

Research on Financial Early Warning of Big Data Enterprises Based on Logistic Regression and BP Neural Network

HongMei Zhang^{1,a}, Jian He^{2,b}

^{1,2}School of Big Data Application and Economics, Guizhou University of Finance and Economics, Guiyang Guizhou 550025, China

^azhm1035@qq.com, ^bdhwgd264nc@qq.com

基于 logistic 回归和 BP 神经网络的大数据企业财务预警研究

张红梅^{1,a} 贺健^{2,b}

^{1,2} 贵州财经大学大数据应用与经济学院, 花溪, 贵阳, 贵州, 550025, 中国

^azhm1035@qq.com, ^bdhwgd264nc@qq.com,

ABSTRACT:—Based on the difference between big data enterprises and traditional enterprises, this paper first constructs two single financial risk early warning models: logistic regression model and BP neural network model, then introduces default probability of logistic regression model output into BP neural network model, and establishes a non-linear combination based on BP neural network model. A new forecasting method is proposed, and an early warning model of financial crisis for large data enterprises is constructed and an empirical study is carried out. The results show that, compared with single model, the combined forecasting model has no significant improvement in the forecasting accuracy of financial early warning for large data enterprises, but the combined forecasting model is more stable. This provides a new idea for the financial risk early warning research of large data enterprises in China.

Key words — combination forecasting model, large data enterprises, financial early warning

摘要—基于大数据企业与传统企业的差异, 本文首先构建了 logistic 回归模型和 BP 神经网络模型两个单一的财务风险预警模型, 然后将 logistic 回归模型输出的违约概率引入到 BP 神经网络模型中, 建立了基于 BP 神经网络模型非线性组合的预测新方法, 构建了大数据企业财务危机预警模型并进行了

实证研究。结果表明, 复合预测模型在对大数据企业财务预警的预测精度上, 与单一模型相比, 预测精度没有显著提高, 但复合预测模型更具有稳定性。这为我国大数据企业财务风险预警研究提供了新思路。

关键词—复合预测模型; 大数据企业; 财务预警

I. 引言

近些年来, 随着整个经济环境的日益复杂, 越来越多的企业正在陷入或已经陷入财务危机, 企业抗风险能力普遍很弱。财务风险是企业面临的最主要的潜在危险之一, 由于财务困境是一个动态的、持续的过程, 所以财务危机的发生是可以提前进行预测的。对于企业财务风险预警这一方向, 国内外学者都有所涉及, 目前企业财务风险预警已经成为了学者们深入探究的一个热门课题。与此同时, 随着人类产生的数据量成几何式地增加, 大数据一词正在被越来越多地被提及, 有关于大数据的构思和概念在不断地提出、完善, 并在实际中被普遍运用, 由此催生出了一大批与大数据直接相关或紧密相关的企业, 我们称其为大数据企业。如华为、小米、腾讯等企业都准备建立基于“大数据”的企业生态系统。和传统企业不同的是, 上市大数据企业有着筹资能力强、规模大等特点, 是我国现代市场经济中最具有发展空间的企业之一。

参考以往文献发现,在研究对象方面,国内在大数据企业的财务预警研究方面较少,已有的文献多集中于大数据企业的运营绩效研究等方面,如孟祎(2018)^[1]采用 DEA 和 Tobit 模型来研究大数据企业运营绩效及其影响因素,张铁山(2016)^[2]则采用因子分析和相关性分析等方法来研究大数据企业的经营绩效;赵传仁(2017)^[3]则利用 Malmquist 生产率指数方法来研究我国大数据企业全要素生产率,并对其时序特征及增长差异进行了分析。在研究方法方面,尚未有学者用组合预测的方法来研究大数据企业的财务风险预警,因此为了有效预测和控制大数据企业的财务风险状况,从而进一步促进大数据企业为我国经济增长添加新动力,实现经济高质量发展,本文分别采用单一模型和组合模型来研究我国上市大数据企业的财务风险状况,为大数据企业的稳定发展提出中肯建议。

II. 文献回顾

目前总体来说对于大数据企业的界定还很模糊,只有部分券商和少量学者对这一概念做出过粗略的划分。如安信证券指出,遵循自身拥有大数据和为客户提供大数据分析运营两条主线选出的企业即为大数据企业。周永祥(2013)^[4]将大数据企业,划分为数据存储和处理、数据运营维护两个类别。然而基于这些学者的研究,只有很少的企业符合大数据企业的标准,本文认为大数据企业的范围应该进一步扩大,并将主营业务、核心能力和未来发展战略与大数据紧密相关的企业定义为大数据企业。

