基于大数据的企业财务预警研究

The Research of Enterprise Financial Early Warning Based on Big Data

宋 彪 朱建明 李 煦

SONG Biao ZHU Jian-ming LI Xu

[摘 要] 当前的财务危机预警研究主要着眼点是基于财务指标建立模型，然而会计舞弊或者会计不作为现象的存在，导致各个财务预警模型在以往的经济危机中纷纷失效。引入了非财务指标的一些模型中，选取的指标相对片面，难以适应对各种非财务指标具有不同敏感度的公司样本。大数据技术的出现，使得获得全面而客观信息成为可能，笔者提出了以网民为企业“传感器”的思想，即基于互联网上的相关在线信息，通过情感分析处理，以及统计网民信息发布频次，融合后形成传感器信号，在此基础上结合财务指标，尝试建立引入大数据指标的财务风险预警模型，并对模型的预测效果进行比较分析。检验显示基于大数据的财务风险预警模型具有更好的有效性。实验结果为相关方面预测我国上市公司财务危机提供了大数据方面的理论依据。

[关键词] 大数据 财务预警 上市公司 支持向量机

[中图分类号] F275 [文献标识码] A [文章编号] 1000 - 1549 (2015) 06 - 0055 - 10

Abstract: The current financial crisis warning research focuses on model which is based on financial indicators. However, in reality accounting fraud or nonfeasance causes the failure of financial early warning model in the past economic crisis. the models selecting one-sided indicators are not suitable for all kinds of sample companies. The development of big data technology makes it possible to obtain comprehensive and objective information. This paper proposes that internet users are regarded as a sensor to enterprise. Based on the relevant online information, sensor signals are formed by mixing the analysis of emotion processing and internet information release frequency statistics together, try to bring the index of big data into financial risk early warning model, and compares the prediction results of the model analysis. We found the financial risk early warning model based on big data has better effectiveness. The results provides a theoretical basis for forecasting listed company financial crisis by big data.

Key words: Big data Financial crisis warning Listed company SVM

[收稿日期] 2015 - 2 - 24

[作者简介] 宋彪，男，1983 年 2 月生，中央财经大学信息学院博士生，内蒙财经大学讲师，研究方向为经济信息分析；朱建明，男，1965 年 8 月生，中央财经大学信息学院院长，教授，博士生导师，研究方向为信息安全；李煦，男，1974 年 5 月生，香港大学经济及工商管理学院会计系副教授，研究方向为会计、行为金融学。

[基金项目] 国家自然科学基金项目“基于博弈论的信息安全理论与方法研究”(61272398)；国家社会科学基金重点项目“大数据时代网络媒介生态环境下个人信息保护体系的构建研究”(13AXW010)；北京市哲学社会科学重点项目“基于大数据的财务预警理论与方法研究”(14JGA001)

感谢匿名评审人提出的修改建议，笔者已做了相应修改，本文文责自负。

一、引言

财务危机风险预警是一个世界性的问题和难题[1]。从 20 世纪 30 年代开始，比较有影响的财务预警方法已经有十几种，但这些方法在经济危机中能够真正预测企业财务风险的很少，还没有查到这方面的有力证据。究其原因，大多数模型中，财务指标是主要的预测依据。但财务指标往往只是财务发生危机的一种表现形式，甚至还有滞后反应性、不完全性和主观性[2]。更为严重的是在基于财务指标预警模型建立过程中，学者们往往都假设财务数据是真实可靠的，但这种假设忽略了财务预警活动的社会学规律，为财务预警模型与现实应用的脱节埋下了伏笔。许多学者建立了结合非财务指标的模型，但所加入的能够起到作用的非财务指标都是依靠试错方法引入的，即都是在危机发生之后，才能够使指标得以确认以及引入模型，下一次经济危机的类型不同，之前建立的财务预警模型便会无法预测甚至可能发生误导。因此靠试错引入的非财务指标具有一定的片面性，忽视了这些指标间的相互作用和相互关系，无法顾及这些指标是否对所有企业具有普遍适用性。

个人网上搜索、发布或者关注行为，以及各种论坛、博客、媒体发布的企业相关信息往往在不经意间透露了一些企业真实的管理状态和走势信息，而且涵盖的范围非常广泛。根据大数据的思维范式[3]，可否把这些信息，通过大数据的处理技术量化后引入到财务风险预警模型当中？基于大数据的财务风险模型是否能够在稳定性、有效性等方面具有一定的改善？这些问题的本质将是当前大数据技术破解财务风险预警难题的一个重要契机。

二、理论基础、文献回顾

风险理论认为，经济活动之所以存在风险，原因在于经济环境存在不确定性。在财务危机预警过程中，为了更准确地预测企业的财务风险，有效的途径就是尽可能寻找全面的可以反映企业自身状况和所处经济环境的指标。

目前主流的财务预警研究方法依然是用财务指标构建模型。“几乎所有的研究都集中于寻找最佳的公开财务指标来预测财务危机”[4]。基于财务指标建模的支持者往往利用实证研究能够得出较高的判别率。如 Altman 在样本公司破产前一年预测准确率是 95%[5]，Deakin 的准确率为 97% 等等。由于财务信息存在滞后性、灰色性和短期性等明显缺点，许多学者已经认识到根据财务指标预测财务风险的局限。Johnson 针对 Altman 的论文使用的预测方法指出：即使比率确实能够提供目前企业状态的信息，它们也并不包含企业的替代战略以及管理层和投资者面临的经济状况，比如企业合并和推迟付款[6]。Ohlson 将国民生产总值、价格指数引入了财务风险预测模型中[7]。EllMoumi 等研究发现，企业董事会的结构和构成可以用于解释财务危机的发生[8]。Campbell 等发现低股票收益率和高股票波动率将增大企业的财务风险[9]。这些研究都各自证明了所提出非财务变量的预警有效性，但在具体操作中，数据的获取难度限制了非财务指标在财务预警方面的系统研究，提出的变量没有统一的标准，也没有考虑变量之外的影响因素，反映不出企业对某一非财务指标敏感程度。企业危机发生后，发现最终导致预警失败的，往往是研究者们未曾考虑过的因素。同时，多数研究者更关注危机发生的根源，以求毕其功于一役，却忽略了财务预警这一行为本身的社会性。简单认为明确了危机发生的原因，就可以进行准确的预测，事实上财务危机每次发生的原因都在变化。奥斯特罗姆的研究指出，信任、声誉与互惠机制来自于人际网络[10]。企业危机的发生，起源于企业所有的相关人在社会网络中的相互作用。大数据技术的出现，使从社会网络角度获得关注会计主体更多一些的细节信息成为可能。所谓大数据：一般认为具有 4V 特征的数据可以称之为大数据，如果从广义的角度拓展，其实大数据是一种思维范式。

