LGB-BAG 在 P2P 网贷借款者信用风险评估中的应用

李淑锦，嵇晓佳
（杭州电子科技大学 经济学院，杭州 310018）

摘 要：立足于 P2P 平台，利用 P2P 平台个人借款人的信息建立了一套系统的信用风险评估指标体系来甄别可能违约的借款人。基于 LightGBM（一种基于决策树的 Boosting 模型）和 Bagging 提出一种新的 LGB-BAG 模型，有效结合了 Boosting 和 Bagging 的优势。结果表明，在决策树的个数增大到一定程度的时候，LGB-BAG 的 F_1 均值（预测效果）要高于 LightGBM 和随机森林；并且 LGB-BAG 的 F_1 方差也要小于其余两种模型。LGB-BAG 的 F_1 均值最高可达到 0.71175，且 LGB-BAG 模型能够显著提高信用风险预测效率。

关键词：集成学习；LightGBM；Bagging；Boosting；P2P 平台
中图分类号：F830.589；F832.479 **文献标志码**：A **文章编号**：1002－980X(2019)11－0117－08

近年来，互联网的发展极大地改变了人民的生活方式，也改变了金融的实现方式，产生出大量的物联网金融产品。这些金融产品受到了广泛的关注，在生活中影响着人民的借贷款模式以及传统的银行业发挥各自的融通资金的作用。自 2007 年第一家 P2P 贷款平台拍拍贷在中国产生后，陆续出现了许多 P2P 平台如人人贷、宜人贷等，在中国更是得到十分的关注。P2P 网贷平台主要搭建中介平台，收取手续费，投资人获得利息，借款人支付利息。这记分满足了中小投资人和借款人的投资和借款的需求。这种互联网金融的借贷模式操作简单、门槛低，解决了很多低信用个人、中小型企业等借款难的问题，是一种极具创新性的借贷方式。

由于中国市场缺乏监管、法律不够完善，以及借款人信用较低等原因，借款人不还款的现象越来越严重，严重影响了投资者利益，使得 P2P 网贷方式的逆向选择和道德风险越加严重，阻碍了这种融资方式继续发展的脚步。因此，对于 P2P 平台上借款人的信息进行分析，排除高风险人群，降低平台的损失，促进其发展，就显得尤为重要。

已有的关于 P2P 网贷平台信用风险的研究十分丰富。部分学者研究了个人借款者成功获得借款的影响因素，如 Pope 和 Sydnor[1] 研究 prosper 平台的数据，发现年纪大的要比年纪小的借款者更可能借到钱，而且网贷平台在借款中还存在种族歧视问题；Ravina[2] 也通过对 prosper 平台数据的实证研究，发现外貌对借款利率有明显影响，外貌较差的借款者会有更高的借款利率；廖理等[3] 利用人人贷的数据进行实证研究，发现我国 P2P 借贷行业中存在地域歧视问题；李悦雷等[4] 通过研究拍拍贷的数据发现国内 P2P 网贷投资者有明显羊群行为。

信用风险评估在金融借贷领域中是非常重要的一环。例如，赖辉等[5] 就曾因为个人信贷客户的违约事件屡见不鲜而提出了 PCBS 方法，且在此基础上建立一种个人信贷客户动态信用评估方法，以此来降低违约造成的风险损失。在 P2P 借贷的个人信用风险评估中，传统的 Logistic 回归模型，因为其简单、快速和足够高的预测的准确度而备受学者和从业人员的青睐。谈超等[6] 使用 Logistic 回归模型对 P2P 平台的羊群行为进行研究，此外还有王修华等[7]、滕晓慧[8]、王文怡和程平[9] 利用 Logistic 回归模型对 P2P 平台的研究。随着人工智能（AI）的发展，出现了如 SVM、ANN 等经典的单分类器，因此国内外学者也用了这些方法来提高个人信用风险评估的预测精度。现有的方法有梯度提升决策树（谭中明等[10]）、决策树（Serrano-Cinca 和 Gutiérrez-Nieto[11]）、SVM（师应来等[12]）、ANN（Korol[13]）等。

集成学习方法由于可以集成基分类器并提高分类效果而已经成为目前研究的一个热点。Nanni 和

收稿日期：2019－10－23
基金项目：国家社会科学基金项目“基于大数据的金融零售信用风险评估与智能决策研究”（17BJY233）；教育部人文社科青年项目“网络借贷信用风险评估的结构化方法及应用研究”（16YJCZH031）
作者简介：李淑锦（1967—），女，山西原平人，金融数学博士，杭州电子科技大学经济学院教授，研究方向：金融工程、风险评估；嵇晓佳（1995—），女，浙江湖州人，杭州电子科技大学硕士研究生，研究方向：信用风险评估。