在财务预警研究方面,早在上个世纪,国外学者就进行了相关研究,最初是 Fitzpatrick(1932)^[5]与 Beaver(1966)^[6]共同提出的单变量判别分析模型,直到 1968 年 Altman^[7]把多元线性判别法应用于企业财务风险预警并进行实证分析。在这之后 Edmisterd(1973)^[8]提出了专门针对小企业的财务预警模型,日本开发银行建立的多变量预测模型等等, Dusan Marcek(2013)^[9]应用了基于统计理论和神经方法的预测模型,提出了一种确定财务风险最小的预测水平的方法。

组合预测模型最早是 J.M.Bates 与 C.W.JGranger 于上个世纪 60 年代提出,这一方法近年来也引起我国学者的重视,如张红梅(2011)^[10]通过构建 logistic 模型和 BP 神经网络模型的组合预测新方法,并应用于我国上市企业,实证表明组合预测方法较单一方法而言预测精度较高,李忠武(2013)^[11]也通过多元判别分析、Logit 分析和支持向量机模型分别建立了三个单项财务风险预警模型,然后利用这三个模型构建了一个组合预测模型来分别研究上市公司

财务状况,实证表明,与单一模型相比,组合预测方法有着较高的稳定性,王秋玮(2017)^[12]也通过决策树和 Microsoft 时序分析法建立了新常态下的企业财务预警模型,结果表明模型预测精度高达 84.6%。

通过以上文献回顾可以看出,目前国内对大数据企业的财务风险预警研究尚属空白。此外关于财务预警研究,我国学者大多数都是使用多变量模型,其可分为两种,一种是如 logistic 模型和 Fisher 判别方法等传统统计模型,另一种是如神经网络模型等人工智能模型。总体来说这两类模型都存在一定缺陷,如 logistic 模型参数最大似然估计量不存在和神经网络的“黑箱性”问题等等。考虑到目前预警方法的局限性,本文在参考全春光^[13]等人的基础上,把两种方法结合起来,采用组合预测模型来构建上市大数据企业财务预警机制。

III. 大数据企业财务风险预警实证分析

3.1 样本选取

目前国际上关于财务困境的界定标准还有较大的分歧,国外学者主要把破产清算当作企业发生财务困境的标志,而大部分国内学者认为企业是否发生财务危机主要取决于上市企业是否被证监会处以 ST。考虑到我国国情,本文也以是否被 ST 当作企业发生财务危机的标准,即 ST 企业为发生财务危机的公司,非 ST 企业为财务状况健康公司。

以往大多数学者都是以某一区域或某一个行业的企业作为研究对象,鲜有学者研究以大数据企业为研究对象的财务风险预警。考虑到数据的可获得性,本文从沪深证券交易所的上市企业中,按照前文所述的大数据企业定义标准,在剔除了存在数据缺失的企业之后,最终选取了我国 2009-2018 年间 112 家符合要求的上市大数据企业,其中包括 22 家财务状况出现问题即 ST 企业和 90 家财务健康的非 ST 企业。将样本集分为训练样本和测试样本,采用分层抽样法从样本集中随机抽取 70%(共 77 个样本)作为训练样本,30%(35 个样本)作为检验样本。另外参考以往文献,发现大多学者以上市企业被 ST 前一年的数据为基础来进行试验。事实上,企业在第 t 年是否被特殊处理是由其 t-1 年的年报数据所决定,这两件事情几乎是同时发生,因此若采用 t-1 年的数据进行分析,即使得出的结论有着较高的预测精度,在实际预测中也没有多大的应用价值。本文通过参考张新红^[14]等人,最终以财务危机公司被宣布 ST 前 3 年所公布的年报数据为基础,以期提高其实际应用价值。

3.2 预警指标选取

大数据企业作为技术密集型企业的典型代表企业，具有研发投入高，成长速度快和盈利能力强等特点，因此在评价指标的选取上，既要考虑能够综合反应企业整体情况的必要指标，又要考虑与大数据企业成长能力和盈利能力相关的指标。本文在考虑大数据企业发展特点的同时，参考邓新策^[5]等人构建的评价指标体系，从企业盈利能力、经营能力、偿债能力、成长能力和现金能力等方面，初步选取了每股收益、净资产收益率、资产报酬率、流动比率、速动比率、营业收入增长率、营业利润增长率、净利润增长率、存货周转率、应收账款周转率、总资产周转率、资产负债率、销售商品劳务收入现金/营业收入、销售现金比率等 14 个指标。本文的样本数据主要来源于 RESSET 金融研究数据库。运用 spss23 软件对 112 家上市大数据企业的数据进行主成分分析，提取主成份因子。

