基于机器学习技术的中小银行风险早期预警系统研究

王博　鄢若兰

（南开大学金融学院　天津市　300350）

摘　要：在宏观经济环境承压的背景下，中小银行风险水平上升，与大银行“大而不能倒”不同，中小银行是“多而不能倒”。本文通过优化银行风险评估标准和预警指标体系，选取2007–2019年我国宏观数据和183家中小银行的财务数据、监管指标，基于C5.0决策树、随机森林算法及支持向量机构建我国中小银行风险早期预警系统。实证研究发现，三种模型对样本外数据集均表现出很高的预测准确率，其中随机森林与支持向量机模型表现较优。这是基于此三种机器学习技术识别我国单体银行风险状态的首次尝试，将为监管部门预防、预警和处置中小银行风险提供实用的技术手段。

关键词：银行风险；早期预警系统；决策树；随机森林；支持向量机

中图分类号：F832.3　**文献标识码**：A　**文章编号**：1007–4392(2022)03–0073–13

一、引言

随着利率市场化改革的推进和金融市场逐步对外开放，银行间竞争必然加剧，银行经营活动的不稳定性将会提高。从总体上看，我国中小银行近年来风险水平陡增主要有如下几种原因：一是公司治理不完善，缺乏公司治理的良好实践和基本原则（周小川，2020）[1]，风险主要来源于股权结构分散导致的内部人控制风险、股东干预公司治理风险和股东资金占用风险。二是经营业务及产品较为单一，高度依赖利息收入，2019年以来尤其是新冠疫情之下，货币政策结构性宽松，利率水平下行，加之大型商业银行贷款利率走低挤压市场空间，对中小银行利润来源形成明显制约。三是以城商行和农商行为主的中小银行定位于区域金融供给，主要服务于当地民营及小微企业，信用下沉程度较高，天然面临着较高的信用风险，受经济周期和信用周期影响较大。如果中小银行产生风险，因其庞大的数量和风险传染机制，

基金项目：本文受到国家自然科学基金面上项目“外部冲击对中国金融稳定的影响机理：不确定性与公共事件冲击视角”(72073076)、国家自然科学基金面上项目“基于大数据的中国金融系统性风险测度及其演化规律研究”(71873070)、国家社科基金重大专项“我国债务危机风险的防范治理与有效缓解对策研究”(18VFH007)、国家社会科学基金重大项目“基于结构化数据分析的中国金融系统性风险防范体系研究”(17ZDA074)资助。

可能导致系统性金融风险的发生。

此外,2008年国际金融危机之后,防范、恢复和处置金融机构危机的重要性被关注。巴塞尔银行监管委员会、金融稳定理事会及英、美、欧盟等主要经济体纷纷制定了恢复和处置计划的监管指引。对中小银行的风险识别和早期预警对于防范银行业系统性风险蔓延十分必要,构建符合我国中小银行经营特点和监管框架的风险早期预警系统对于增强银行风险管理、强化银行审慎经营和防范化解金融风险都具有重要意义。

二、文献综述

(一)银行业系统性风险及银行风险预警指标体系

银行系统性风险具备负外部性、风险与收益不对称、传染性、损害实体经济和干预投资者信心五个特征,主要通过宏观经济的外部因素,银行间存款、信贷和支付体系等实际业务和由信息引发的银行危机的溢出效应三种渠道进行传播(范小云,2004)[2]。在西方主要经济体经历了数次银行业危机后,各国的中央银行意识到了银行业"太大而不能倒(Too-big-to-fail)"的问题,规模较大的银行倒闭容易引发银行系统性风险(徐超,2013)[3]。1984年9月,在救助伊利诺伊大陆银行后,"太大而不能倒"的重要性含义逐步被监管当局明确,该原则逐渐在银行领域流行起来,1991年颁布的《联邦存款保险公司促进法案》在立法上最终确立了该原则。但此后以规模评判系统重要性的救助标准一直被学界争议,监管当局和学界关注的焦点转向其他判断银行系统重要性和救助标准的方法。Zhou(2009)[4]在一个互联金融系统中测度一个金融机构的系统重要性,提出金融机构的规模作为"太大而不能倒"的系统重要性的衡量标准不是一直有效的,并提供了在多元极值理论框架下系统重要性的测度方法。Rajan(2009)[5]提出阻止金融机构变得"太过系统重要而不能倒(Too-systemic-to-fail)"的三种方法以防范系统性金融风险的发生和蔓延。Acharya和Yorulmazer(2007)[6]发现各国的银行倒闭政策同样存在一个隐含的"太多而不能倒(Too-many-to-fail)"的问题,当倒闭银行数量很多时,监管机构往往会救助部分或全部的倒闭银行。但当倒闭银行数量较少时,倒闭的银行往往由健康存活的银行收购。这样的隐含政策倾向增强了银行从众和选择一起倒闭的动机,且对小银行的影响更大并引发小银行中的羊群效应。Brown和Dinc(2009)[7]使用竞争风险模型研究新兴市场国家的银行破产问题,发现当银行体系较为脆弱时,政府接管或直接关闭破产银行的可能性较小,这种"太多而不能倒"的效应是稳健存在的。2008年金融危机提醒了人们系统性风险的主要来源是金融机构之间的相互关联性,监管当局和学者们开始关注"过于关联而不能倒(Too-connected-to-fail)"的系统重要性银行和关联性问题。Gabrieli(2012)[8]发现在2007年8月之后,银行的相互关联性影响增大,并从正外部性转为负外部性,很好地解释了本次金融危机影响面广、传染性强的原因。范小云等(2012)[9]构建网络模型分析银行间负债关联程度对于银行系统重要性的影响,发现我国部分中