维克托·迈尔-舍恩伯格认为，大数据在分析过程中需要全部数据样本而不是抽样[11]，目标上关注相关性而不是因果关系。人机互动产生的数据是大数据最主要的来源。这些人们在互联网世界里留下的各种“数据足迹”，并不是有意识地留下的数据，而是机器之间相互处理交互时沉淀下来的数据，从社会学角度来看，获得的数据无意识性越强，或者说人为主观参与度越小，这种数据就越客观，也越接近事物的本质，危机的爆发总是具有出乎意料的特征，具备更强无意识度的信息对危机的预测具有更重要的意义。

目前利用从互联网获得的大数据进行研究，一般从分析群体情绪和网民行为规律来实现。对网络上公众的情绪进行分析是一项高难度工程，国外在相关方面取得了一系列进展，美国印第安纳大学的约翰·博伦（Johan Bollen）等人对 Twitter 进行研究，将实验中获得的“冷静”情绪指数后移 3 天，可以和道琼斯指数获得惊人的一致。同时该研究还测试了一个基

于自组织模糊神经网络的股市预测模型，当模型仅输入股市数据时，模型可以达到 73.3% 的准确率，而在加入“冷静”的情感信息后，准确率明显上升至 86.7%。麻省理工的张雪等人，根据情绪词将推文标定为正面或负面情绪[12]。结果发现，无论是如“希望”的正面情绪，或是“害怕”、“担心”的负面情绪，其占总推文数的比例，都预示着道琼斯指数、标准普尔 500 指数、纳斯达克指数的下跌。美国佩斯大学的亚瑟·奥康纳，发现 Facebook 上的粉丝数、Twitter 上的听众数和 Youtude 上的观看人数，都和股价密切相关。品牌的受欢迎程度，还能预测股价在 10 天、30 天之后的上涨情况①。Tobias Preis 等人研究发现，当特定的关键词在某一时间内被大量搜索时，在股票市场上可能有买入或者卖出的行为发生[13]。

三、基于大数据的企业危机预警模型

从在线信息获取的企业相关大数据，其内容可以包含导致企业财务危机方方面面的因素，甚至包含人们尚未认识到的危机根源。大数据体现了群体智慧的特征，有价值的信息密度非常低，这使一些人为的修改在群体行为的均衡下，信息的价值往往不受太大的影响，可以避免仅依靠信息提供者而受到蒙蔽的现象。大数据信息比以往通过公司公告、调查、谈话等方式获得的信息更为客观和全面，而且这些信息中可以囊括企业在社会网络中的嵌入性影响。

在社会环境中，企业存在的基础在于相关者的认可，这些相关者包括顾客、投资者、供应链伙伴、政府等等。考虑到企业的经营行为，或者企业关联方的动作都会使企业的相关者产生反应，进而影响到网络上的相关信息。因此本文把所有网民看作企业分布在网络上的“传感器”，这些“传感器”有的反应企业的内部运作状态，有的反应企业所处的整体市场环境，有的反应企业相关方的运行状态等等。由此构建基于大数据进行企业财务预警的模型:

大数据企业财务预警系统不排斥财务报告上的传统指标，相反，传统的财务指标应该属于大数据的一部分。互联网上网民对企业的相关行为，包含了线下的人们和企业的接触而产生对企业的反应，这些反应由于人们在社会网络中角色的不同，涵盖了诸如顾客对产品的满意度、投资方的态度、政策导向等各种可能的情况。所有这些信息通过线下向互联网映射，在互联网中通过交互作用，由网民的情绪形成了相关企业的企业网络舆情。

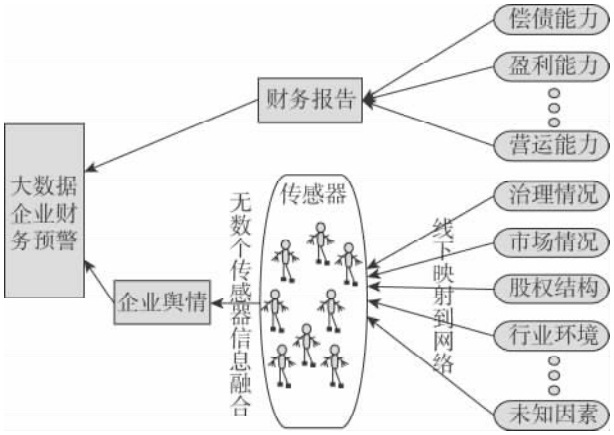

图 1 大数据企业财务预警原理

在具体处理过程中，可以对网络舆情信息进行语义分析，通过情绪指标对舆情进行量化，形成各种行为的一个融合后的综合性指标，具体数据处理过程如图 2 所示。

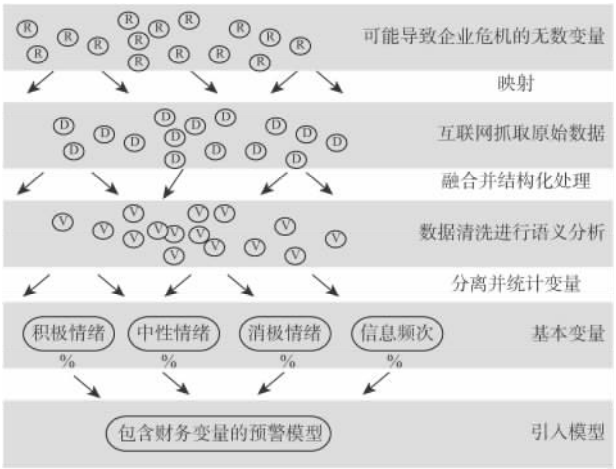

图 2 大数据企业财务预警数据处理机制

起到企业“传感器”作用的网民，由于在线下和企业有着各种各样的角色关系。这些角色和企业的相互作用会产生不同的反应，从而刺激这些角色对企业产生不同的情绪。群体的情绪通过映射到互联网，才使这些信息能够被保存下来并被我们获取，这些不同的情绪经过网络上交互过程中的聚集、排斥和融合作用，最后会产生集体智慧，这些群体智慧能反应企业的某种状态。

因此得出如下命题:

① New study finds link between social media popularity and stock prices http: //www. famecount. com/news/ new-study-finds-link-between-social-media-popularity-and-stock-prices-242652.