(1)KMO 检验和 Bartlett 球度检验

在用主成分分析法之前，首先需要进行 KMO 检验和 Bartlett 球形度检验，以此来分析样本数据是否适合采用主成分分析法。结果如表 1 所示。

表1 KMO和Bartlett检验

KMO取样适切性量数		0.565
巴特利特球形度检验	近似卡方	915.592
	自由度	91
	显著性	0.000

根据 Kaiser 研究表明，KMO 值大于 0.5 时，数据是比较适合做主成分分析的，当 KMO 值小于 0.5，数据就不适合做主成分分析。从表 1 可得，KMO 值为 0.565，适合做主成分分析。

(2)主成分因子提取

表2 总方差解释

成分	初始特征值		旋转载荷平方和	
	方差百分比	累积 %	方差百分比	累积 %
1	24.285	24.285	22.889	22.889
2	18.524	42.809	17.808	40.697
3	9.483	52.292	10.767	51.465
4	8.837	61.129	9.442	60.907
5	8.276	69.405	8.498	69.405

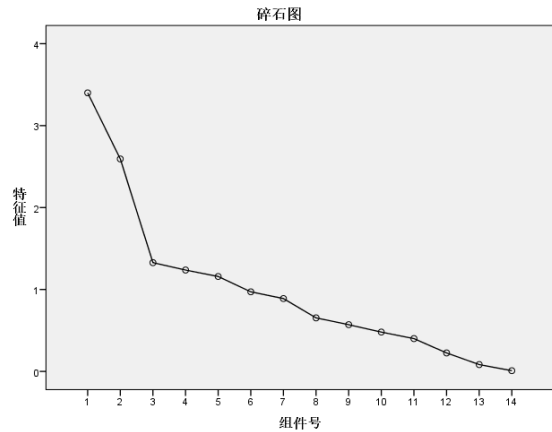


图1 碎石图

根据图 1 和表 2 可知，这 5 个主成分因子解释了总方差的 69.405%，解释程度较高，故本文最终选取前 5 个主成分作为变量来进行建模，每一个主成分在原始变量中所占权重如表 3 所示。

表3 成分得分系数矩阵

	成分				
	1	2	3	4	5
销售商品劳务收入现金/营业收入 X1	0.023	0.062	-0.183	0.613	0.003
销售现金比率 X2	0.028	0.067	0.195	0.455	-0.197
每股收益 X3	0.275	0.014	-0.113	0.019	0.036
净资产收益率 X4	0.186	0.031	-0.091	-0.313	-0.158
资产报酬率 X5	0.227	-0.033	0.114	-0.038	-0.017
流动比率 X6	-0.033	0.411	0.041	0.135	0.068
速动比率 X7	-0.027	0.409	0.038	0.145	0.087
营业收入增长率 X8	0.029	-0.037	0.457	-0.067	-0.079
营业利润增长率 X9	0.274	0.001	-0.111	0.041	0.063
净利润增长率 X10	0.280	-0.033	-0.020	0.066	0.015
存货周转率 X11	-0.004	0.034	-0.035	-0.071	0.541
应收账款周转率 X12	-0.129	0.048	0.614	-0.014	0.083
总资产周转率 X13	0.003	-0.022	0.039	0.047	0.657
资产负债率 X14	-0.034	-0.258	0.053	0.223	0.162

按照表 3 成分得分系数矩阵, 通过公式 $FAC_i = \sum C_{ij}$ 得出各因子表达式为:

$$F1 = 0.023X_1 + 0.028X_2 + 0.275X_3 + 0.186X_4 + 0.227X_5 - 0.033X_6 - 0.027X_7 + 0.029X_8 + 0.274X_9 + 0.280X_{10} - 0.004X_{11} - 0.129X_{12} + 0.003X_{13} - 0.034X_{14}$$

其他主成分因子表达式同理可得。

3.3 单模型预测结果分析

本文首先用 logistic 模型和 BP 神经网络两种单一模型来研究上市大数据企业财务风险状况, 此为建立组合预测模型的基础, 之后将 logistic 回归模型输出的违约概率引入到 BP 神经网络模型中, 从而构建出非线性组合预测模型并进行比较分析。