小银行如果倒闭，能够诱发严重的银行危机，因此在银行风险监管中，应该考虑“过于关联而不能倒”（即关联性）问题。我国中小银行数量众多，近些年来，业务同质化严重，贷款集中度过高，资本充足水平低，经营管理理念激进，行业竞争激烈，导致系统性风险水平不断上升。特别是近年来同业、理财和金融市场等经营模式提高了中小银行的网络中心度和关联性，扩大了风险传染的网络效应，使得风险在银行间更迅速地传播，形成了中小银行“太多而不能倒”和“过于关联而不能倒”的局面。为了有效降低传染风险，就必须在风险诱导因素发生前进行事前预防（马君潞等，2007）[10]。

（二）银行风险早期预警模型构建

国外学者对风险早期预警模型的研究开始较早，采用的方法众多，研究脉络相对完整和成熟。Altman（1968）[11]使用多重判别分析方法，建立5变量Z-score模型来判断企业破产的可能性。Martin（1977）[12]引入Logisitic回归分析方法研究银行的违约概率。West（1985）[13]使用因子分析和Logit回归方法，借鉴CAMELS评级法对美国银行资产负债表和银行监管报告中提取的变量进行分析。Lane等（1986）[14]使用Cox模型预测银行倒闭的时间。Chiaramonte等（2015）[15]比较了Z-score模型和Probit模型识别银行金融风险的有效性。Rosa和Gartner（2017）[16]采用多元Logisitic回归方法构建巴西国家所有和州所有银行预测窗口为1年的风险早期预警模型。Ferriani等（2019）[17]使用Logit模型计算意大利非重要金融机构的违约概率，构建预测范围为4-6个季度的早期预警模型。Brauning等（2019）[18]采用决策树算法构建欧洲银行中小银行危机早期预警模型。Suss和Treitel（2019）[19]使用英国金融管理局的机密监管数据和风险监管评估分数，比较了混合Logistic回归、线性随机效应模型、随机森林及Boosting技术、K最近邻算法（KNN）、支持向量机（SVM）以及组合模型对风险状态的预测能力。

国内相关研究中，杨保安和季海（2001）[20]利用人工神经网络前向三层BP网络方法提取财务指标构造商业银行信贷风险预警系统，通过BP网络的输出向量判断银行信贷风险警情。贺晓波和张宇红（2001）[21]运用聚类分析法筛选指标，使用熵值法和层次分析法确定指标权数，计算综合分值，使用信号灯显示法进行银行风险状态判定。毛锦等（2006）[22]根据对银行信贷交易的风险分析构建商业银行信用风险预警的概念模型。中国银行业监督管理委员会银行风险早期预警综合系统课题组（2009）[23]综合了扩散指数、合成指数、降级距离和百分位排序方法并进行有效性检验，构建短期和中长期单体银行风险预警模型。罗晓光和刘飞虎（2011）[24]使用主成分分析法对商业银行风险预警指标进行赋权并计算综合评价分数，构建Logistic预警模型。陆静和王捷（2012）[25]采用贝叶斯网络测算各类指标对银行风险的影响程度，通过信号灯模型显示银行风险状态。王伟（2013）[26]采用专家咨询法和层次分析法构建银行危机预警系统指标体系，通过指标映射法计算单项指标警情分值，加权汇总后形成

银行危机综合预警分值。丁德臣(2016)[27]利用多专家协商机制确定预警指标权重,使用灰色评价模型预测商业银行风险评级。

综上,国外学者对银行风险早期预警系统的研究起步较早,研究脉络完整,不断更新技术提高对银行风险状态的预测能力,目前研究已经较为成熟。但是国内学者在银行风险早期预警研究上比较滞后。在技术上,我国学者还大多采取较为主观、定性的方式对风险指标进行赋权,计算综合风险分值,且不能实现对风险的事前预警,不能为银行经营及时纠偏和监管部门尽早干预预留时间窗口,机器学习技术在此领域的应用尚不成熟。在数据方面,我国银行业尤其是中小银行对经营数据和监管指标的披露不完全,从公开渠道获取的数据极为有限。因此本文旨在采用合适的银行风险状态评估标准,针对国内中小银行构建风险早期预警模型,引入决策树、随机森林算法和支持向量机,在研究方法上填补空白。

三、研究设计

(一)国内银行风险状态评估标准

为向监管者和政策制定者提供有效的银行风险早期预警模型,首先银行风险状态评估标准要符合国内银行经营特点,其次须为政府监管和干预预留时间窗口。如果仅将银行风险状态定义为银行破产倒闭或流动性危机,那么模型只对银行危机的事后干预有效而没有预测意义。此外,也有学者提出不以发生恐慌而以银行资本紧缩作为判断银行陷入风险状态,因为即使没有发生恐慌,银行股本大幅下跌仍会导致严重的信贷紧缩和产出缺口,恐慌只是银行发生风险事件的结果而非原因。有时恐慌因为监管机构宽容、债权人隐性担保或政府强有力干预而免于发生,但银行依然处于资本不足、放贷能力严重受损的风险状态(Baron等,2020)[28]。目前国内仅有1998年海南发展银行破产清算、2019年包商银行被接管两例明显银行危机状态,银行破产在国内较为罕见。