命题 1：引入企业相关网络大数据信息正面情绪指标的财务风险预警模型可以提高预警的有效性。

命题 2：引入企业相关网络大数据信息中性指标的财务风险预警模型可以提高预警效果。

命题 3：引入企业相关网络大数据信息负面情绪指标的财务风险模型可以提高预警效果。

命题 4：引入企业相关网络大数据信息正面和负面交互影响指标的财务风险模型可以提高预警效果。

命题 5：引入企业相关大数据产生频次指标的财务风险预警模型可以提高预警的有效性。

四、研究样本、研究变量和研究设计

（一）研究样本

利用聚焦网络爬虫，收集了从 2009 年 1 月 1 日到 2013 年 12 月 31 日的关于 60 家企业的所有相关全网网络数据，包括新闻、博客、论坛等信息，经过在线过滤删重，最终获得有效信息共 7 000 万余条。来自网络的上市公司相关大数据主要是非结构化的文本信息，而且包含大量重复信息。为了验证大数据反映的相关情绪能够有效提高财务风险预警模型的性能，首先要把这些信息进行数值化处理，过滤掉大量无效数据，并且进行基于财经领域词典的文本情绪倾向计算。同时对相关上市公司的有效信息进行频次统计，以便验证大数据有效信息频次对财务风险预警模型的影响。考虑到不同行业的财务特征并不相同，而制造业在上市企业中所占比例最大，正常企业的数量远大于具有危机风险的企业。研究中把制造业作为模型研究的样本企业，将会使模型在实际应用中具有更好的代表性。

在沪深两市的危机企业中，危机企业的数量远远小于正常企业。如果按照资产规模 1∶1 配对抽样，会破坏样本的随机性，导致模型效果虚高，夸大了模型的预测精度[14]，同时以资产规模为配对原则缺乏有力的理论依据，本文后面对资产规模进行检验也发现在对危机的判断中并不显著。因此本文将危机企业和正常企业按 1∶2 的方式进行随机抽样配比（注意不是配对）。共搜集 60 家企业，其中危机企业 20 家，正常企业 40 家。危机企业样本来源于 2012、2013 年被沪深两市特别处理的工业制造业企业，2012 年危机企业 11 家，正常企业 22 家，2013 年危机企业 9 家，正常企业 18 家。筛选采用的标准是：上市以来首次被处理，上市时间已经超过 5 年，因为连续两年亏损而被特别处理。正常企业的样本采用随机抽取，在沪深两市中上市超过 5 年，上市以来从未被特殊处理的工业制造业企业。样本企业的财务指标的采集和计算源自 RESSET 瑞斯金融研究数据库。

（二）研究变量

财务危机预警模型主要有两个核心的工作：一个是预警指标的确定；一个是预警模型算法的选择。前者是财务预警信息的深层次挖掘，后者是预警算法技术的应用，两者同时对企业财务危机预警的精度产生影响。也就是说，财务危机预警模型的效果不仅取决于模型的学习和泛化能力，也取决于选择的财务预警模型输入和输出变量。

财务风险预警理论的研究一直缺乏系统的经济理论支持，在实际预警工作中没有特别直接的理论基础。因此现阶段的预警工作多数从实际数据出发，采用经验研究的方法，用试错的方法，逐一考察变量的组合在实际样本数据中的表现，筛选出突出判别能力的变量组合来构建最终的预测模型。

财务指标方面的变量选取参考目前财务危机预警大量的研究成果，选出 32 个变量作为备选考察变量，如表 1 所示。

表 1　　财务指标

	符号	指标名称
企业偿债能力	X1	current ratio
	X2	quick ratio
	X3	working capital/assets
	X4	asset-liability ratio
	X5	cash ratio
企业盈利能力	X6	EBIT/ general assets
	X7	Sales net profit rate
	X8	profit rate to net worth
	X9	rate of earnings on shareholders equity
	X10	return on equity
	X11	Operating margins
	X12	profit margin of the primary business
企业现金流量	X13	Recovery rate of The total assets
	X14	Sales cash ratio
企业资本结构	X15	Ratio of fixed assets
	X16	Intangible assets ratio
企业成长能力	X17	The growth rate of net assets
	X18	Net profit growth rate
	X19	The main business revenue growth rate
	X20	Operating profit growth rate
	X21	Total profit growth rate
	X22	The growth rate of total assets
	X23	The main business profit growth rate
	X24	Operating cash flow per share growth rate
	X25	Growth rate of netassets per share
	X26	growth rate Eps
	X27	Capital increment rate
	X28	Retained earnings/assets

续前表

	符号	指标名称
企业营运能力	X29	The rate of inventory turnover
	X30	Accounts receivable turnover rate
	X31	turnover of current assets
	X32	turnover of total capital

其中 X1 ~ X5 是反映企业偿债能力的指标，X6 ~ X12 是反映企业盈利能力的指标，X13 ~ X14 是反映企业现金流量的指标，X15 ~ X16 为反映资本结构的指标，X17 到 X28 是反映企业成长能力的指标，X29 ~ X32 是反映企业营运能力的指标。

输入变量由假设形成的 5 个变量结合财务指标变量构成。关于输出变量，考虑企业发生财务危机风险的界定，学者的观点并不统一，多数学者都采用将 ST 作为企业陷入财务危机，具有财务风险的标志。这样的界定标准符合我国的现实情况，而且便于学者间的成果相互比较，因此把 ST 作为模型输出变量。

（三）研究方法

研究过程分为 3 个部分，第一个部分，确定有效的财务变量指标，建立基于财务指标的财务危机预警模型。第二部分，加入大数据变量，考察模型在加入大数据变量的基础上对模型效果的影响，如图 3 所示。