Lumini[14]利用日本、澳大利亚和德国的信用数据，运用 Random Subspace、Bagging、Class Switching 和 Rotation Forest 对银行的信用问题进行了研究；Abellán 和 Castellano[15]利用 6 个国家的实际信用数据，使用 5 种集成学习方法构建模型，结果表明与单个模型相比，集成学习模型具有更强的预警能力和稳健性。此外还有 Florez-Lopez 和 Ramon-Jeronimo[16]、Tsai 等[17]关于这些模型的研究。而操玮等[18]则对这些不同的集成学习方法进行了对比。

通过梳理国内外相关文献可知，集成学习方法要优于单个学习器，但是大多数学者只对基学习器使用单一的集成学习方法，却没有将多种集成学习方法结合起来使用。在这些集成学习方法中，Bagging 可以减少模型的方差，却很难降低模型的偏差；而 Boosting 则能够有效降低模型的偏差。因此本文提出一种新的基于 LightGBM 和 Bagging 的模型，称为 LGB-BAG，有效结合了 Boosting 和 Bagging 两种集成学习策略的优势，提高了模型的分类效果。

1 模型介绍

首先介绍集成学习方法——Bagging 和 Boosting，以及作为基学习器的 LightGBM，最后基于 LightGBM 和 Bagging 的融合创建新的模型——LGB-BAG 模型。

1.1 Bagging 和 Boosting 集成学习法

Bagging[19]是一种把多个不同的基分类器集成为一个集成分类器的方法。Bagging 基于自主采样法(bootstrap sampling)，重复地取样而取得不同的数据集，并在不同的数据集上训练得到有高泛化能力和较大差异度的基分类器。从一群基分类器得到的预测集合体在预测一个类标时，采用投票的方法，将票数最多的类标确定为该样本的预测类标。此算法也是一种并行的集成学习方法，从而提高算法的时间效率。

Boosting[20]则是一种串行的集成学习方法，其函数模型是叠加型的。后一个基学习器会不断修正或提高前一个基学习器的结果，最终将各个集学习器叠加而成。其中 Gradient Boosting 是 Boosting 当中的一个重要方法，它会在迭代的时候选择梯度下降的方向来保证最后的结果最好。基于此方法有很多著名的算法如 GBDT、XGBoost、LightGBM。

1.2 LightGBM

LightGBM[21]（Light Gradient Boosting Machine)是微软亚洲研究所 DMTK 团队的一个开源的算法，是一种基于决策树和 Gradient Boosting 的改进模型，可以用于常见的分类、回归等问题。LightGBM 和 XGBoost 算法被分别称为机器学习中的“倚天剑”和“屠龙刀”，都是非常优秀的算法。LightGBM 有着很多的优点：使用基于直方图的算法，有着更快的训练速度和更高的效率；更少的内存占用；支持并行计算，并且由于在训练时间上的缩减而拥有处理大数据的能力。

LightGBM 有两个重要的创新点是使用直方图算法和带深度限制的 Leaf-wise 的叶子生长策略：

(1)直方图算法。LightGBM 的一大创新点是基于直方图算法提出的。在计算的时候，模型会将浮点型的数值转化成离散数值，从而生成了一个直方图。将离散数值作为索引在图中累计统计量，这样的结果就是能极大地降低内存占用来进行遍历找出最佳分割点。

(2)带深度限制的 Leaf-wise 的叶子生长策略。大部分的决策树使用 Level-wise 策略。但是 Level-wise 是一种低效算法，因为它不加区分的对待同一层的叶子，带来了很多没必要的开销，如图 1 所示。

相较于 Level-wise 策略，Leaf-wise 则更为高效，它有着这样的一个循环：每次从当前所有叶子中，找到分裂增益最大的一个叶子，再进行分裂。所以在分裂次数相同的情况下，Leaf-wise 误差更低、精度更高。Leaf-wise 如图 2 所示。

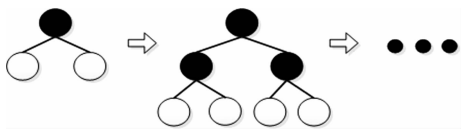

图 1 Level-wise 策略

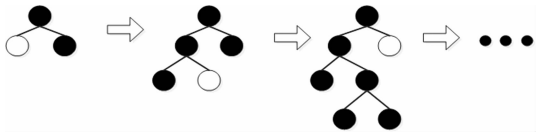

图 2 Leaf-wise 策略

带深度限制的 Leaf-wise 的叶子生长策略的缺点是：当样本量较小的时候，Leaf-wise 可能会长出比较深的树而导致了过拟合。因此 LightGBM 增加了一个最大深度限制来防止过拟合。

1.3 LGB-BAG

在基分类器是相互独立的前提下，由 Hoeffding 不等式[22]推导可知，随着集成当中学习器数目的增加，集成的错误率会呈指数级下降，最后趋向于零。Bagging 通过重新选取训练集从而增加分类器集成的差异度，并提高泛化能力，减少过拟合的