因此本文不采用仅将大规模恐慌、银行违约、被兼并收购、被接管或政府援助作为银行陷入风险状态的传统做法,而是结合国内银行经营特点和监管框架及条例,从经营数据和监管指标的视角来考察银行风险事件,一是能提供客观、定量、实时的银行风险事件评估标准;二是能挖掘被遗忘和忽视的"沉默"银行风险事件,扩大银行风险事件研究样本;三是可以更敏锐地捕捉银行陷入风险状态的早期迹象,有利于提高银行风险早期预警系统的预测精度。

在宏观审慎评估体系(MPA)中,对资本和杠杆情况与定价行为两个方面要求最严格,只要有一项不及格,MPA考核便不达标。以资本充足率考核为例,如果某银行资本充足率大于宏观审慎标准,该项指标为满分80分,如果某银行资本充足率小于宏观审慎标准但超过4%,得分在48至80分之间,如果不超过4%,则得分为0。在央行取消4%的缓冲范围之后,只要银行达到宏观审慎标准就获满分,否则为0分,这种评价方法较为严格,能敏感地判别银行风险状态。因此,本文也采用这样严格的评价标准判别银行风险状态,有利于捕捉银行的风险预警信号。

根据《商业银行资本管理办法》《商业银行风险监管核心指标（试行）》《商业银行风险监管核心指标一览表》《商业银行监管评级内部指引》，本文认定银行触发下列条件即被认为银行风险事件：1.拨备充足率<100%；2.贷款损失准备充足率<100%；3.拨备覆盖率<120%；4.拨贷比<1.5%；5.存贷款比例>75%；6.资产利润率<0.6%；7.资本利润率<11%；8.流动性比例<25%；9.不良贷款率>5%；10.单一客户贷款集中度>10%；11.成本收入比>35%；12.核心一级资本充足率<5%，资本充足率<8%。

（二）数据来源

本文样本覆盖国内股份制银行、城市商业银行、农村商业银行和村镇银行183家银行，从Wind数据库获取2007–2019年的银行财务数据、重要业务指标和银行业专项分析指标，从国泰安数据库、Wind数据库获取宏观经济数据。

（三）制定风险预测指标

鉴于我国《股份制商业银行风险评级体系（暂行）》参考了CAMELS评级法，本文也采用CAMELS评级法的框架，参考我国商业银行监管条例选取微观公司数据和宏观经济数据作为风险预测指标。基于数据可得性，初始选取60个指标，分别进行如下回归，删去方差膨胀因子VIF≥10的指标，最终选取预测指标见表1。

$$X_1=\alpha+\beta_2 X_2+\cdots+\beta_k X_k, VIF_1=\frac{1}{1-R_1^2} \quad (1)$$

（四）实证分析

1.使用C5.0决策树算法。决策树是在机

表1　预测指标选取

指标类型	指标名称
资本充足性	资本充足率
	核心一级资本充足率
	ln（归属于母公司所有者权益）
	拨备覆盖率
	拨贷比
	资产负债率
资产质量	不良贷款比例
	平均风险权重
	单一最大客户贷款比例
管理水平	成本收入比
	NtoI
	StoA
盈利水平	资产利润率
	资本利润率
	存贷款比例
	净息差
	净利差
	生息资产 / 总资产，生息负债 / 总负债
	贷款 / 生息资产，同业资产 / 生息资产 存款 / 计息负债，同业负债 / 计息负债
流动性	流动性比例
市场风险敏感度	AVGNtoI
	AVGStoA
宏观经济数据	GDP增速（季度 / 半年度）
	财政赤字 /GDP
	3个月上海银行间同业拆放利率（SHIBOR）–3个月国债到期收益率
	10年期国债到期收益率–3个月国债到期收益率
	房地产业对金融业的超额收益率（上证地产指数 – 上证180金融股指数）
	上证综指收益率（季度 / 半年度）
	上证综指30天历史波动率

注：NtoI= 非利息收入 / 利息收入×100%，StoA= 证券资产 / 总资产×100%，AVGNtoI= 市场平均非利息收入 / 利息收入×100%，AVGStoA= 市场平均证券资产 / 总资产×100%

器学习中被广泛使用的分类预测算法之一。决策树的生长过程本质上是将训练样本根据各个特征逐次划分在多个类别下的过程(即分枝),在所有可能的分枝中,选择结果最好的一种,分枝的质量和评价预测结果的好坏标准由节点随机变量的纯度决定,纯度由节点随机变量的信息熵来度量。不确定性越大,信息熵越大,纯度越低。在决策树分枝过程中,设T为待训练样本集合,有k个不同的类别,类别集合为$\{c_1,c_2,\cdots,c_k\}$,则类c_i中包含的信息量表示如下:

$$I(c_i)=-\log_2 p_i \tag{2}$$

样本集合T中属于类c_i下的子集为R_i,R_i中包含的样本个数为r_i,样本集合T中包含的样本总数为|T|,那么待分类的数据集合T中用来分类的子集所包含的信息定义如下:

$$I(r_1,r_2,\cdots,r_k)=-\sum_{i=1}^{k} p_i \log_2 p_i \tag{3}$$

I(r)表示样本集合T包含的随机变量的信息,p_i表示任意抽取样本属于类c_i下的概率,$p_i=\frac{r_1}{|T|}$。

假定某一属性M有互不重合的n个取值$\{m_1,m_2,\cdots,m_n\}$,把样本集合T分为n个子集$\{T_1,T_2,\cdots,T_n\}$,其中T_j是样本集合T中当属性M取m_j时的子集。在决策树分类过程中,属性M将样本集合T中各样本分在不同类别中,T_{ij}表示T_j子集中属于类c_i的样本个数,那么属性M对分类$\{c_1,c_2,\cdots,c_k\}$的熵为:

$$H(M)=\sum_{i=1}^{n}\frac{T_{1j}+T_{2j}+\cdots,+T_{kj}}{|T|}I(T_{1j},T_{2j},\cdots,T_{kj}) \tag{4}$$