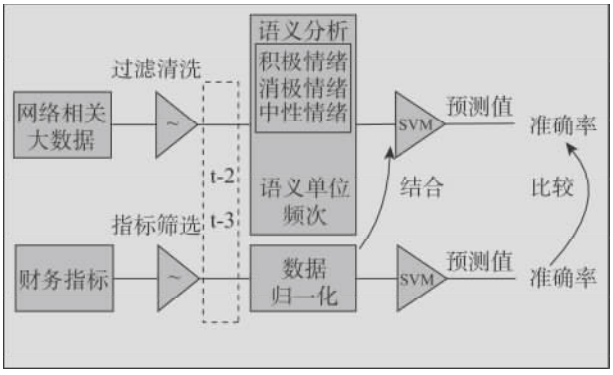

图 3 模型验证过程图

（四）模型的构建

构建模型前的数据处理：对引入模型的各项指标进行显著性检验和多重共线性分析。

1. 财务指标的正态分布检验

在进行显著检验之前，需要考察样本的分布情况，并以此来确定使用哪种方法来进行预警指标的差异显著性检验。运用 SAS 软件来进行财务指标分布情况的考察。采用 Shapiro-Wilk 正态分布检验。Shapiro-Wilk 检验适用于样本量小于 2 000 的正态分布检验，适用于探索连续型随机变量的分布。对选取的企业各个财务指标进行检验，结果如表 2 所示。

表 2 正态分布检验

X1	W =	0. 792 24	Pr < W	<0. 000 1
X2	W	0. 708 566	Pr < W	<0. 000 1
X3	W	0. 976 706	Pr < W	0. 683 2
X4	W	0. 933 197	Pr < W	0. 067 1
X5	W	0. 463 693	Pr < W	<0. 000 1
X6	W	0. 229 876	Pr < W	<0. 000 1
X7	W	0. 229 395	Pr < W	<0. 000 1
X8	W	0. 229 048	Pr < W	<0. 000 1
X9	W	0. 222 516	Pr < W	<0. 000 1
X10	W	0. 209 483	Pr < W	<0. 000 1
X11	W	0. 244 886	Pr < W	<0. 000 1
X12	W	0. 480 229	Pr < W	<0. 000 1
X13	W	0. 913 662	Pr < W	0. 012 3
X14	W	0. 854 439	Pr < W	0. 000 4
X15	W	0. 956 127	Pr < W	0. 200 4
X16	W	0. 923 127	Pr < W	0. 150 0
X17	W	0. 350 579	Pr < W	<0. 000 1
X18	W	0. 959 847	Pr < W	0. 255 9
X19	W	0. 287 268	Pr < W	<0. 000 1
X20	W	0. 740 375	Pr < W	<0. 000 1
X21	W	0. 861 226	Pr < W	0. 000 6
X22	W	0. 275 577	Pr < W	<0. 000 1
X23	W	0. 814 331	Pr < W	< 0. 000 1
X24	W	0. 532 709	Pr < W	<0. 000 1
X25	W	0. 306 387	Pr < W	<0. 000 1
X26	W	0. 905 474	Pr < W	0. 007 4
X27	W	0. 350 579	Pr < W	<0. 000 1
X28	W	0. 520 976	Pr < W	<0. 000 1
X29	W	0. 616 608	Pr < W	<0. 000 1
X30	W	0. 219 643	Pr < W	<0. 000 1
X31	W	0. 838 908	Pr < W	0. 000 2
X32	W	0. 886 62	Pr < W	0. 002 4

根据 Shapiro-Wilk 判定原理，多数财务指标不服从正态分布，只有 X3、X4、X15、X16、X18 服从正态分布。这个结果和国外对财务指标的实证研究分析结果大致相同。对不服从正态分布的财务指标的差异显著性检验要采用非参数的检验方法，而对服从正态分布的财务指标使用参数 T 检验方法。

2. 财务指标的差异显著性检验

采用 Wilcoxon rank-sum 检验法，对 2 个独立样本进行非参数检验。找出对分辨 ST 公司和非 ST 公司没有贡献的财务指标。检验结果见表 3。

表 3　　差异显著性检验

X1	Chi-Square	4. 411 8	Pr > Chi-Square	0. 035 7
X2	Chi-Square	6. 743 8	Pr > Chi-Square	0. 009 4
X5	Chi-Square	9. 568 8	Pr > Chi-Square	0. 002 0
X6	Chi-Square	2. 218 3	Pr > Chi-Square	0. 136 4
X7	Chi-Square	3. 793 4	Pr > Chi-Square	0. 051 5
X8	Chi-Square	5. 250 4	Pr > Chi-Square	0. 021 9
X9	Chi-Square	1. 890 1	Pr > Chi-Square	0. 169 2
X10	Chi-Square	3. 360 2	Pr > Chi-Square	0. 066 8
X11	Chi-Square	4. 252 8	Pr > Chi-Square	0. 039 2
X12	Chi-Square	0. 001 8	Pr > Chi-Square	0. 714 5
X13	Chi-Square	7. 560 5	Pr > Chi-Square	0. 006 0
X14	Chi-Square	5. 250 4	Pr > Chi-Square	0. 021 9
X17	Chi-Square	2. 218 3	Pr > Chi-Square	0. 136 4
X19	Chi-Square	0. 001 5	Pr > Chi-Square	0. 969 5
X20	Chi-Square	1. 493 4	Pr > Chi-Square	0. 221 7
X21	Chi-Square	0. 373 4	Pr > Chi-Square	0. 541 2
X22	Chi-Square	3. 793 4	Pr > Chi-Square	0. 051 5
X23	Chi-Square	2. 218 3	Pr > Chi-Square	0. 345 1
X24	Chi-Square	0. 472 5	Pr > Chi-Square	0. 491 8
X25	Chi-Square	0. 373 4	Pr > Chi-Square	0. 541 2
X26	Chi-Square	0. 472 5	Pr > Chi-Square	0. 491 8
X27	Chi-Square	2. 218 3	Pr > Chi-Square	0. 136 4
X28	Chi-Square	1. 890 1	Pr > Chi-Square	0. 169 2
X29	Chi-Square	8. 423 9	Pr > Chi-Square	0. 003 7
X30	Chi-Square	0. 005 8	Pr > Chi-Square	0. 939 1
X31	Chi-Square	0. 036 5	Pr > Chi-Square	0. 848 6
X32	Chi-Square	4. 906 2	Pr > Chi-Square	0. 026 8