风险。

本文以 LightGBM(基于决策树和 Boosting)为基分类器,以 Bagging 为集成学习方法构建 LGB-BAG 模型。LGB-BAG 是一种基于决策树、结合了 Boosting 和 Bagging 的算法。Bagging 可以减少模型的方差,Boosting 则能够有效降低模型的偏差。因此理论上讲,LGB-BAG 既可以减少模型的方差,又可以有效降低模型的偏差。

LGB-BAG 算法具体如下:重复 T 次,每次从大小为 m 的训练样本集中随机放回地抽取 m 个样本,用基分类器 LightGBM 进行训练。利用相同的方法得到 T 个基分类器,并得到一个分类函数序列 $h_1(x),h_2(x),\cdots,h_T(x)$,最后的分类函数 $H(x)$ 采用投票的方法,将票数最多的类标确定为该样本的预测类标。对未知样本 x 进行分类时,每个基分类器得到一个分类结果,T 个基分类器进行投票,将票数最多的类标确定为样本 x 的预测类标。LGB-BAG 的算法流程如图 3 所示。

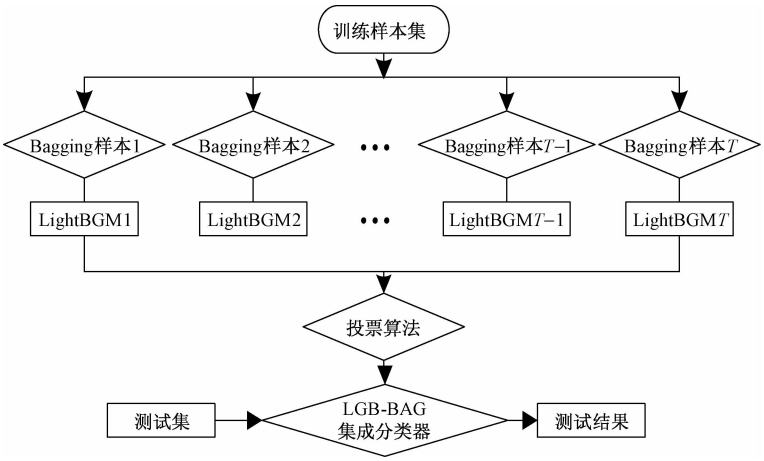

图 3　LGB-BAG 的算法流程

LGB-BAG 算法的具体描述如下。

输入:训练集 $D(x)=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$;

算法 Bagging 和 LightGBM;

训练轮数 T。

过程:

1. For t=1,2,…,T;

2. 对训练集 D 使用自主采样法,有放回地随机抽取 m 个样本,创建样本集 D_t;

3. 使用样本集 D_t 和算法 LightGBM 训练,得到基分类器 $h_t(x)$。

输出:$H(x)=\operatorname{argmax}\sum_{t=1}^{T}h_t(x)$

使用集成分类器 $H(x)$ 对未知样本 x 进行分类:

1. 未知样本 x 进行分类时,每个分类器 $h_t(x)$ 得到一个分类结果,T 个分类器进行投票,将票数最多的类标确定为样本 x 的预测类标;

2. 采用投票算法输出分类结果 $H(x)=\operatorname{argmax}\sum_{t=1}^{T}h_t(x)$。

2　信用风险评估流程

运用 LGB-BAG 模型对 P2P 平台的个人借款者进行信用风险评估是一个系统性的决策问题,为了达到较好的评估效果,需要对评价指标进行科学合理的处理以及赋值。因此,对个人借款者信用风险评估的研究流程为:评估指标选取—数据预处理—数据分析—模型设计—衡量指标选择。

2.1　评估指标选取

在这个大数据信息化的时代,每一个信息都传递着重要的信用价值,因此本文基于人人贷网站可获得的个人借款者的信息,选取了评估个人借款者信用风险的指标体系。具体的评估指标以及赋值见表 1。

2.2　数据预处理

(1)样本的采样。人人贷是国内的领袖级网贷平台,本文使用 Python 网络爬虫程序爬取人人贷个人借款者的实际交易数据。经处理后的数据集当中,共有 1048575 个样本,其中未违约人数为 1035133,违约人数为 13442,二者的比例是 77∶1。由于原有的数据量过于庞大,因此本文使用随机选取的方法,以 0.05 的比例,随机选出 52429 个样本,其中未违约人数为 51769,违约人数为 660,二者的比例是 78∶1。具体的样本分布情况见表 2。