其中,令子集T_j在样本集合T中所占的比重为$W_j=\frac{T_{1j}+T_{2j}+\cdots+T_{kj}}{|T|}$,属性M的各个取值对应的类$c_i$包含的期望信息量$I(T_{1j},T_{2j},\cdots,T_{kj})=-\sum_{i=1}^{k} p_{ij}\log_2 p_{ij}$,$p_{ij}=\frac{T_{ij}}{T_j}$。

由上述公式,属性M作为分类属性的信息增益(Information Gain)为:

$$Gain(T,M)=I(T_1,T_2,\cdots,T_k)-H(M) \tag{5}$$

不同决策树算法的区别在于分枝准则,C5.0算法是根据特征变量的信息增益率(Information gain ratio)来决定分枝的选择。如果信息增益率够大,就分裂为左右子树;如果信息增益率很小,就停止分裂,这个节点直接作为叶子节点。信息增益率(Information Gain Ratio)由信息增益(Information Gain)和分裂信息度量(Split Information)共同定义:

$$\text{Split Information}(T,M)=-\sum_{i=1}^{k}\frac{|T_i|}{|T|}\log_2\frac{|T_i|}{|T|} \tag{6}$$

$$\text{Information Gain Ratio}=\frac{Gain(T,M)}{\text{Split Information}(T,M)} \tag{7}$$

通过一系列的特征选取和决策规则,节点分枝达到最终的分类叶子节点,分类算法经过递归最终输出结果是一棵决策树。对于本文的预测实证分析来说,该过程是从根节点开始,基于一系列预测指标的值创建一条路径直到到达叶子节点,最终判断银行是否处于风险状态。

本文模型设定为:一是采用二元分类,若银行财务数据及监管指标触发本文对银行风险状态评估标准条目,视为银行陷入风险状态,标记为1,否则为0。二是设置预测期为6个月,以银行陷入风险状态的前一期状态作为目标变量,以保证模型具有前瞻性

和预测性。三是考虑到监管当局往往会采取保守和谨慎的态度，宁愿发出错误的预警警报，也不愿遗漏报告银行的风险状态而使事态恶化，本文设置不对称的误判成本，将第一种错误（遗漏报告风险状态）的误判成本设置为第二种错误（错误报告风险状态）的误判成本的 2 倍（Brauning 等，2019）[18]，误判成本矩阵见表 2。

表 2　误判成本矩阵

实际＼预测	正常	遇险
正常	0	1
遇险	2	0

通过数据预处理，最终得到 183 家中小银行的数据，首先将全部样本集按照 4:1 分为训练集和预测集，使用训练集数据通过 K 折交叉验证法确定参数最优值，然后使用预测集数据进行样本外检验，得到模型对预测集的分类准确率和 AUC 值。

2.使用随机森林算法。随机森林是将决策树和 Bagging 算法（Bootstrap aggregating）相结合的一种算法，采用随机有放回的选择训练集构造分类器的集合学习，Bagging 算法通过降低基分类器的方差从而改善泛化误差，从而提高算法的准确率和稳定性，避免过拟合的发生。随机森林的基本单元是决策树，它建立多个决策树并整合每一棵决策树的分类结果以得到更准确和稳定的预测。随机森林算法流程见图 1。

随机森林模型的两个主要参数是构建决策树节点分枝选择的特征变量个数和随机森林模型中决策树的数量，仍然使用训练集数据通过 K 折交叉验证法确定最优的特征变量个数和决策树数量，然后使用预测集数据进行样本外检验，使用分类准确率和 AUC 值来评估模型的预测能力。

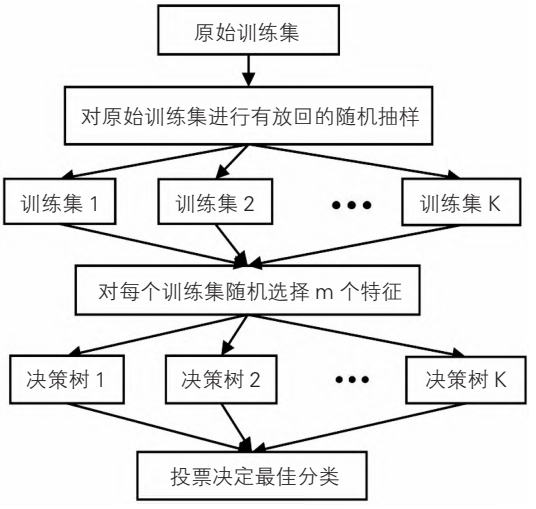

图 1　随机森林算法流程图

3.使用支持向量机。支持向量机是一种二元分类模型，当有两个及以上的预测变量时，使用超平面来分离两个类别。它的基本模型是定义在特征空间上间隔最大的线性分类器，也可以通过设置不同形式的核函数处理线性不可分的训练数据，形成非线性支持向量机。支持向量机的学习策略就是使间隔最大化，并将其形式化为一个求解凸二次规划的问题，等价于正则化的合页损失函数的最小化问题。常见的核函数包括线性核、多项式核、高斯核、Sigmoid 核等。由于在以前的银行破产预测研究中，James 等（2013）[29]发现径向基核函数表现要优于其他核函数形式，因此本文采用的是径向基核函数。支持向量机的两个主要参数是成本即分类错误的代价和影响径向基核函数的复杂度 γ，同样使用训练集数据应用 K 折交叉验证法进

