嘉兴学院学报
Journal of Jiaxing University
第 31 卷第 4 期 2019 年 7 月
Vol.31 No.4 2019.7

DOI:10. 3969/j. issn. 1671-3079. 2019. 04. 010

基于机器学习方法的小微企业信用风险评估问题研究
——关于嘉兴市小微企业 2013-2017 年金融信用信息的实证分析

张榕薇
(中国人民银行 嘉兴市中心支行，浙江嘉兴 314050)

摘 要： 以嘉兴市 2013-2017 年发生信贷业务的小微企业为研究对象，对 2646 家小微企业 2013-2017 年的金融信用信息进行实证分析，构建了 5 类小微企业信用风险评估模型，结果显示，随机森林模型更适用于嘉兴市小微企业信用风险评估，根据该方法将小微企业风险分为“高档”“中档”“低档”3 类，提出了政策建议。

关键词： 机器学习；嘉兴；小微企业；信用风险；实证分析
中图分类号： F279.243 **文献标志码：** A **文章编号：** 1671-3079(2019)04-0064-08

On Credit Risk Assessment of Small and Micro Enterprises Based on the Method of Machine Learning
——An Empirical Study of the Financial Credit Information of the Small and Micro Enterprises in Jiaxing from 2013 to 2017

Zhang Rongwei
(Center Branch of Jiaxing, People's Bank of China, Jiaxing, Zhejiang 314050)

Abstract: Taking the small and micro enterprises with credit business in Jiaxing from 2013 to 2017 as the research objects, we carry out an empirical analysis of the financial credit information of 2646 small and micro enterprises among them. Five kinds of credit risk assessment models are established and the results show that the random forest model is more suitable for the assessment of the small and micro enterprises in Jiaxing. According to this method, small enterprises are classified into three categories, namely, "the high-grade", "the mid-grade" and "the low-grade", and policy suggestions are put forward.

Key words: machine learning; Jiaxing; small and micro enterprise; credit risk; empirical study

小微企业融资难、融资贵问题一直受到党中央、国务院的高度重视。2014 年 8 月，国务院办公厅发布了《关于金融支持小微企业发展的实施意见》，将大数据理念首次纳入国家解决小微企业融资难的政策中。2017 年 9 月，颁布新修订的《中小企业促进法》第七条规定，建立社会化的信用信息征集与评价体系；第二十三条规定，国家征信机构发展针对中小企业融资的征信产品和服务。如何通过企业金融信用信息合理有效评估小微企业的信用风险状况，对于解决中小企业融资难问题起着关键性的作用。

收稿日期：2019-06-11
作者简介：张榕薇（1989- ），女，浙江嘉兴人，中国人民银行嘉兴市中心支行经济师。
网络出版时间：2019-07-02 10:21:46 网络出版地址：http://kns.cnki.net/kcms/detail/33.1273.Z.20190702.0757.004.html

机器学习（machine learning）是指“计算机利用经验自动改善系统自身性能的行为”。[1] 小微企业金融信用信息数据样本数量大、特征种类多，基于机器学习方法的大数据分析方法因自我学习能力强，可以将数量庞大、种类多样的数据在短时间内进行有效采集和处理，适用于小微企业的信用风险评估。本文依托机器学习方法，建立实证研究模型，以提高数据分析的效率和准确性，帮助小微企业降低融资成本和风险。

国外学者普遍认为，融资难是制约小微企业发展的重要问题。Stiglitz 和 Weiss 发现信息不对称情况下，金融机构无法了解小微企业真实信息，无法准确评估其风险，因而制约了金融机构对小微企业的融资供给。Diamond 认为，在信用匹配模型中，信用的良好与否影响借款人获得贷款的成功率。Viktor Mayer-Schnberger 认为，随着大数据时代的来临，人们应转变数据处理上的思维模式，更加注重数据间的相关关系，而非因果关系。[2-4]