检验结果发现财务指标 X1、X2、X5、X8、X11、X13、X14、X29、X32 的 Pr > Chi-Square 都小于 0. 05，通过显著性检验，而其余指标因为没有通过显著性检验被剔出。

采用 T 检验方法，对样本进行参数检验，检验结果如表 4 所示。

表 4　　T 检验

Variable	DF	t Value	Pr > | t |
X3	32	1. 11	0. 274 1
X4	32	19. 64	< 0. 000 1
X15	32	9. 27	< 0. 000 1
X16	31	8. 02	< 0. 000 1
X18	32	1. 95	0. 060 5

发现 X4、X15、X16 的 Pr > | t | 都小于 0. 05，通过显著性检验，而 X3、X18 被剔出。

其中 X1、X2、X4、X5 是反映企业偿债能力的指标；X8、X11 是反映企业盈利能力的指标；X13、X14 是反映企业现金流的指标；X15、X16 是反映企业资本结构指标；X29、X32 是反映企业营运能力的指标。

3. 财务指标的多重共线性检验

从数据中看出有一定的多重共线性，如表 5 所示。

表 5　　多重共线性检验

Label	VarianceInflation	ConditionIndex
X1	36. 718 58	1. 582 19
X2	47. 121 27	2. 259 39
X4	2. 343 52	2. 628 34
X5	19. 552 20	3. 160 79
X8	3. 117 71	4. 239 16
X11	2. 660 26	5. 236 78
X13	5. 270 03	7. 079 14
X14	6. 924 02	8. 031 54
X15	1. 837 56	9. 477 31
X16	1. 218 37	13. 398 07
X29	2. 802 42	22. 418 26
X32	1. 338 56	29. 855 44

存在多重共线性的模型用于预测时，往往不影响预测结果。

对大数据指标的检验，命名积极情绪指数 P1，中性情绪指数 P2，消极情绪指数 P3，频次指数 C1，交互情绪指数 P1_ P3。根据分布状态分别进行显著性检验，见表 6。

表 6　　显著性检验

P1	Chi-Square	6. 095 4	Pr > Chi-Square	0. 013 6
C1	Chi-Square	8. 874 0	Pr > Chi-Square	0. 002 9

可以发现 P1 和 C1 显著。

从表 7 可以发现，样本数据中，p2，p3，p1_ p3 不显著。

表 7　　T 检验

Variable	DF	t Value	Pr > | t |
p2	32. 1	0. 78	0. 439 8
P3	34. 3	1. 03	0. 312 3
P1_ p3	37. 8	-1. 72	0. 092 9

4. 选用的模型

支持向量机（Support Vector Machine，SVM）是 Vapnik 等人在多年研究统计学习理论基础上提出的，是一种新型机器学习方法。支持向量机凭借统计学习理论，奠定了坚实的基础，其设计源于结构风险最小原则和有限样本假设，在过学习、局部收敛、高维灾难等问题方面克服了传统机器学习（如神经网络）的缺陷，支持向量机在学习能力和泛化性能方面具有

不可比拟的优势。SVM 分类方法基于结构风险最小化理论，在特征空间中建构最优分割超平面，使得学习器得到全局最优化，基本原理如图 4。

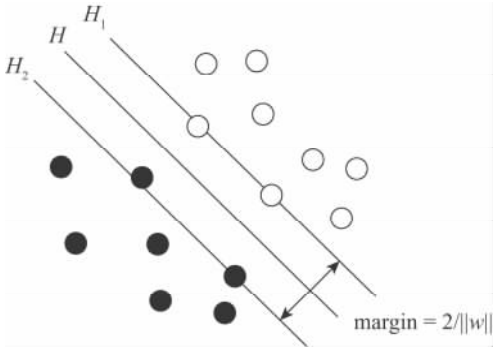

图 4 SVM 超平面图

图 4 中黑点和白点代表两个类别样本，H 表示超平面，H1，H2 是平行于超平面的直线，分别表示两个类别中离超平面最近的样本。H1 与 H2 的距离称为分类间隔（margin）。能将两个类别正确分开，并使分类间隔最大的平面即最优分类平面。SVM 超平面可用方程表示为：$H: \omega \cdot x + b = 0$，对于样本 $(x_i, y_i) i = 1,2,3\cdots,n$ 进行最优分类超平面构造，用如下规划问题表示：

$$\min_{\omega,b} \frac{1}{2}||W|| = \min_{\omega,b} \frac{1}{2} W^T W$$

$$\text{s. t.} \quad \gamma_i[(\omega \cdot x_i) + b] - 1 \geqslant 1 \quad i = 1,2,3\cdots,n$$

引入拉格朗日系数 α_i，得到对偶形式：

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j \gamma_i \gamma_j (x_i x_j)$$

$$\text{s. t.} \quad \sum_{i=1}^{n} \gamma_i \alpha_i = 0 \quad \alpha_i > 0 \quad i = 1,2,3\cdots,n$$

经过求解，获得决策函数：

$$f(x) = \text{sgn}\{\sum_{i=1}^{n} \gamma_i \alpha_i (x \cdot x_i) + b\}$$

sgn(·) 代表一种符号函数，b 表示阈值。

线性不可分的情况，一般进行非线性变换，通过把样本转化为在高维特征空间使其线性可分。一般处理过程中采用核函数 $K(x, x_i)$ 来实现样本变换，引入常数 $C(C > 0)$ 使最优分类问题转化为二次规划问题：

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j \gamma_i \gamma_j K(x_i, x_j)$$

$$\text{s. t.} \quad \sum_{i=1}^{n} \gamma_i \alpha_i = 0 (0 \leqslant \alpha_i \leqslant C, i = 1,2,3\cdots,n)$$

