远,任若恩. Logistic 违约率模型的最优样本配比与分界点研究[J]. 财政研究,2005 (9): 38 – 48.

[15] 陈静. 上市公司财务恶化预测的实证分析[J]. 会计研究,1999 (4): 31 – 38.

[16] 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(3): 102 – 110.

作者简介

方匡南,男,1983 年生,浙江台州人,2010 年毕业于厦门大学计划统计系获经济学博士学位,现为厦门大学经济学院教授、博士生导师,厦门大学数据挖掘研究中心副主任。研究方向为数据挖掘、计量经济学。

范新妍,女,1990 年生,河北保定人,2013 年毕业于武汉理工大学理学院获理学学士学位,现为厦门大学统计学在读博士研究生。研究方向为数理统计。

马双鸽,男,1978 年生,内蒙古呼伦贝尔人,2004 年 7 月获得美国威斯康辛大学统计学博士学位,现为美国耶鲁大学生物统计系副教授、厦门大学经济学院讲座教授、厦门大学数据挖掘研究中心副主任。研究方向为数理统计、数据挖掘、生物统计。

(责任编辑: 曹　麦)