行调参,使用最优参数对预测集数据进行样本外检验,查看模型的分类准确率和 AUC 值。

四、模型预测结果及分析

(一)半年度数据训练及预测结果

首先使用训练数据集进行样本内简单交叉验证,使用默认 Boosting 迭代次数 1,得到模型的预测准确率为 84%,AUC 值为 0.863;然后通过 K 折交叉验证确定最优 Boosting 迭代参数为 21,模型的预测准确率提升至 86%,AUC 为 0.944。接着使用随机森林算法,通过 K 折交叉验证确定决策树节点分枝变量个数为 9,决策树数量为 2400,样本内模型预测准确率为 90.53%,样本外数据集预测准确率为 88%,AUC 为 0.911,较决策树模型预测准确率均略有提高。最后使用支持向量机算法,通过 K 折交叉验证确定最优参数 Gamma 为 0.031,Cost 参数为 1,样本内模型预测准确率为 93.06%,样本外数据集预测准确率为 89.23%,AUC 为 0.938。此外,本文还采用 Logit 回归模型作为基准模型,以衡量各个算法模型的预测准确率,Logit 模型对半年度数据的预测准确率为 83.21%,AUC 为 0.869,比较可得机器学习技术在预测能力均要优于 Logit 回归模型,在对银行半年度数据样本进行风险状态的分类预测中,支持向量机的预测准确率是最高的,AUC 值也较高,模型性能是最优的,预测结果见表 3,ROC 曲线见图 2 至图 5。

表 3 中小银行半年度数据预测结果

	样本内数据预测准确率	样本外数据预测准确率	AUC
Logit 回归	–	83.21%	0.869
默认 C5.0 决策树	–	84.00%	0.863
BoostingC5.0 决策树	88.89%	86.00%	0.944
随机森林	90.53%	88.00%	0.911
支持向量机	93.06%	89.23%	0.938

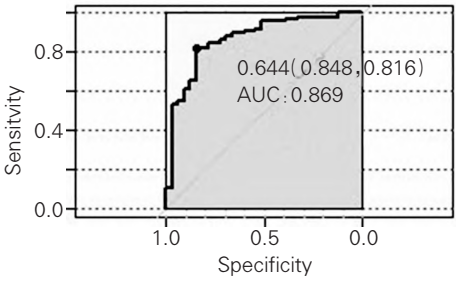

图 2 Logit 模型 ROC 曲线

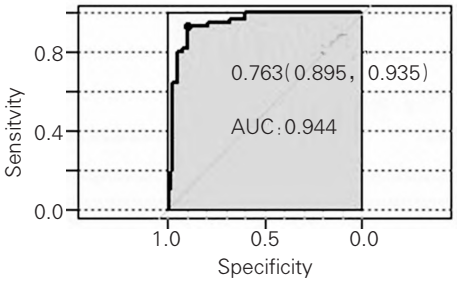

图 3 C5.0 决策树模型 ROC 曲线

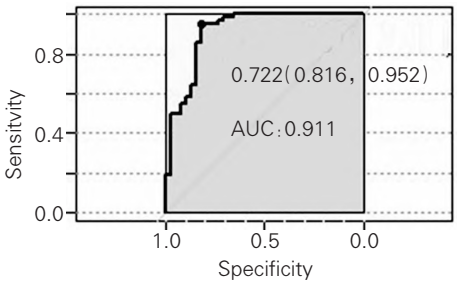

图 4 随机森林模型 ROC 曲线

(二)模型预测结果分析

1.决策树模型预测结果分析。C5.0 算法构建的决策树不仅提供早期识别风险的手段,也能揭示影响银行风险状态的重要变量。在本文构建的决策树中,根节点的分枝变量是 GDP 增长率,说明宏观经济状况对于银行经营有着巨大的影响。余下的较为重要的变量覆盖了几种最重要的银行风险的指

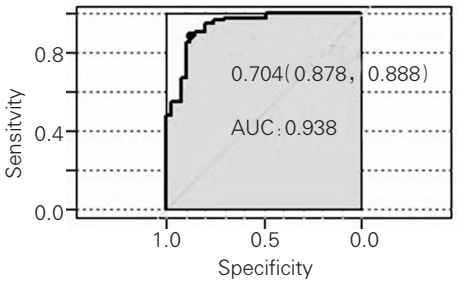

图5　支持向量机模型ROC曲线

标，比如信用风险，包括不良贷款率、平均风险权重、单一最大客户贷款比例；流动性风险，包括流动性比例；市场风险，包括市场平均非利息收入与利息收入比率、银行非利息收入与利息收入比率；资本风险，包括资产负债率、资本充足率等。