经过求解，得：

$$f(x) = \text{sgn}\{\sum \alpha_i \gamma_i K(x_i \cdot x) + b\}$$

在企业财务预警模型的应用中，假设预警指标有 m 个，即 *SVM* 的输入为 m 维变量，判别企业是否危机的指标为 n 个，即 *SVM* 的输出为 n 维变量。取 p 个企业数据作为研究总体 $S = \{x_i, y_i \mid i = 1, 2, 3\cdots, p)\}$ $x \in Rm$，$y \in Rn$。对于总体 S，取出一部分作为训练数据，另一部分样本作为验证数据。因此，企业财务预警问题的本质即针对训练样本空间，寻求最优分类面，继而确定核函数和参数，从而获得问题的决策函数。该决策函数的学习能力可以通过训练数据检验，泛化预测能力一般根据验证数据检验。在复杂社会环境中，企业与利益相关者形成了复杂的社会网络，各者之间存在着各种各样的关系，企业的物流、信息流和资金流等存在不确定性，种种因素决定了企业财务危机具有复杂性的特征。所以预测上市公司财务状况模型的输入变量，与预警输出变量是非线性关系，因此引入满核函数进行非线性映射处理。通常而言，RBF 核是合理的首选。这个核函数将样本非线性地映射到一个更高维的空间，与线性核不同，它能够处理分类标注和属性的非线性关系。并且，线性核是 RBF 的一个特例（Keerthi and Lin 2003），因此，使用一个惩罚因子 C 的线性核与某些参数（C，γ）的 RBF 核具有相同的性能。同时，Sigmoid 核的表现很像一定参数的 RBF 核（Lin and Link 2003）。RBF 函数：$K(x_i, x_j) = e - \gamma \|x_i - x_j\| 2, \gamma > 0$，通过选择核函数和其他参数，经过训练得到企业财务危机预警的非线性支持向量机模型，其决策函数为：

$$f(x) = \text{sgn}\{\sum_{i=1}^{n} \alpha_i \gamma_i K(x \cdot x_i) + b\}$$

选择 SVM 模型，以 BP 神经网络为对比对象，SVM 具有一定的优势：首先，SVM 无需考虑太多的参数，只有惩罚函数和核函数，而 BP 神经网络的控制参数较多。如隐藏层的数量、节点的数量、学习的速率等等，这些参数一般凭经验选择。获得最优预测效果的参数组合相对很困难。其次，支持向量机的学习训练，能够保证获得的全局的最优解是唯一的。而 BP 神经网络，训练的结果无法确保唯一性。因此在比较不同指标引入预测模型效果的时候，BP 算法不易确定不同指标对结果的影响。

（五）实证研究和结果分析

为了确保模型在应用过程中有稳定的预警能力和能够满足大数据的程序处理，对 SVM 所选择的核函数处理非线性关系能力和 SVM 并行算法进行测试。

在 matlab 环境中构建具有非线性关系的散点如图 5，其中可以看到两种类型的散点各 900 个，由于目标是测试核函数在支持向量机中的有效性，所以在选择散点时，无需考虑散点的具体含义，即这些散点没有具体的经济意义，但从散点分布上可以看出具备一定的非线性可区别关系。

SVM 选择 RBF 核函数，其中 C = 1，sigma = 0.1，对上述散点进行分类，即利用 SVM 对散点进行边缘划线区分，分类结果如图 6 所示。从图中可以直观看出，选择 RBF 核函数，虽然边缘附近有一些区分误差，但总体上可以对具有非线性区别关系的变量进行很好的区分，因此该核函数应该能够满足引入大数据指标的财务预警模型应用。

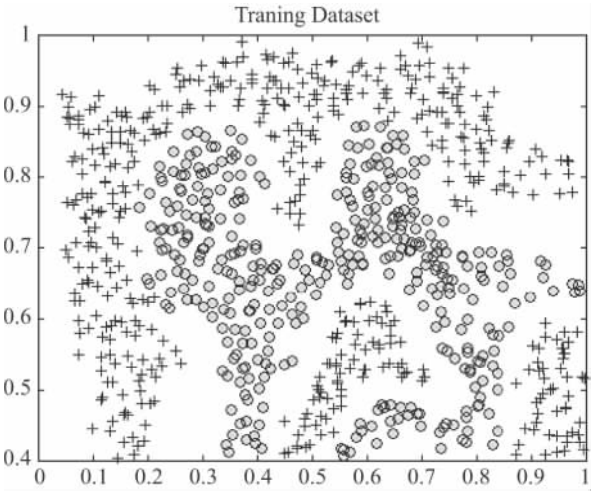

图 5　随机散点图

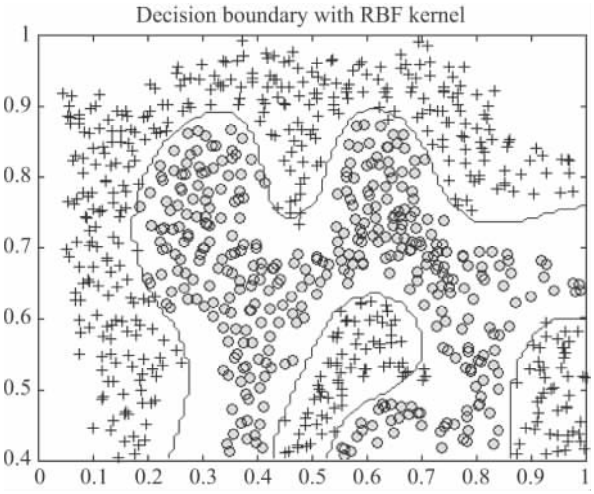

图 6　SVM 分类结果图

将相关指标具体数值引入模型，通过 SVM 模型得到的结果如表 8 以及图 7 所示。

表 8　　效果对比情况

模型	财务指标型	引入积极情绪和频次模型
T－2	90.909 1%	93.939 4%
T－3	69.697 0%	78.787 9%

从图 7 中左图可以看出，当采用财务指标进行财务预测时，根据 t－3 年的数据，正常企业中有 8 个被误判，判正率达到 63%，ST 企业中有 2 个被误判，判正率达到 81%。通过图 7 中右图可以看出，采用融入积极情绪的指标进行预测时，根据 t－3 年的数据，正常企业中有 5 个被误判，判正率达到 77%，ST 企业中有 2 个被误判，判正率达到 81%。