3.3.1 logistic 回归模型的应用

选取前 5 个主成分变量作为协变量来建立 logistic 模型, 所得模型结果如下表所示。

表4 霍斯默-莱梅肖检验

步骤	卡方	自由度	显著性
1	4.218	8	0.837

表5 方程中的变量

	B	标准误差	瓦尔德	显著性	Exp(B)	
步骤 1	FAC1	0.296	0.210	1.980	0.159	1.344
	FAC2	0.832	0.397	4.385	0.036	2.299
	FAC3	-0.342	0.202	2.877	0.090	0.710
	FAC4	0.050	0.251	0.039	0.844	1.051
	FAC5	0.070	0.264	0.071	0.790	1.073
	常量	1.614	0.283	32.455	0.000	5.024

根据运行结果, 可得 logistic 回归模型如下:

$$Y = \frac{1}{1 + e^{-(1.614 + 0.296F1 + 0.832F2 - 0.342F3 + 0.05F4 + 0.07F5)}} \quad (1)$$

参考胡宇^[6]等人的文献, 以 0.5 作为 ST 企业和非 ST 企业的分界点, 若模型的预测概率大于 0.5, 则认定该企业为 ST 企业, 反之则认定为财务状况健康企业, logistic

回归模型建模结果如表 6 所示(本文把财务健康企业定义为 1, ST 企业定义为 0)。

表6 logistic回归预测结果

实测		预测			
		财务状况		正确百分比	
		0	1		
步骤 1	财务状况	0	3	19	13.6%
		1	1	89	98.9%
	总体百分比		82.1%		

由表 6 可知, logistic 回归模型的总体预测准确率达到 82.1%, 其中 ST 企业被正确预测的概率为 13.6%, 第一类错误率为 86.4%, 非 ST 企业被正确预测的概率为 98.9%, 第二类错误率为 1.1%。该模型对 ST 企业预测的准确率偏低, 第一类错误率明显高于第二类错误率, 这可能跟 ST 大数据企业的样本量较少有关。

3.3.2 BP 神经网络模型的应用

同样的选取前 5 个主成分变量作为协变量来建立 BP 神经网络模型, 所得模型结果如下表所示。

表7 BP神经网络预测结果

样本	实测	预测		
		0	1	正确百分比
训练	0	6	12	33.3%
	1	6	53	89.8%
	总体百分比	15.6%	84.4%	76.6%
检验	0	3	1	75.0%
	1	1	30	96.8%
	总体百分比	11.4%	88.6%	94.3%

由表 7 可得, 该模型将训练样本中的 18 家 ST 企业中的 12 家判定为财务状况健康企业, 6 家判定为 ST 企业, 正确率为 33.3%; 59 家财务状况健康企业中的 6 家被判定为 ST 企业, 53 家判定为财务正常企业, 正确率为 89.8%。将检验样本中的 4 家 ST 企业中的 3 家判定为 ST 企业, 1 家被判定为财务状况健康企业, 正确率为 75%; 31 家财务状况健康企业中的 1 家被判定为 ST 企业, 30 家被判定为财务正常企业, 正确率为 96.8%。从该结果可得出, 训

训练样本的平均准确率为 61.55%，总体正确率为 76.6%，检验样本的平均准确率达到 85.9%，总体正确率为 94.3%。

3.4 复合预测分析

复合预测模型是将 logistic 模型输出的违约概率引入到 BP 神经网络中，这样网络的输入就包括前 5 个主成分变量和 1 个表示上市大数据企业违约概率的变量，共 6 个输入。组合预测模型结果见表 8。

表8 组合模型预测结果

样本	实测	预测		
		0	1	正确百分比
训练	0	6	12	33.3%
	1	4	55	93.2%
	总体百分比	13.0%	87.0%	79.2%
检验	0	2	2	50.0%
	1	1	30	96.8%
	总体百分比	8.6%	91.4%	91.4%

由上表显示结果可知，该模型把训练样本中的 18 家 ST 企业中的 12 家判定为财务状况健康企业，另外 6 家判定为财务危机企业，预测准确度为 33.3%；59 家非 ST 企业的 4 家判定为 ST 企业，其余 55 家判定为财务状况健康企业，预测正确率为 93.2%。把检验样本中的 4 家 ST 企业中的 2 家判定为财务危机企业，另外 2 家判定为财务正常企业，预测正确度为 50%；31 家非 ST 企业中的 1 家判定为财务危机企业，30 家判定为财务状况健康企业，预测正确度为 96.8%。可以看出，总体来说采用复合预测模型对企业财务风险预警的预测精度较高，训练样本的平均预测正确度达到 63.25%，总体正确率为 79.2%，检验样本的平均预测正确度达到 73.4%，总体正确率为 91.4%。