(2)数值型指标的标准化。将逾期的客户标记为 1,未逾期的客户标记为 0。所有数值型的数据标准化,使其服从标准正态分布,以防止不同数据的数值大小对模型的影响。

表 1　信用风险评估的指标体系及其赋值

类别	指标	变量缩写	相关赋值说明
基本信息	性别	*gender*	男:0;女:1
	年龄	*age*	实际值
	婚姻状况	*marriage*	其他:0;已婚:1
工作状况	公司规模	*officeScale*	≤10:0;11～100:1;101～500:2;>500:3
	工作年限	*workYears*	1 年以下:0;1～3 年:1;3～5 年:2;5 年以上:3
	收入	*salary*	≤1000:0; 1001～2000:1; 2001～5000:2; 5001～10000:3; 10001～20000:4;20001～50000:5;≥50000:6
房车信息	房产	*hasHouse*	无房:0;有房:1
	车产	*hasCar*	无车:0;有车:1
	房贷	*houseLoan*	有房贷:0;无房贷:1
	车贷	*carLoan*	有车贷:0;无车贷:1
信用认证	信用评级	*creditLevel*	HR:0;E:1;D:2;C:3;B:4;A:5;AA:6
	信用分数	*sumCreditPoint*	实际值
借款信息	借款量	*borrowAmount*	实际值
	借款次数	*totalCount*	实际值
	成功次数	*successCount*	实际值
	利率	*interest*	实际值
	剩余月数	*leftMonth*	实际值
	借款月数	*month*	实际值

表 2　随机选取前后的样本分布情况

	正常还款记录	违约记录	总记录
处理前	1035133	13442	1048575
处理后	51769	660	52429

(3)指标缺失数据的填充。本文共有 4 个指标有缺失值:*officeScale*、*workYears*、*salary*、*marriage*,其中 *marriage* 的缺失率最少,为 3.17%;*officeScale* 的缺失率最大,为 17.59%。4 个指标的缺失率都较高,且数据量很大,如果删除指标会造成有效信息的缺失。由于这 4 个变量全为非数值型变量,本文采用 CART 算法来填充缺失值。文章一共有 18 个指标,使用剩余 14 个指标的数据,对 4 个缺失数据的指标(分别作为因变量)的缺失值进行预测并填充。

2.3　数据分析

2.3.1　相关性分析

对数据集当中的各个特征进行相关性检验,结果显示,变量的特征之间的相关性并不高,无需使用主成分分析法等方法对其进行降维,因此可以直接输入模型进行训练。

2.3.2　部分指标的分析

分析年龄(*age*),如图 4,发现借款人无论违约或不违约,多集中于 30 岁左右,而且明显呈现右偏态的情况;对性别(*gender*)分析的结果如图 5,发现男性违约率明显要高于女性很多,也符合大众的认知。

表 3 对信用风险评估的一些主要指标进行了统计分析。

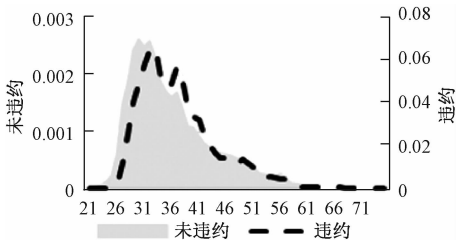

图 4　借款人的年龄分布

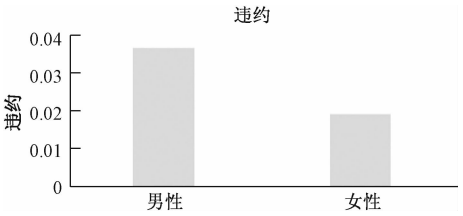

图 5　性别与违约分布

2.4　模型设计

本实验基于机器学习方法,通过个人借款者的信息来检验集成学习算法的预测性能以及基分类器对集成预警性能的影响。LightGBM 是一种基于决策树和 Boosting 的模型;随机森林(random forest,RF)是一种基于决策树和 Bagging 的模型;选择集成学习算法 Bagging、基分类器 LightGBM,构造了 LGB-BAG 模型,说明 LGB-BAG 模型是一种基于决策树、结合 Boosting 和 Bagging 两种集成学习方法的模型。

本文设置对模型的结果有重大影响的几个超参数。3 个模型都是基于决策树的模型,3 个模型当中

表 3　主要指标的统计信息

变量	均值	标准差	最小值	中位数	最大值
borrowAmount	39741.736	65018.794	0.000	8000.000	9000000.000
sumCreditPoint	92.982	83.637	0.000	67.000	3220.000
interest	12.053	2.592	3.000	12.000	24.400
leftMonths	14.420	11.653	0.000	12.000	48.000
months	22.981	11.774	1.000	24.000	48.000

决策树的个数 N 是非常重要的一个超参数;此外,LGB-BAG 中基学习器 LightGBM 的个数 M 也是非常重要的超参数。在本次实验当中,对比不同的 N,分别设置为 10、50、100、200、500、800、1000、1200、1400、1600、1800、2000、2200。而 LGB-BAG 的 M 依次设置为 2、5、10。为了避免随机性对结果的影响,使用了 5 折交叉检验,并取 F_1[精确率(*precision*)和召回率(*recall*)的调和平均值]的平均值。