国内学者也对中小企业融资难进行了研究。林毅夫认为，解决中小企业融资难问题的有效方法是设立专门服务小微企业的融资机构，向非国有中小企业金融机构开放市场，进行市场化竞争，从根本上缓解小微企业融资难问题。巴曙松认为，资金供需双方之间信息不对称和风险管理上的激励不相容是造成小微企业融资困境的两大根源。而大数据时代的到来，增强了解决这两个根源的可能性。李先瑞以拍拍贷为例，认为大数据征信通过全面分析法，更加注重事物之间的相关性，能有效解决小微企业融资难的缺信息、缺信用问题。范晓忻、冯文芳认为，通过国家为小微企业融资创造良好的政策环境、消除信用信息孤岛、培育多元化的征信机构和加快征信制度建设等方式，可解决小微企业融资难问题，助力小微企业发展。[5-8]

在已有企业大数据研究中，大多应用大数据对企业进行定性分析；少有定量的分析，在定量分析中，又多以“上市公司”作为研究对象，数据样本量基本在 200 家左右，很少有对金融信用信息数据的挖掘分析。[9-10] 本文将构建有效的小微企业信用风险评估模型，对 2013-2017 年嘉兴市小微企业的金融信用信息进行实证分析。

一、嘉兴市小微企业的融资和信用现状

截至 2018 年末，嘉兴市小微企业数量为 14.57 万家，2017 年全市新增小微企业 2.49 万家，小微企业贷款余额 2 004.63 亿元，同比增长 7.56 %，占全部贷款余额 51.66 %。2013-2017 年，嘉兴市辖内小微企业数量稳步增长，见图 1。

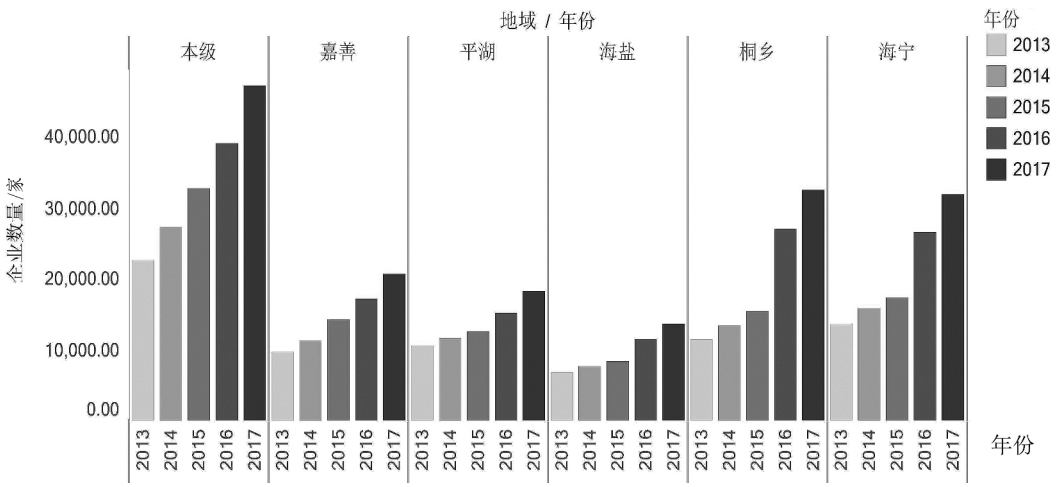

图 1 2013-2017 年嘉兴市小微企业数量统计

数据来源：嘉兴市统计局

选取嘉兴范围内2013-2017年贷款发生额100万元以上的小微企业为研究对象，涉及16 105家企业。

（一）小微企业获得融资仍主要依靠抵押、保证

从表1可以看出，小微企业信贷业务类型中，抵押方式占比为48.2%，保证方式占比为31.3%，小微企业融资仍高度依赖抵押、保证等方式，对于部分轻资产的小微企业而言，融资问题比较突出。