3.5 模型预测效果比较分析

通过这三个模型对我国 112 家上市大数据企业进行分析，得出不同的预测结果，根据表 6、表 7、表 8 可知，logistic 回归模型平均预测精度为 56.25%，BP 神经网络模型平均预测精度为 73.72%，复合模型平均预测精度为 68.33%，由此可见，复合预测模型和 BP 神经网络模型的预测精度都在 70%左右，对上市大数据企业的财务预警精度较高，logistic 模型的财务风险预警效果次之，但也在 60 左右%，这说明我国上市大数据企业财务数据是真实有效的且具有较强的预测性。

IV. 结论

本文选取我国 112 家上市大数据企业为样本进行实证分析，通过企业的盈利能力和成长能力等方面选取了 14 个评价指标，通过主成分分析得到了五个主成分因子，先通过 logistic 回归模型和 BP 神经网络模型对样本进行单模型预测，然后通过将 Logistic 模型输出的违约概率引入到 BP 神经网络中，构建非线性组合预测模型来研究大数据企业财务预警，得出如下结论：

(1)为了降低指标之间的多重相关性，本文通过主成分分析法将 14 个指标分为 5 个主因子，并把 5 个主成分变量作为协变量，而不是传统地将某些单个指标作为协变量，大数据企业财务状况作为因变量，结果表明模型在研究大数据企业财务预警分析上具有可行性。

(2)通过上述实证可以看出，BP 神经网络模型预测精度较高，平均预测精度达到了 73.72%，随后是复合预测模型，平均预测精度为 68.33%，之后再是 logistic 回归模型，平均准确度为 56.25%。另外从数据的对比中可以看出，复合预测模型在判断财务正常企业时的预测精度是最高的。

(3)从表中可以看出，预测精度变化最小的是复合预测模型，为 12.2%，随后是 BP 神经网络模型，为 17.7%。因此，从模型稳健性的角度来说，复合预测模型的稳健性较好。

致谢

本研究得到国家自然科学基金地区项目《贷款风险补偿资金对科技型中小企业信贷配给的影响机理研究》(71263011)、《基于文本信息的科技型中小企业信用风险识别机理研究》(71861003)的大力资助。

参考文献

[1]孟沛.基于 DEA-Tobit 模型的大数据类上市公司运营绩效及影响因素研究.当代会计, vol.5,pp.3-4,2018.
 [2]张铁山,邓新策.基于上市大数据企业的经营绩效与研发投入关系研究.工业技术经济, vol.35,issue 9,pp.77-84,2016.
 [3]赵传仁,戴俊丽.中国大数据企业全要素生产率增长及收敛研究.西北大学学报(哲学社会科学版), vol.47,issue 2,pp.60-66,2017.
 [4]周永祥.关注大数据概念股.股市动态分析, pp.21-28,2013.

- [5]Fitzpatrick P J,A comparison of ratios of successful industrial enterprises with those offailed firms,Certified Public Account,1932.
- [6]Beaver WH, Financial ratios as predictors of failure,Journal of Accounting research(Supplement),1966.
- [7]Edward I Altman,Financial ratios, discriminant analysis and the prediction of bankruptcy, Journal of Finance,1968.
- [8]Edmister Robert O,Financial ratios as discriminant predictors of small business failure,Journal of Finance, 1972.
- [9] Dusan Marcek, "Risk Scenes of Managerial Decision-Making with Incomplete Information: An Assessment in Forecasting Models Based on Statistical and Neural Networks Approach," Journal of Risk Analysis and Crisis Response,vol.3,issue 1,pp.13-21,2013.
- [10]张红梅,刘文蕊.组合预测在上市公司财务预警中的应用.会计之友, vol.6,pp.101-102,2011.
- [11]李忠武.基于组合预测技术的上市公司财务困境预测问题研究.山东财经大学,2013.
- [12]王秋玮.新常态下 ST 公司财务困境预警机制研究.浙江工业大学,2017.
- [13]全春光,程晓娟.基于组合预测理论的上市公司财务危机预警研究.财会通讯, vol.15,pp.32-33,2009.
- [14]张新红,王瑞晓.我国上市公司信用风险预警研究.宏观经济研究, vol.1,pp.50-54,2011.
- [15]邓新策.基于上市大数据企业的经营绩效与研发投入关系研究.北方工业大学,2017.
- [16]胡宇.组合预测模型在财务预警中的应用.长沙理工大学,2013.