文章的实验通过 Python 当中的 sklearn 和 lightgbm 包实现。

2.5　判别评估效果的衡量指标选择[22]

对于二分类问题,可以将样本根据真实的类别和分类器预测类别组合划分。定义 TP 为真正例(true positive),FP 为假正例(false positive),TN 为真反例(true negative),FN 为假反例(false negative)。由此得到二分类问题的混淆矩阵,见表 4。

表 4　二分类问题的混淆矩阵

	预测正类	预测负类
实际正类	*TP*	*FN*
实际负类	*FP*	*TN*

通常预测分类器准确程度的指标有准确率、精准率、召回率等。在普通的二分类问题中,准确率是衡量分类器性能最常见的指标,但是这种指标只是一种简单的衡量。正如上文所提到的,在不平衡问题当中,即使将所有样本分成负类,准确率依然很高的问题。在不平衡分类问题的判别中,通常不能使用准确率。准确率的定义为公式:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN}。\tag{1}$$

精确率表示在所有预测为正类的样本中,实际正类样本所占的比例,定义见公式(2);召回率表示在所有实际为正类的样本中,预测正类样本所占的比例,定义如公式(3)。可以看出精确率和召回率是一对矛盾的度量。一般地说,精确率高的时候,召回率往往偏低。

$$P = precision = \frac{TP}{TP + FP};\tag{2}$$

$$R = recall = \frac{TP}{TP + FN}。\tag{3}$$

由于精准率、召回率呈现出信息利用不足的问题,又出现了预测准确程度的指标——F_1。能够在一定程度上综合反应精确率和召回率的分类性能,具体定义为

$$F_1 = \frac{2PR}{P + R}。\tag{4}$$

为了验证文章提出的模型的分类效果,需要将 LGB-BAG 和 LightGBM、RF 的 F_1 的均值和方差进行比较,查看是否确实有提升的效果。

3　实验结果和分析

本文使用 5 折交叉检验(在给定的建模样本中,拿出大部分样本(训练集)进行建模,留小部分样本(测试集)用已经建立的模型进行预测),对 LGB-BAG 模型和 LightGBM、RF 的 F_1 均值和 F_1 方差进行比较,以确定各个模型在信用风险预测的准确程度。表 5 显示了 LGB-BAG 模型在不同 N 和 M 情况下的 F_1 均值和 LightGBM、RF 的 F_1 均值。

由表 5 的结果表明:

(1)在 N 较小的时候,LGB-BAG 的 F_1 均值不能稳定地大于 LightGBM 和 RF 的 F_1 均值,其值会随着 N 的增加产生波动。

(2)但是当 N 增大到一定程度的时候,即 $M=2$ 以及 $N \geqslant 1800$,$M=5$ 以及 $N \geqslant 1400$,$M=10$ 以及 $N \geqslant 1200$,LGB-BAG 的 F_1 均值都会稳定大于其余两种模型 F_1 均值。

(3)随着 M 的增大,N 的临界点随之降低(分别 N 为 1800、1400、1200),即 M 增大后,LGB-BAG 的

表 5　不同 N 值下各个模型 F_1 均值

N	LGB-BAG ($M=2$)	LGB-BAG ($M=5$)	LGB-BAG ($M=10$)	LightBGM	RF
200	0.68869	0.69682	0.68588	0.69899	0.67868
500	0.68942	0.69640	0.69220	0.69903	0.68828
800	0.69658	0.70637	0.69816	0.69746	0.68403
1000	0.69966	0.69107	0.68484	0.69601	0.68468
1200	0.70097	0.69161	0.70197	0.69436	0.6803
1400	0.69571	0.70259	0.69725	0.69319	0.68509
1600	0.69052	0.69871	0.69896	0.69348	0.69073
1800	0.69715	0.70695	0.70506	0.69255	0.68781
2000	0.70077	0.71123	0.71175	0.69202	0.67645
2200	0.70363	0.70729	0.69928	0.69134	0.68283

F_1 均值能更早地稳定地大于其余两种模型的 F_1 均值。

(4)随着 M 的增加，LGB-BAG 的最大 F_1 均值也增加，即 $M=2$、5、10 时，LGB-BAG 的最大 F_1 均值分别为 0.70363、0.71123、0.71125，且三种 LGB-BAG 的最大 F_1 均值显然要大于其余两种模型的最大 F_1 均值(分别为 0.69903 和 0.69073)。

下面进一步给出 $M=2$ 以及 $N\geqslant 1800$，$M=5$ 以及 $N\geqslant 1400$，$M=10$ 以及 $N\geqslant 1200$ 时 LGB-BAG 的 F_1 均值和其余两种模型 F_1 均值的对比图，具体如图 6～图 8 所示。