探究决策树的主要生长路径，当宏观经济运行较差时，银行是否陷入风险状态很大程度上取决于银行的盈利能力，如果银行的净息差和资产利润率都很高，那么银行在宏观经济状况不好时仍然可以顺利存活。反之，如果银行的净息差较低，那么即便有较大比例的生息资产，但并未发挥其为银行盈利的作用，银行也会陷入风险。对于区域性中小银行来说，生息资产主要为地方企业贷款，当宏观经济状况不好时，当地企业在银行中的贷款质量会迅速下降，甚至成为呆账坏账，从而给银行带来风险。如果此时银行的存款占全部负债的比例较高，有较为稳定的存款基础，那么银行仍然可以处于较为安全的状态。

当银行生息资产占总资产比例较低时，资产质量的相关指标则决定了银行是否处于风险状态。如果银行资产的平均风险权重较高，盈利能力方面净利差也较低，那么银行就会陷入风险状态之中。如果银行的资产的平均风险权重较低，不良贷款率也较低，那么此时银行的低成本收入比说明银行有着较为良好的管理经营水平，银行也会免于陷入风险。当银行的不良贷款率较高时，如果银行个体的非息收入占比和存贷比较高，不过于依赖传统的放贷业务进行盈利，那么银行也会处于相对安全的状态。

在GDP增长率较高即宏观经济运行较为良好的状态之下，决定银行经营状态的关键变量也以资产质量和盈利能力为主。如果银行的资本利润率较高，净利差较高，而单一最大客户贷款比例和不良贷款率较低，说明银行的盈利水平和经营的安全性较高，银行陷入风险的可能性较小；反之，如果银行的资产质量指标单一最大客户贷款比例和不良贷款率均达到很高水平，银行就会陷入风险状态。而当银行的盈利水平较低时，如果银行的拨备覆盖率较高，银行的计息负债比例较高，存款基础较为稳定，银行也会依然处于健康状态；如果银行的计提拨备不充足，主要收入来源又是传统的贷款业务，银行就会陷入风险。

2.随机森林模型预测结果分析。在评估预测指标的重要性时，算法提供平均精度下降度和平均Gini指数下降度两种指标，见表4。

平均精度下降度是把一个变量的取值变为随机数，改变前后随机森林预测准确率的降低程度，平均精度下降度越大，该变量越重要。Gini指数表示节点纯度，Gini指数越大纯度越低，平均Gini指数下降度通过

表 4 随机森林模型预测指标重要性程度排序

预测指标	Mean Decrease Accuracy	Mean Decrease Gini
资本利润率	61.235253	24.551588
拨备覆盖率	44.813303	15.127262
GDP 增速	43.735355	26.841066
财政赤字 /GDP	42.137423	25.915383
不良贷款比例	29.891249	10.267250
资产利润率	25.187294	11.762404
AVGNtoI	24.001650	7.830385
归属于母公司所有者权益	23.767137	6.060924
核心一级资本充足率	19.791031	4.992221
存贷比	18.562662	5.370811
AVGStoA	17.413571	7.148856
存款 / 负债	16.594435	4.575057
净息差	15.542459	5.132352
StoA	14.806797	4.361616
单一最大客户贷款比例	14.363566	5.096307
净利差	14.357932	4.704245
上证综指 30 天历史波动率	13.487708	3.003139
房地产行业对金融行业的超额收益率	13.431647	2.823727
平均风险权重	12.930525	4.497146
NtoI	12.312152	4.739333
10 年期国债到期收益率 -3 个月国债到期收益率	12.092386	2.527500
拨贷比	11.782055	4.082415
杠杆率	10.942747	5.490261
3 个月 SHIBOR-3 个月国债到期收益率	10.870771	2.421353
上证综指收益率	10.460897	1.888179
成本收入比	9.380949	4.309349
资本充足率	8.493954	2.938980
贷款 / 生息资产	6.487688	2.873970
生息资产 / 总资产	6.098651	2.875493
同业资产 / 总资产	5.961944	4.891391
流动性比例	3.612707	2.660655
生息负债 / 总负债	2.853556	2.951511

Gini 指数计算每个变量对决策树每个节点上观测值异质性的影响得到。平均 Gini 指数下降度越大,那么变量越重要。

从分类结果来看,平均精度下降和平均 Gini 指数下降的趋势大致相同。首先,盈利水平类指标对银行风险状态影响最大,资本利润率、资产利润率、存贷款比例、净息差和净利差指标重要性排序较为靠前,如果银行盈利能力较强,则可以较好地抵御风险。资本充足性指标如拨备覆盖率、归属于母公司所有者权益以及核心一级资本充足率等指标的重要性排序也比较靠前,资本充足水平是银行破产的最后一道防线,能够冲销经营过程中因不确定性造成的损失,提高了银行经营的安全性和稳定性。紧随其后的是两个宏观经济变量即 GDP 增长率和财政赤字与 GDP 比率,意味着宏观经济环境对于中小银行的生存和风险有巨大影响,中小银行所依赖的区域性民营企业和小微企业的经营是顺周期的,因此中小银行资产质量和风险水平因此也是顺周期的。相应地,资产质量类指标,如不良资产比例,对银行风险状态的准确分类也非常重要。值得注意的是,市场风险敏感度指标也较为重要,印证了 Hirakata 等(2017)[30]的研究结论,虽然单个银行对金融资产更多地配置会提升银行资产组合异质性,从而降低银行风险,但是当市场间银行总体持有证券资产的配置在总资产中占比较高时,市场内银行总体资产组合配置会趋同,银行系统性风险上升,从而使得单个银行风险不降反升。类似地,单个银行非息收入占比提高有利于银行盈利能力增强,