表1　2013-2017年嘉兴市小微企业各信贷业务类型累计发生额统计　%

类型	抵押	保证	质押	贷款	承兑汇票	票据贴现	信用证	贸易融资	公开授信
占比	47.8	33.1	12.0	4.9	1.3	0.4	0.2	0.1	0.1

数据来源：企业金融信用信息统计

（二）小微企业不良贷款分布行业较为集中

小微企业不良贷款主要集中在住宿和餐饮业、制造业及批发和零售业领域，如图2所示，且住宿和餐饮业远高于后两大行业。具体来看，制造业虽然是融资主力军，但产生不良贷款的家数不多，不良贷款主要集中在一些自身并无竞争实力的企业，这与不同行业面临的风险和生长周期密切相关。

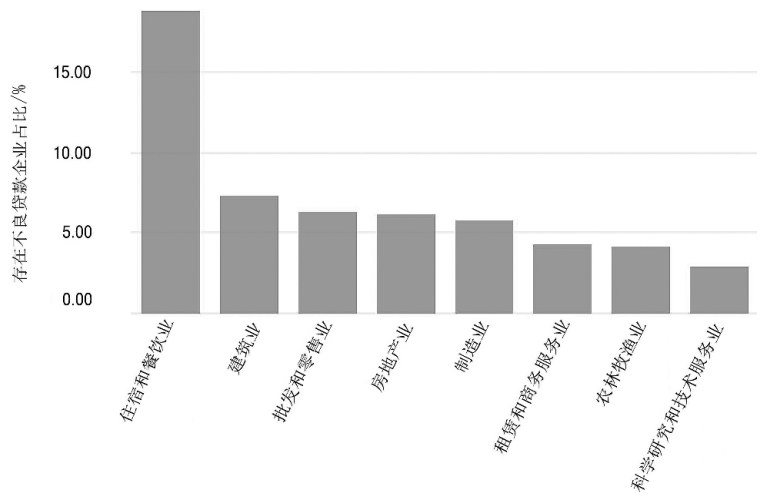

图2　2013-2017年嘉兴市小微企业各行业不良贷款累计发生额占比统计

数据来源：企业金融信用信息统计

分析不良贷款中次级、可疑和损失的占比中，损失类贷款排名前三的行业为制造业、批发和零售业及农林牧渔业，如表2所示；分析不同行业逾期时间长度中，排名前三的行业为住宿和餐饮业、农林牧渔业以及批发和零售业，如表3所示。

表2　2013-2017年嘉兴市小微企业各行业不良贷款累计发生额分类占比表　%

	制造业	批发和零售业	农林牧渔业	建筑业	房地产业	租赁和商务服务业	住宿和餐饮业
次级	55.94	8.96	1.01	10.32	7.19	5.49	10.99
可疑	59.70	14.25	0.44	14.49	9.01	0.91	1.21
损失	87.10	12.13	0.75	0.02	0.00	0.00	0.00

数据来源：企业金融信用信息统计

表3 2013-2017年各行业小微企业贷款逾期累计发生额情况统计表 %

行业名称	长期逾期金额率	中期逾期金额率	短期逾期金额率	合计
住宿和餐饮业	3.047	1.769	0.105	4.920
农林牧渔业	1.739	0.683	0.222	2.645
批发和零售业	0.346	0.299	0.191	0.836
建筑业	0.396	0.249	0.248	0.894
制造业	0.407	0.218	0.168	0.792
房地产业	0.713	0.146	0.183	1.042
租赁和商务服务业	0.180	0.126	0.155	0.461
科学研究和技术服务业	0.247	0.105	0.007	0.358

数据来源：企业金融信用信息统计

二、小微企业信用风险评估模型的构建及分析

小微企业信用风险评估有主观和客观两类建模方式。主观方法为专家评分法，即根据评价对象的具体要求选定若干评价项目，由业务代表性专家凭借自身经验根据评价项目制订出评价标准，形成评价指标体系，然后对全量样本进行评判。它一般适用于个体信息量化程度不高和特征量较少的情况。客观方法为机器学习法，即运用机器学习技术在目标领域大规模历史数据基础上，推演出评判模型。这类方法的好处在于可以从数据出发，能够客观公正地进行评判，且机器学习模型能不断纳入新信息，对历史数据不断自学习，能够动态地更新调整内部权重。