为了衡量模型预测结果的波动性，表 6 则显示了在不同 N 下 LGB-BAG 模型和 LightGBM、RF 的 F_1 的方差。

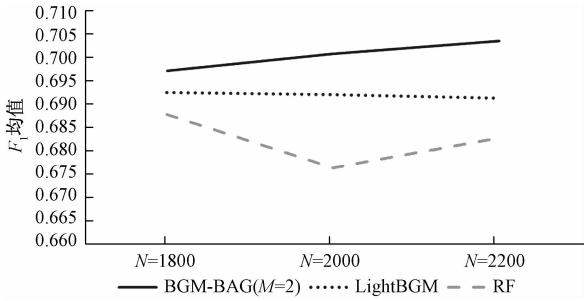

图 6　LGB-BAG($M=2$)和其余两种模型的 F_1 均值

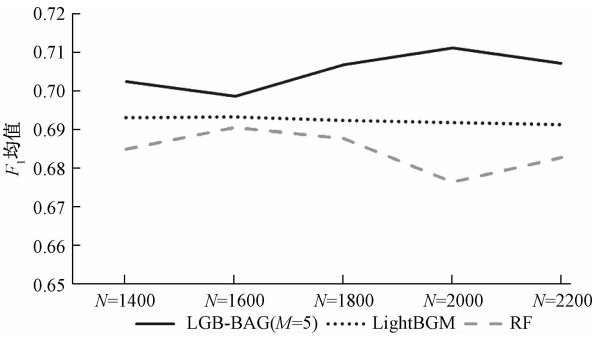

图 7　LGB-BAG($M=5$)和其余两种模型的 F_1 均值

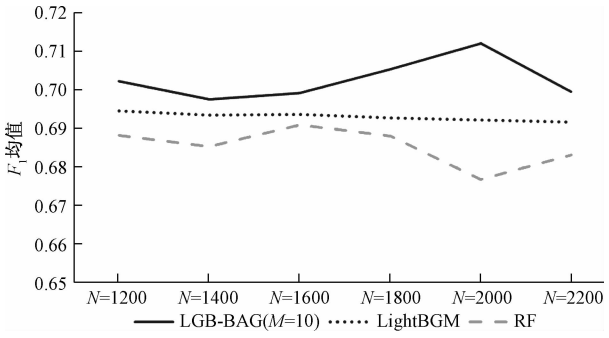

图 8　LGB-BAG($M=10$)和其余两种模型的 F_1 均值

表 6　不同 N 下各个模型 F_1 方差

N	LGB-BAG ($M=2$)	LGB-BAG ($M=5$)	LGB-BAG ($M=10$)	LightBGM	RF
100	0.00068	0.00060	0.00126	0.00064	0.00117
200	0.00073	0.00108	0.00269	0.00081	0.00117
500	0.00104	0.00159	0.00079	0.00150	0.00191
800	0.00132	0.00055	0.00143	0.00162	0.00226
1000	0.00052	0.00110	0.00108	0.00159	0.00114
1200	0.00127	0.00123	0.00139	0.00158	0.00174
1400	0.00102	0.00126	0.00120	0.00173	0.00137
1600	0.00069	0.00059	0.00181	0.00190	0.00196
1800	0.00036	0.00173	0.00146	0.00233	0.00213
2000	0.00096	0.00102	0.00136	0.00225	0.00162
2200	0.00181	0.00202	0.00177	0.00216	0.00206

表 6 的实验结果说明：

(1)在 N 较小的时候，LGB-BAG 的 F_1 方差不能稳定小于 LightGBM 和 RF 的 F_1 方差，随着 N 的增加会产生波动。

(2)当 N 增大到一定程度的时候，即 $M=2$ 以及，$M=5$ 以及，$M=10$ 以及，LGB-BAG 的 F_1 方差都会稳定小于其余两种模型 F_1 方差。

(3)LGB-BAG 的方差趋势和 LightGBM 方差趋势很类似，可能是因为 LightGBM 是其基学习器的原因。

为了进一步说明问题，给出 $M=2$、$M=5$、$M=10$ 以及时 LGB-BAG 的 F_1 方差和其余两种模型 F_1 方差对比，如图 9～图 11 所示。

上述结论都说明在当 N 足够大的时候，LGB-BAG 对个人借款者违约预测更优于 LightGBM 和 RF。LGB-BAG 对比 LightGBM，虽然两者都采用 Boosting，但是结合了 Boosting 和 Bagging 的 LGB-BAG 算法比单纯只用 Boosting 的 LightGBM 算法的预测效果更好；LGB-BAG 对比 RF，虽然两者都采用 Bagging，但是基础分类器的不同导致了最终分类效果的不同，结合了 Boosting 和 Bagging 的 LGB-BAG 算法要比仅用 Bagging 的 RF 算法的预测误差小。