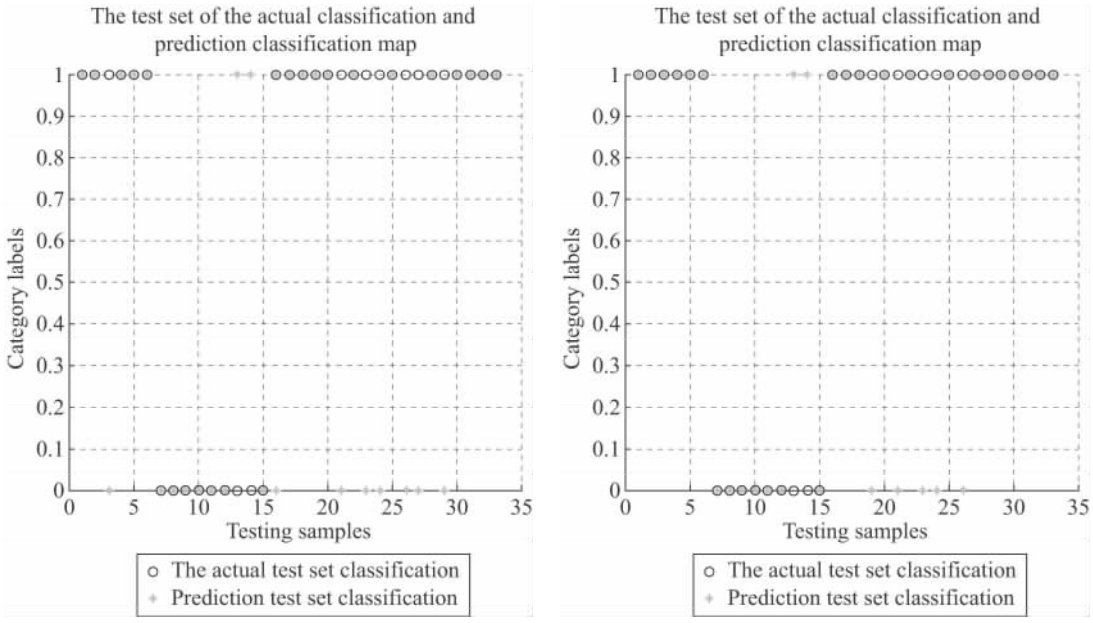

图 7　t－3 年效果比较

从图 8 中左图看出，采用财务指标进行预测，根据 t－2 年的数据，正常企业中有 2 个被误判，判正率达到 90%，ST 企业中有 1 个被误判，判正率达到 90%。通过图 8 右图可以看出，采用融入积极情绪的指标进行预测时，根据 t－2 年的数据，正常企业的判正率达到 95%，ST 企业的判正率达到 90%。

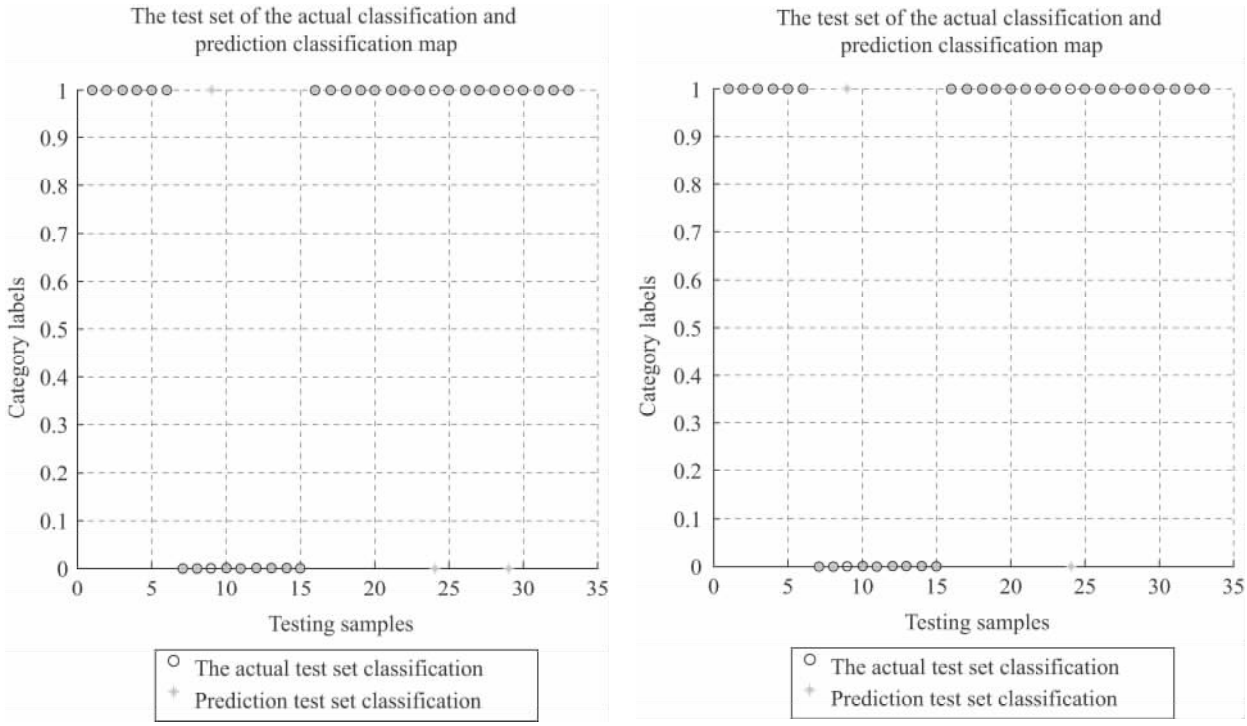

图 8 t－2 年效果比较

虽然中性情绪、负面情绪和交互情绪指标都不显著，但有一定现实意义，因此逐个带入 t－2 年模型，中性情绪和负面情绪对已经加入大数据指标的模型没有影响，结果发现交互情绪对已经引入大数据指标的模型预警效果有所提高，如表 9。