本文使用（tableau数据分析软件、sklearn数据建模工具），建模目标是分析小微企业各维度信用信息，对小微企业信用风险进行预判。

（一）样本选取

在企业样本选取过程中，样本信息完整性、个体代表性和整体丰富性都是重要考量因素。为了避免信息项缺失对分析造成误差，先对嘉兴市现有16 105家小微企业的信息做筛选过滤，剔除数据不完整企业，选取余下2 646家企业作为建模样本对象，这些企业样本规模相当、行业覆盖面较全，具有一定的代表性。

根据各信贷业务种类五级分类的结果，将五级分类中标记为“可疑”“次级”“损失”的企业，划定为“存在信贷违约风险”，作为正样本；把标记为“正常”“关注”的企业，划定为“不存在信贷违约风险”，作为负样本。统计得到样本集合存在信贷违约风险比例为18.7 %，正负样本规模相当。

（二）指标体系

通过数据积累并结合已有研究，根据企业中征码、企业名称等标识字段，将企业基本信息、行业信息、财务信息、纳税信息、水电信息、信贷信息、融资信息、对外担保及被担保信息等各维度数据关联整合，形成小微企业信息项汇总表。经过数据清洗、特征数值平滑处理、标准化处理、独热编码处理和构建组合特征等步骤特征工程处理后，形成最终特征矩阵。

（三）模型选择

小微企业征信数据具有样本数量大、特征种类繁多、样本特征缺省情况较多的特点，为了避免模型过度拟合，尝试使用线性支持向量机、非线性支持向量机、逻辑回归、随机森林和梯度提升树五类主流分类模型构建小微企业信用风险评估模型。

1. 线性及非线性支持向量机模型

令超平面上的分隔函数为：

$$f(x) = \omega^T x + b$$

目标值用+1和-1分别表示正负样本的标签值：

$$y=\begin{cases}+1，f(x)>0，\text{高信贷违约风险企业}\\-1，f(x)<0，\text{低信贷违约风险企业}\end{cases}$$

引入松弛因子ξ，作为损失值相加，使得损失最小化，令常量C等于0.5，表示松弛因子对全局优化过程中的惩罚程度。得到优化问题：

$$\min_{\omega，b，\xi_i}\frac{1}{2}\omega^T\omega+C\sum_{i=1}^{n}\xi_i^2$$

$$s.\ t，y_i=\omega^Tx+b+\xi_i，i=1，2，\cdots，n$$

2. 逻辑回归模型

将函数$h_\theta(x)$用于拟合企业产生不良贷款的概率，具体表示如下：

$$h_\theta(x)=\frac{1}{1+e^{-\theta^Tx}}$$

这里假设给定的阈值是0.5，目标值用1和0分别表示正负样本的标签值：

$$y=\begin{cases}1，h(x)>0.5，\text{低信贷违约风险企业}\\0，h(x)<0.5，\text{高信贷违约风险企业}\end{cases}$$

用对数似然损失函数作为逻辑回归的损失函数，使损失函数最小化：

$$\min_\theta J(\theta)$$

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{m}Cost(h_\theta(x^{(i)})，y^{(i)})$$

其中

$$Cost(h_\theta(x)，y)=\begin{cases}-\log(h_\theta(x))>0.5，\text{if } y=1\\-\log(1-h_\theta(x))<0.5，\text{if } y=0\end{cases}$$

3. 随机森林及梯度提升树

这两类是基于决策树优化的模型，为了防止过度拟合，每棵树的深度尽量小，本文设置最大深度为4。为了使模型更精准，创建决策树的数量要达到一定的量级，本文设置数量为1 000。

（四）检验方式

通过交叉验证法验证现有模型。即将样本分成N份，每一份都可以作为测试集，而剩余的N-1份作为训练集，以此减少突发性和偶然性。考虑到训练样本量的丰富性及测试样本量具备的统计意义，将N设置为4进行验证。