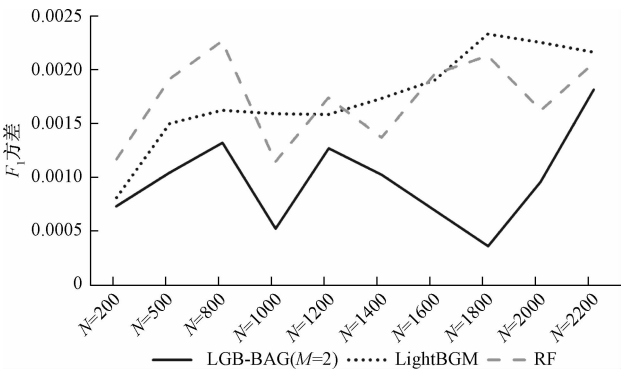

图 9　LGB-BAG($M=2$)和其余两种模型的 F_1 方差

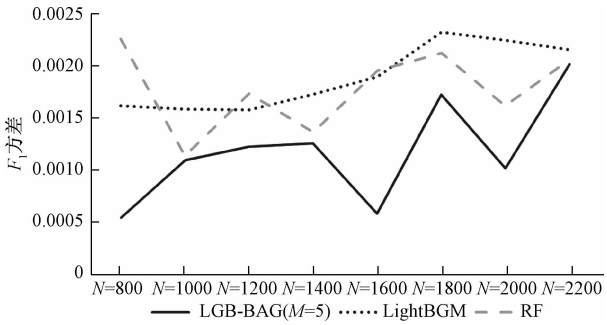

图 10　LGB-BAG(M=5)和其余两种模型的 F_1 方差

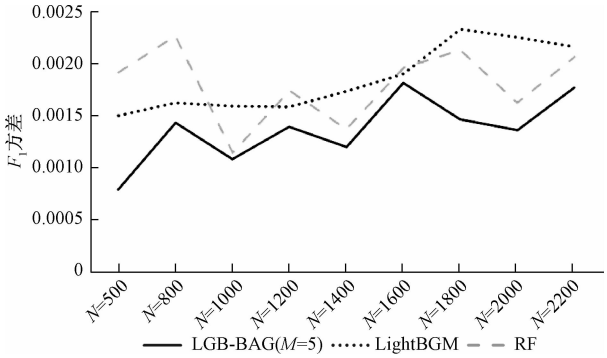

图 11　LGB-BAG(M=10)和其余两种模型的 F_1 方差

4　结论与展望

本文借鉴国内外信用风险评估研究,结合合理性、科学性等原则,根据人人贷网贷平台的特点,集合借款者的信息并结合计算机技术方法设计了借款人的信用风险评估体系。为了提高 P2P 网贷借款人违约预测的准确性,提出集成的 LGB-BAG 模型。LGB-BAG 使用 LightGBM 作为基学习器,借助 LightGBM 能够有效降低模型的偏差的优点,再借助 Bagging 可以减少模型的方差的优点,进一步降低预测的波动性(F_1 方差),使得 LGB-BAG 模型具有较小的偏差与方差,进一步提高模型预测的效果。

本文利用 Python 网络爬虫程序抓取了人人贷个人借款者的实际交易数据进行实证研究。结果表明,当 N 增大到一定程度的时候,LGB-BAG 的预测精度(F_1 均值)要大于其他两个模型,而其方差要小于其他两个模型。证明了本文提出的 LGB-BAG 模型对个人借款者信用风险预测优于 LightGBM 和随机森林。

本文的结果同时说明,集成学习方法可以显著地提高分类的效果,也表明利用集成学习方法来评估个人借款人的信用风险是可行的。

综上所述,基于大数据的背景,结合计算机技术和数学规划,利用借款者的各类信息进行集成学习来预测借款者的信用风险,可以得到很好的预测效果。这类方法可以广泛应用到包括各类 P2P 平台的借款者的信用风险评估中,有助于缓解因征信体系的缺乏以及对风险把控不力而导致的平台危机,促进我国互联网金融的稳健发展。

参考文献

[1] POPE D G, SYDNOR J R. What's in a picture? Evidence of discrimination from prosper[J]. Journal of Human Resources, 2011, 46(1): 53-92.

[2] RAVINA E. Love & loans: the effect of beauty and personal characteristics in credit markets[J]. SSRN Electronic Journal, 2012. DOI: 10.2139/ssrn.1107307.

[3] 廖理, 李梦然, 王正位. 中国互联网金融的地域歧视研究[J]. 数量经济技术经济研究, 2014, 31(5): 54-70.

[4] 李悦雷, 郭阳, 张维. 中国 P2P 小额贷款市场借贷成功率影响因素分析[J]. 金融研究, 2013(7): 126-138.