但是当银行总体非息收入占比提升时，各个银行主营业务模式和盈利模式趋同，也会导致银行系统性风险提升，对单个银行造成不利影响。总而言之，银行投资组合对市场风险敞口增加以及与市场活动相关的非传统收入来源的更大依赖尽管会降低单个银行的风险，但会增加银行系统性风险，因此表现为“银行间系统性的羊群效应”，在共同的冲击来临时，提供了一种通过证券资产持有量和收入来源的风险传导方式。

（三）季度数据训练及预测结果

考虑到季度数据能够更准确地反映银行整体经营状况和风险状态变化，本文还使用季度数据来检验上述模型的结论并验证模型分类预测的有效性。

维持原模型设定不变，将预测期更改为3个月，利用148家中小银行26个预测指标的数据分别通过3种机器学习算法构建银行风险早期预警模型，使用Logit模型预测结果作为基准，得到模型预测结果如下表。同样可以发现，三种机器学习算法的样本外预测能力优于Logit分类回归模型，AUC值较高，模型性能较好。在对银行季度数据样本进行风险状态的预测中，随机森林算法对样本外数据的泛化能力最优，预测结果见表5，ROC曲线见图6至图9。

表5　中小银行季度数据预测结果

	样本外数据预测准确率	样本外数据预测准确率	AUC
Logit回归	–	84.57%	0.894
默认C5.0决策树	–	85.09%	0.546
BoostingC5.0决策树	89.32%	88.20%	0.929
随机森林	89.78%	90.68%	0.950
支持向量机	96.43%	90.12%	0.878

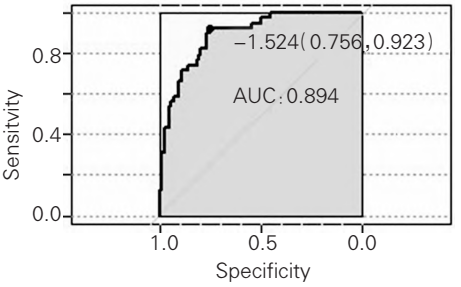

图6　Logit模型ROC曲线

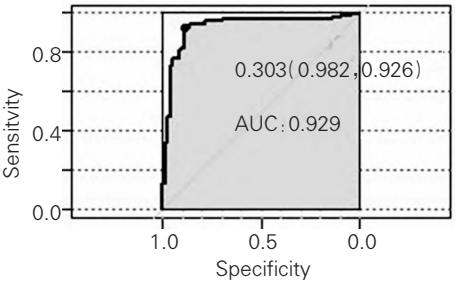

图7　C5.0决策树模型ROC曲线

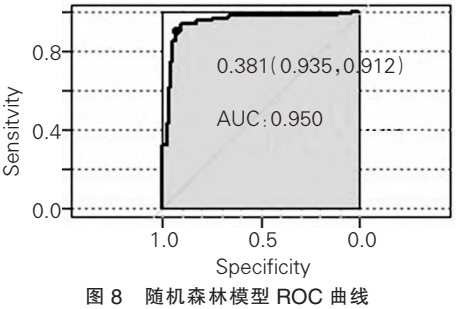

图8　随机森林模型ROC曲线

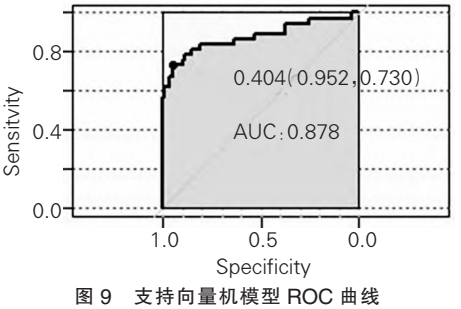

图9　支持向量机模型ROC曲线

五、结论与建议

本文使用2007–2019年我国近200家中小银行的半年度数据和季度数据2个样本集，结合国内外银行监管框架及我国监管

部门颁布的法律法规及指导性文件,确定了符合国内银行经营特点的、较为广泛且严格的银行风险状态评估标准,应用C5.0决策树、随机森林和支持向量机三种机器学习技术构建国内中小银行风险早期预警模型,实现对银行潜在风险状态的提前预警,并验证了三种算法能够将国内中小银行准确地分类为遇险和正常两种状态,能够人为设定预测时间窗口,为银行经营者及时纠偏和监管部门早期干预防范银行风险提供了实用的技术手段。三种算法构建的模型预测准确率均高于传统的Logit模型,且模型稳定性较高,其中随机森林算法和支持向量机表现更优。

第一,在外部宏观经济环境承压和我国利率市场化改革等金融改革的大背景下,对风险水平不断提高、经营发展水平良莠不齐且"太多而不能倒"和"太过关联而不能倒"的中小银行群体构建风险早期预警模型有利于银行风险的事前防范、事中控制和事后监督,也能够正确引导中小银行改革发展,不仅有利于监管机构实施外部监管,也为银行自律监管和风险控制提供参考,对完善现代金融监管体系,健全金融风险预防、预警、处置、问责的制度体系,防范化解系统性金融风险具有重要意义。

第二,考虑到中小银行区别于系统重要性银行的区域性经营特点,如果要进一步提升中小银行风险早期预警模型的预测精度,中央及地方金融监管机构应该建立全面、完善、动态、公开的行业数据库,运用大数据、云计算等技术挖掘中小银行经营中的关键指标和动态数据,不局限于现有银行监管条例,依据"适配性"监管原则构建与中小银行经营业务相匹配的监管指标。仿照对国有大型商业银行"一行一策"的监管标准,对中小银行采取差异化的适配性监管原则并设置不同监管标准,有利于降低监管成本和中小银行合规成本,激励中小银行探索健康发展和稳健经营模式,实现对风险状态更为精准的预警。

参考文献:

[1]周小川.公司治理与金融稳定[J].中国金融,2020(15):9-11.