表 9 效果对比表

模型	引入积极情绪和频次模型	引入交互情绪和频次模型
T－2	93.939 4%	96.969 7%

从图 9（详见下页）左图可以看出，正常企业的判正率达到 95%，ST 企业的判正率达到 90%，当加入交互情绪指标后，正常企业的判正率没有变化，ST 企业的判正率达到 100%。大数据模型在预测有效度上有很大程度的提高，在实验样本中 T－2 期和 T－3期的判正率均比财务指标模型要高一些。针对实验样本来说，T－3 期的大数据指标引入的影响更大一些，T－2 期的大数据指标引入的影响就逐渐变小。这说明，引入大数据指标的预测模型在中短期内有一定程度的提高，同时在长期预测能力方面，明显高于财务指标模型。

五、结论

本文以网民为企业传感器，利用网络在线的大数据信息，依靠大数据涵盖范围广泛、体现群体智慧和不易被修改的特点，引入了大数据指标建立财务预警模型。通过对 2012、2013 年 60 家企业进行全网信息的过滤和爬取，进行了企业相关大数据信息指标的整理，通过与财务指标的结合，对研究假设进行实际数据验证，发现引入大数据指标的财务预警模型，相对财务指标预警模型，在短期内对预测效果有一定提高，从长期来看，对预测效果有明显提高，大数据指标在误警率和漏警率上比财务指标表现明显要好，从而验证了在复杂社会环境中，依靠大数据技术加强信息搜寻是提高财务预警有效性的重要路径这一观点。

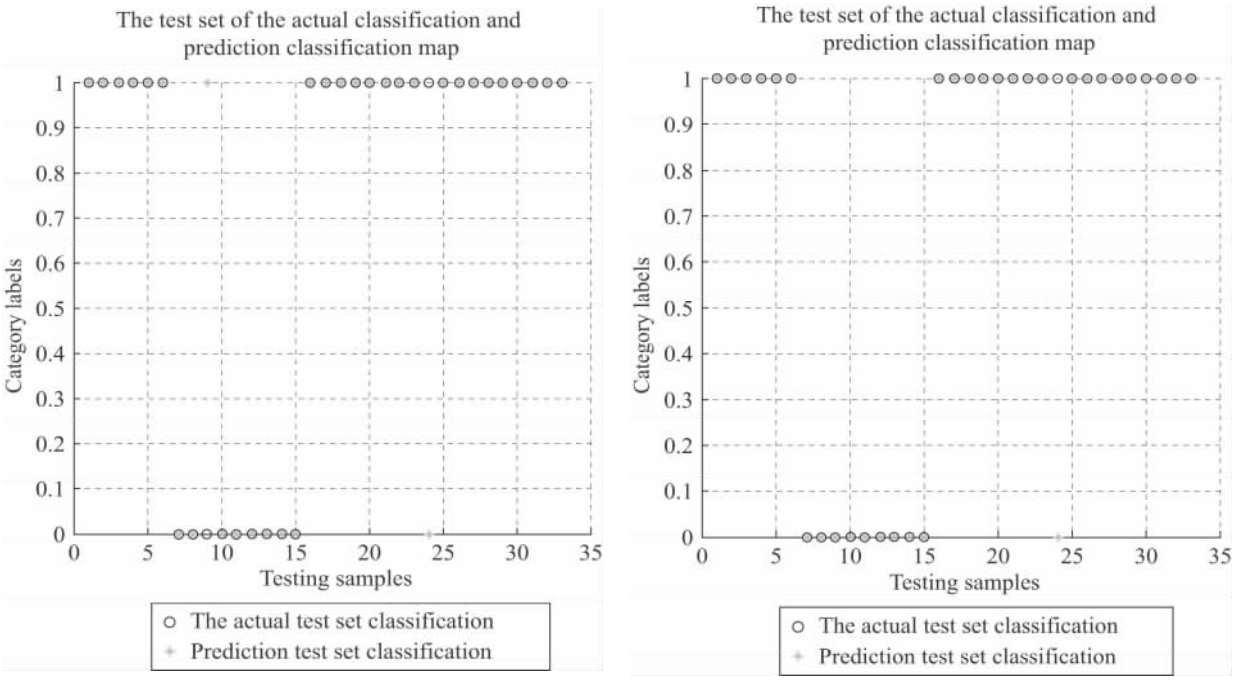

图 9　t－2 年含交互情绪比较

参考文献

[1] 吴星泽．财务危机预警研究：存在问题与框架重构 [J]．会计研究，2011（2）：59－65.

[2] 吴星泽．财务预警的非财务观 [J]．当代财经，2010（4）：122－128.

[3] 李国杰．大数据研究：未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考 [J]．中国科学院院刊，2012（6）：647－656.

[4] 张鸣．企业财务预警前沿 [M]．北京：中国财政经济出版社，2004.

[5] Altman. E1I1，Financial Ratios Discriminant Analysis and the Prediction of Corporate Bankruptcy [J]. Journal of Finance，1968，23（4）：589－6091.

[6] Johnson Craig G. Ratio Analysis and the Prediction of Firm Failure [J]. The Journal of Finance，1970，25（5）：1166－1168.

[7] Ohlson，J1，Financial Ratios and the Probabilistic Prediction of Bankruptcy [J]. Journal of Accounting Research，1980，18（1）：109－1301.

[8] Elloumi F，Gueyie J P. Financial Distress and Corporate Governance：An Empirical analysis [J]. Corporate Goernance，2001（5）：15－23.

[9] Compbell J Y，Hilscher J，Szilagyi J. In Search of Distress Risk [J]. The Journal of Finance，2008，63（6）：2899－2939.

[10] E. Ostrom. Building Trust to Solve Comments Dilemmas：Taking Small Steps to Test an Evolving Theory of Collective Action [M]. New York：Springer，2008：211－216.

[11] Viktor Mayer-Schonberger. Big Data：A Revolution That Will Transform How We Live，Work，and Think [M]. 浙江人民出版社，2013.

[12] Zhang X，Fuehres H，Gloor P. Predicting Stock Market Indicators Through Twitter－“I Hope it is not as Badas I Fear”. In COIN Collaborative Innovations Networks Conference，2010：1－8.

[13] Tobias Preis1. Quantifying Trading Behavior in Financial Markets Using Google Trends [R]. Scientific Reports，2013.

[14] Zmijewski Mark E，Dietrich J. Richard. Methodological Issues Related to the Estimation of Financial Distress Pre-diction Models [J]. Journal of Ac-counting Research，1984.

（责任编辑：韩　嫄）