（五）实证分析

为了同时验证机器学习算法在信用评估上的性能和有效性，分别测试线性支持向量机及非线性支持向量机、逻辑回归、随机森林和梯度提升树算法，计算在不同模型下数据集的表现，并通过交叉验证法多维度验证模型性能，并对其进行比较。

1. 模型性能分析

在传统分类方法中常用准确度（Accuracy）这一评价指标。为了更准确地区分各模型优劣，在此基础上引入混淆矩阵（confusion matrix），矩阵的含义如表4所示。

表4 混淆矩阵（confusion matrix）

项目	预测为正类	预测为负类
实际为正类	真正（TP）	假负（FN）
实际为负类	假正（FP）	真负（TN）

误差（error，ERR）可以理解为预测错误样本的数量与所有被预测样本数量的比值，而正确率（accuracy，ACC）的计算方法则指正确预测样本的数量与所有被预测样本数量的比值：

$$ERR = \frac{FP + FN}{FP + FN + TP + TN}$$

预测准确率也可通过误差直接计算：

$$ACC = \frac{TP + TN}{FP + FN + TP + TN} = 1 - ERR$$

受试者工作特征曲线（receiver operator characteristic，ROC）是基于模型真正率（TPR）和假正率（FPR）性能指标进行分类模型选择的有效工具，假正率和真正率可以通过移动分类器的分类阈值来计算。如对于完美的分类器，其真正率为1，假正率为0，这时ROC曲线即为横轴0与纵轴1组成的折线。

$$TPR = \frac{TP}{P} = \frac{TP}{FN + TP}$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

对5类分类模型产生的ROC曲线进行比较。通过ROC曲线的线下区域（area under the curve，AUC）评判分类模型的整体性能。

准确率（precision，PRE）和召回率（recall，REC）是与真正率和真负率相关的性能评价指标：

$$PRE = \frac{TP}{TP + FP}$$

$$REC = TPR = \frac{TP}{P} = \frac{TP}{FN + TP}$$

在实践中，使用准确率和召回率的组合F1分数衡量模型过滤负样本的性能。

$$F1 = 2 \times \frac{PRE \times REC}{PRE + REC}$$

精度-召回率曲线（precision recall curve，PRC）描述了分类阈值计算得到的准确率和召回率的变化情况，随着正样本召回率的增加，准确率减少的速度越慢，反映出模型过滤负样本的性能越高。

通过ROC曲线和PRC曲线的对比，可以发现在本文选取的训练样本中，随机森林的ROC曲线更接近边界值，PRC曲线有更高的准确率，曲线平滑程度高，具有较强的抗过拟合能力，相比于另外4类模型更适用于评估小微企业的信用情况。

表5为各模型算法在各衡量指标上的结果对比情况。

表5 模型各维度参数对比表

模型	ACC	REC	AUC	F1-score
逻辑回归	0.886	0.942	0.850	0.933
随机森林	0.917	0.995	0.921	0.953
梯度提升树	0.916	0.960	0.903	0.951
线性支持向量机	0.889	0.947	0.866	0.935
非线性支持向量机	0.892	0.989	0.848	0.939

2. 预测结果分析

随机选取一批嘉兴市小微企业样本，使样本集包含正负样本的数量相当，进一步分析小微企业信用风险分布情况。将基于随机森林模型预测得到的发生不良贷款的概率probability（x）保留2位小数倒置，再进行等比例放大，作为小微企业X信用评分Credit_ Index（x），企业信用评分越高，表示其信贷违约风险越低，即

$$Credit\_\ Index(x) = \text{probability}(x)100.\ 0$$

$$Credit\_\ Index(x) \in [0, 100]$$

计算样本集合中每家小微企业的信用评分，可绘制小微企业信用评分直方图。由于模型将概率0.5作为划分正负样本的阈值，在直方图中沿用该阈值将Credit_ Index（x）≥50认定为高信用评分小微企业，将Credit_ Index（x）<50认定为低信用评分小微企业，如图3所示。