[2]范小云.金融结构变革中的系统性风险分析[J].经济学动态,2002(12):21-25.

[3]徐超."太大而不能倒"理论:起源、发展及争论[J].国际金融研究,2013(8):89-96.

[4]Zhou,C..Are Banks Too Big to Fail? Measuring Systemic Importance of Financial Institutions.International Journal of Central Banking,2010,6(4):205-250.

[5]Rajan.Too Systemic to Fail:Consequences,Causes and Potential Remedies.Written Statement to the Senate Banking Committee Hearings,2009.

[6]Acharya,Viral V.,and Yorulmazer,T..Too many to fail--An analysis of time-inconsistency in bank closure policies[J].Journal of Financial Intermediation,2007,16(1):1-31.

[7]Brown,C O.,and Din? I.S..Too Many to Fail? Evidence of Regulatory Forbearance When the Banking Sector Is Weak[J].The Review of Financial Studies,2011,24(4):1378-1405.

[8]Gabrieli,S..Too-Connected Versus Too-Big-To-Fail:Banks' Network Centrality and Overnight Interest Rates.Banque de France Working Paper No.

398,2012.

[9]范小云,王道平,刘澜飚.规模,关联性与中国系统重要性银行的衡量[J].金融研究,2012(11):16-30.

[10]马君潞,范小云,曹元涛.中国银行间市场双边传染的风险估测及其系统性特征分析 [J]. 经济研究,2007(1):68-78+142.

[11]Altman, E. I.. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance,1968.23(4):589-609.

[12]Martin Daniel.Early warning of bank failure:A logit regression approach.North-Holland,1977,1(3).

[13]West,R.C..A Factor -analytic Approach to Bank Condition.Journal of Banking & Finance,1985,9(2):0-266.

[14]Lane William R.,Looney Stephen W.,and Wansley James W..An Application of the Cox Proportional Hazards Model to Bank Failure.Journal of Banking & Finance,1986,10(4):0-531.

[15]Chiaramonte L.,Croci,E.,and Poli,F..Should We Trust the Z-score? Evidence from the European Banking Industry.Global Finance Journal,2015,28:111-131.

[16]Rosa,P.S.,and Gartner I.R.,Ricardo,I..Financial Distress in Brazilian Banks:An Early Warning Model.Revista Contabilidade & Finan?as,2018,29:312-331.

[17]Ferriani,F.,Cornacchia,W.,Farroni,P.,Ferrara,E.,Guarino,F.,and Pisanti,F..An Early Warning System for Less Significant Italian Banks.Questioni Di Economia E Finanza.2019.

[18]Brauning,M.,Malikkidou,D.,Scricco,G.,and Scalone,S..A New Approach to Early Warning Systems for Small European Banks.ECB Working Paper Series No 2348,2019.

[19]Suss,J.,and Treitel,H..Predicting Bank Distress in the UK with Machine Learning.Bank of England Working Papers,2019.

[20]杨保安,季海.基于人工神经网络的商业银行贷款风险预警研究[J].系统工程理论与实践,2001(5):70-74.

[21]贺晓波,张宇红.商业银行风险预警系统的建立及其实证分析[J].金融论坛,2001(10):32-35.

[22]毛锦,周鹏,蔡淑琴.商业银行信用风险预警支持模型及其系统[J].金融理论与实践,2006(8):3-5.

[23]中国银监会银行风险早期预警综合系统课题组.单体银行风险预警体系的构建[J].金融研究,2009(3):39-53.

[24]罗晓光,刘飞虎.基于 Logistic 回归法的商业银行财务风险预警模型研究[J].金融发展研究,2011(11):55-59.

[25]陆静,王捷.基于贝叶斯网络的商业银行全面风险预警系统[J].系统工程理论与实践,2012(32):225-235.

[26]王伟.后金融危机时代商业银行危机预警系统构建与警情分析——以 A 股上市银行为例 [J].中国经济问题,2013(1):92-99.

[27]丁德臣.经济新常态下商业银行风险预警系统研究[J].宏观经济研究,2016(4):124-134.

[28]Baron,M.,Verner,E.,and Xiong,W.,.Banking Crises Without Panics.The Quarterly Journal of Economics,2020,136(1):51-113.

[29]James,Gareth,Witten,D.,Hastie,T.,and Tibshirani,R..An Introduction to Statistical Learning.2013.Vol.112.Springer.

[30]Hirakata,N.,Kido,Y.,and Jie,L.T..Empirical Evidence on 'Systemic as a Herd':The Case of Japanese Regional Banks.Bank of Japan Working Paper Series,2017.

责编:何青　　校对:魏鹏飞