[5] 赖辉, 帅理, 周宗放. 个人信贷客户信用评估的一种新方法[J]. 技术经济, 2014, 33(9): 97-103.

[6] 谈超, 孙本芝, 王冀宁. P2P 网络借贷平台的羊群行为研究——基于 Logistic 模型的实证分析[J]. 南方金融, 2014(12): 30-37, 53.

[7] 王修华, 孟路, 欧阳辉. P2P 网络借贷问题平台特征分析及投资者识别——来自 222 家平台的证据[J]. 财贸经济, 2016(12): 71-84.

[8] 滕晓慧. 基于 Logistic 模型的 P2P 信贷中同群效应的研究[J]. 中国商论, 2018(19): 44-48.

[9] 王文怡, 程平. 基于 Logistic 和决策树模型的 P2P 网络借贷信用风险研究——以 HLCT 为例[J]. 上海立信会计金融学院学报, 2018(3): 42-55.

[10] 谭中明, 谢坤, 彭耀鹏. 基于梯度提升决策树模型的 P2P 网贷借款人信用风险评测研究[J]. 软科学, 2018, 32(12): 136-140.

[11] SERRANO-CINCA C, GUTIERREZ-NIETO B. The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending[J]. Decision Support Systems, 2016, 89: 113-122.

[12] 师应来, 张冰洁, 姜昊. P2P 网贷平台风险甄别研究[J]. 统计与决策, 2018, 34(16): 153-156.

[13] KOROL T. Early warning models against bankruptcy risk for Central European and Latin American enterprises[J]. Economic Modeling, 2013, 31: 22-30.

[14] NANNI L, LUMINI A. An experimental comparison of ensemble classifiers for bankruptcy prediction and credit scoring[J]. Expert Systems with Applications, 2009, 36(2): 3028-3033.

[15] ABELLÁN J, CASTELLANO J G. A comparative study on base classifiers in ensemble method for credit scoring[J]. Expert Systems with Applications, 2016, 73: 1-10.

[16] FLOREZ-LOPEZ R, RAMON-JERONIMO J M. Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment: a correlated-adjusted decision forest proposal [J]. Expert Systems with Applica-

tions, 2015, 42(13): 5737-5753.
[17] TSAI C F, HSU Y F, YEN D C. A comparative study of classifier ensembles for bankruptcy prediction[J]. Applied Soft Computing, 2014, 24: 977-984.
[18] 操玮, 李灿, 贺婷婷, 等. 基于集成学习的中国 P2P 网络借贷信用风险预警模型的对比研究[J]. 数据分析与知识发现, 2018, 2(10): 65-76.
[19] BREIMAN L. Arcing classifiers[J]. The Annals of Statistics, 1998, 26(3): 801-824.
[20] FREUND Y, SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 1997, 55(1): 119-139.
[21] KE G, MENG Q, Finley T, et al. Lightgbm: a highly efficient gradient boosting decision tree[C]//31st Conference on Advances in Neural Information Processing Systems. Long Beach, CA: Nearal Information Processing Systems Foundation, 2017: 3146-3154.
[22] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012: 18-20.

Application of LGB-BAG in Credit Risk Assessment of Borrowers from P2P Lending

Li Shujin, Ji Xiaojia
(College of Economics, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: Based on the P2P platform, this paper uses the information of the individual borrowers of the P2P platforms to establish a credit risk assessment indicator system to identify borrowers who may default. Therefore, this paper proposes a new LGB-BAG model based on LightGBM (a tree-based Boosting model) and Bagging, which effectively combines the advantages of Boosting and Bagging. The results show that when the number of decision trees increases to a certain extent, the F_1 mean of LGB-BAG is higher than that of LightGBM and random forest; and the F_1 variance of LGB-BAG is also smaller than that of other two models. The F_1 mean of LGB-BAG can reach up to 0.71175, and the LGB-BAG model could improve identification of credit risks of borrowers more effectively.

Keywords: ensemble learning; lightGBM; bagging; boosting; P2P platforms

(上接第 99 页)

Will Banking Structure Affect the Credit Capacity of SMEs? Empirical Study Based on Hierarchical Linear Model

Chi Renyong, Chen Jie, Liao Yaya, Hu Qianqian
(College of Management, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: The financing problem of small and medium enterprises(SMEs) in China has caused widespread concern. Based on the wind database, this paper uses the hierarchical linear model to examine the impact of enterprise size and banking structure on the credit capacity of SMEs, and introduces commercial credit environment variables to study whether the impact will change under different commercial credit environments. It is found that the larger the enterprise size and the more dispersed the banking structure, the stronger the credit capacity of the SMEs. But the impact is weakened with the improvement of the commercial credit environment. This study can be great reference in improving the credit capacity of small and medium enterprises(SMEs).

Keywords: banking structure; commercial credit environments; enterprise size; credit capacity of SMEs