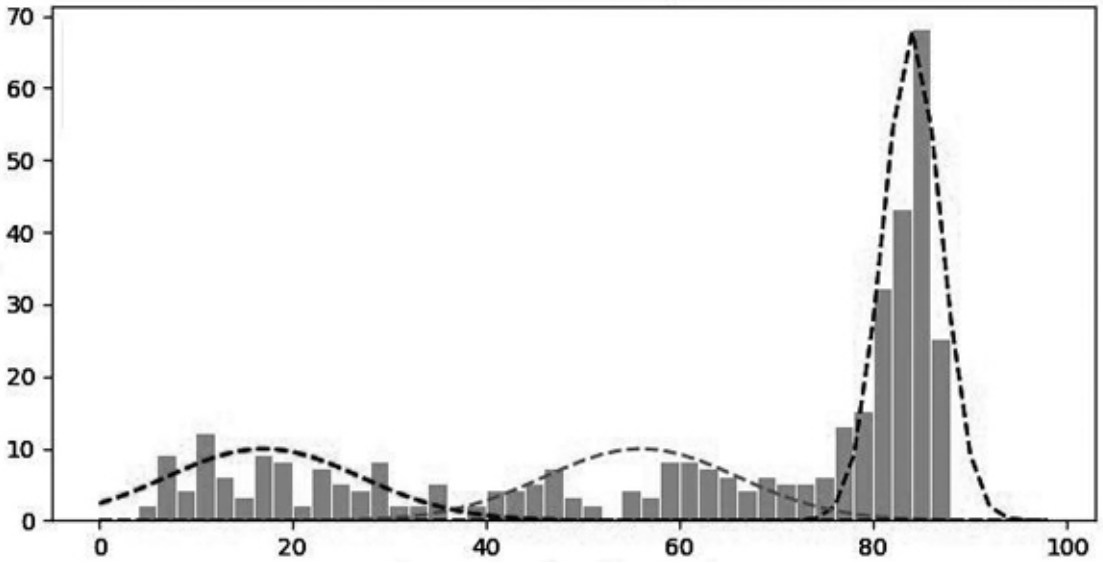

图3 小微企业信用评分直方图（高斯分布拟合结果）

从图3可以发现，小微企业信用评分的分布呈现“三段式分布”，分别在信用评分的高、中、低三个区域相对集中。假设小微企业的信用评分分布情况满足独立同分布（IID）特性并服从混合高斯分布（Mixture Gaussian Distribution），通过混合高斯模型（Gaussian Mixed Model）拟合其分布情况，样本在3个区域呈现高斯分布情况。

通过混合高斯模型拟合的结果不难发现，混合高斯模型能较好地拟合小微企业信用评分的分布情况，可以将小微企业划分为高、中、低三档，使其在各个区域分布的方差和最小化。

各分布区域的波峰位置，即各区域小微企业信用评分的数学期望。分布在高档信用评分区域的小微企业，均被随机森林模型标记为“低信贷违约风险”，在总小微企业数量中的占比最高；分布在低档信用评分区域的小微企业，均被随机森林模型标记为“高信贷违约风险”，占总小微企业数量的一小部分；而在中档信用评分区域内，两类标记同时存在，属于模型划分正负样本的临界点，对应实际情况，这部分小微企业中有一定比例企业已发生了不良贷款，剩余企业虽未出现不良贷款，却存在发生不良贷款的隐患。本文通过以上模型，测算得出了嘉兴市小微企业信用评分，并据此划分出了“高档”“中档”和“低档”信用风险的小微企业。

三、结论

基于嘉兴市2 646家小微企业微观数据，构建包含行业、财务、纳税、水电、信贷、担保、融资各维度的小微企业信用指标体系。分别运用5类机器学习方法拟合小微企业信贷违约风险概率，通过对比发现，随机森林模型的总体准确性与稳定性高于其余4类模型，对嘉兴市小微企业有更优的检测评估效果，且有明确的计算过程和可解释的决策规则，适用于对嘉兴市小微企业信贷风险的定量评估。

小微企业和商业银行之间的信息不对称，使得大部分的小微企业在融资中需要采取抵押或保证方式，因此，运用混合高斯模型寻求小微企业信用风险分布的规律，从定性结果反映定量原因，透过小微企业发生不良贷款的现象，可发现小微企业存在信贷风险的本质，避免了“一刀切”的评判方式，能够更加精准地对小微企业的信用风险进行评估。银行在充分获得小微企业金融信用信息数据的前提下，可以采用随机森林模型进行数据建模分析，得出每家小微企业的信用“分档”，从而有针对性地

采取“分而治之”的融资策略。为信用信息不对称、信用信息获取成本高问题提供切实的解决途径，并且能够解决金融机构的信贷政策配给问题。同时，机器学习方法提供较先进的技术支持，加快小微企业融资的进程。

小微企业仅仅是嘉兴市企业的一部分，信用风险评估的测算模型还可推广到全市其他企业。而信用风险评估的推广，有利于全社会营造“守信者受益、失信者惩戒”的氛围，引导企业加强信用建设和信用管理，为推动嘉兴市社会信用体系建设、构筑良好的社会信用环境、促进经济社会发展做出贡献。

四、政策建议

1）借助机器学习技术，改进传统小微企业信用风险评估方法，将评价方法引入企业征信系统，得出小微企业在行业内的相对位置。低档信用小微企业，可采取适当措施清退；中档信用小微企业，由于违约和正常情况均存在，存在潜在发展空间，应重点关注和扶持；高档信用小微企业，可给予完善保障措施，加强事后监管。

2）加强机器学习技术学习，改进传统信用风险评估方法。建议商业银行加大技术投入，特别是地方性金融机构，要深入学习大数据、云计算等先进技术，在有效整合小微企业信用信息数据的基础上，实现数据更大规模、更加自主地搜集和处理，在数据存储、使用等方面进一步创新。

3）加强顶层设计，完善小微企业的信用信息采集。当前，小微企业的信用信息分散在多个部门，信息资源共享机制不够成熟，小微企业联动管理的难度较大。金融机构由于缺少对小微企业信息全面的掌握，在政策上很难形成对小微企业的有效针对性措施，加之小微企业缺少大企业的自身优势，使小微企业处于金融机构的政策青睐之外。因此，建立小微企业信用信息共享机制，推动小微企业信用信息在金融机构之间的共享，对推动小微企业的融资和发展非常有益。

参考文献：

[1] TOM MITCHELL. 机器学习 [M]. 北京：机械工业出版社，2008：11-50.
[2] STIGLITZ JE，WEISS A Credit Rationing in Markets with Imperfect Information [J]. The American Economist，1981（3）：393-410.
[3] DIAMOND. Financial Intermediation and Delegated Monitoring [J]. Review of Economic Studies，1984（51）：393-414.
[4] 维克托·迈尔-舍恩伯格，肯尼斯·库克耶. 大数据时代——生活、工作与思维的大变革 [M]. 杭州：浙江人民出版社，2013：20-56.
[5] 巴曙松. 大数据可解小微企业融资瓶颈 [J]. 中国经济报告，2013（6）：29-31.
[6] 李先瑞. 大数据征信破解小微企业融资困境探讨——以拍拍贷为例 [J]. 会计之友，2015（13）：52-55.
[7] 冯文芳. 互联网金融背景下小微企业大数据征信体系建设探析 [J]. 国际金融，2016（3）：74-78.
[8] 范晓忻. 大数据征信与小微企业融资 [J]. 观察思考，2014（22）：81-82.
[9] 肖斌卿，柏巍，姚瑶，等. 基于 LS_ SVM 的小微企业信用评估研究 [J]. 审计与经济研究，2016（6）：102-111.
[10] 王庆，姚康. 机器学习方法在中小企业信用评估中的应用研究 [J]. 特区经济，2018（348）：145-147.

（责任编辑　王连桥）