计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201000013

基于深度森林的 P2P 网贷借款人信用风险评估方法

王萧萧[1]　王亭雯[1]　马玉玲[2]　范佳奕[3]　崔超然[1]
1 山东财经大学计算机科学与技术学院　济南 250014
2 山东建筑大学计算机科学与技术学院　济南 250101
3 青岛大学商学院　山东 青岛 266000
(xiaoxiao.wangq@aliyun.com)

摘　要　P2P 网络借贷是近年来新兴的一种金融业务模式,具有投资门槛低、交易方便快捷、融资成本低等优点。但在快速成长的同时,借贷过程中的信用风险问题也日益凸显,层出不穷的借款人跑路乃至诈骗事件给行业留下重大阴影。针对该问题,提出一种基于深度森林的网贷借款人信用风险评估方法。首先从借款人的基本信息和历史借款信息两类数据中提取特征;然后通过多粒度扫描和级联森林模块构建深度森林模型,对借款人进行违约预测,同时使用基尼指数计算随机森林的特征重要性评分,并使用波达计数法进行排序融合,从而对模型的预测结果给出一定的解释。在 LendingClub 和拍拍贷两个公开数据集上,将所提出的方法与支持向量机、随机森林和广而深的网络等方法进行了对比,实验表明该方法具有更好的性能,并且特征重要性评分符合人们的直观理解和客观认识。

关键词:P2P 网络借贷;信用风险评估;深度森林;特征重要性;不平衡数据集
中图法分类号　TP391

Credit Risk Assessment Method of P2P Online Loan Borrowers Based on Deep Forest

WANG Xiao-xiao[1],WANG Ting-wen[1],MA Yu-ling[2],FAN Jia-yi[3] and CUI Chao-ran[1]
1 School of Computer Science and Technology,Shandong University of Finance and Economics,Jinan 250014,China
2 School of Computer Science and Technology,Shandong Jianzhu University,Jinan 250101,China
3 School of Business,Qingdao University,Qingdao,Shandong 266000,China

Abstract　P2P online lending is an emerging financial business model in recent years,which has many advantages of low investment threshold,convenient transaction and low financing cost. However,at the same time of rapid growth,the credit risk problem in the lending process has become increasingly prominent,and the endless stream of borrowers running away and even fraud have left a heavy shadow on the industry. Aiming at this problem,a credit risk assessment method of P2P online loan borrowers based on deep forest is proposed. Firstly,the features are extracted from the basic information and the historical loan information of the borrower. Then,the deep forest model was constructed by multi-granularity scanning and cascade forest module to predict the default of borrowers. At the same time,Gini index was used to calculate the feature importance score of random forest,and Borda count method was used to sort and fusion,so as to give a certain explanation to the prediction results of the model. On the two public datasets of LendingClub and Paipaidai,the proposed method was compared with methods such as support vector machines,random forests,and wide and deep networks. Experiments show that the method has better performance,and the feature importance rating is consistent with people's intuitive understanding and objective understanding.

Keywords　Per-to-per lending,Credit risk assessment,Deep forest,Feature impertance,Unbalanced dataset

1　引言

在“互联网金融”的推动下,P2P(per-to-per lending)网络借贷作为一种新型贷款方式应运而生。该贷款方式也称为“个人对个人信贷”,其典型的运行模式为:借贷双方在借贷平台自由竞价,交易成功后贷款人获取利息,平台收取中介费。P2P 网络借贷具有投资门槛低、交易方便快捷、融资成本低等优点。但在快速成长的同时,借贷过程中的信用风险问题也日益凸显,层出不穷的借款人跑路乃至诈骗事件给行业留下重大阴影。据网贷之家统计,截止到 2020 年 6 月,全国平台总数累计 6611 家,其中存在跑路、提现困难、经侦介入的问题平台数目累计 2924 家。

针对该问题,借款人信用风险评估研究受到广泛关注。例如,Ohlson[1] 首次利用逻辑回归(Logistic Regression)搭建信用分类模型;Xiao 等[2] 使用 BP 神经网络构建网贷借款人信用评估模型等研究。尽管已取得了一些卓有成效的成果,但现有研究仍存在以下两个局限性问题:1)多数研究采用支持向量机、决策树等浅层模型进行建模预测,所得模型的性能

基金项目:国家自然科学基金(61701281,62077033)
This work was supported by the National Natural Science Foundation of China(61701281,62077033).
通信作者:崔超然(crcui@sdufe.edu.cn)

仍尚待提升;2)个别研究基于深度学习的方法构建信用风险评估模型,虽然取得了较好的预测性能,但对模型的预测结果缺乏解释。

为解决以上问题,本文提出一种基于深度森林[3]的 P2P 网贷借款人信用风险评估方法,使用借款人的基本信息和历史借款信息构建以深度森林算法为核心的借款人信用风险评估模型,同时以基尼指数[4]构建特征重要性度量,求解模型的特征重要性评分,再从特征重要性度量的角度对预测结果做出一定的解释。深度森林是南京大学 LAMDA 研究组于 2017 年提出的一种以决策树为基分类器的深度模型。该模型具有参数少、易于调参等优点,在不同规模数据集上表现出较强的鲁棒性;另外,作为一种基于决策树的机器学习方法,深度森林相比于基于神经网络的深度学习模型更容易进行理论分析,易于对模型做出一定的解释。本文的主要贡献包括:

1)提出一种基于深度森林的 P2P 网贷借款人信用风险评估方法,这是首次将深度森林应用到该问题中;

2)在两个公开数据集上的实验结果证明了本文方法的优越性,其准确率、召回率和 F_1 值等指标均高于支持向量机、广而深的网络(Wide and deep networks, Wide & Deep)等对比方法,尤其在 LendingClub 数据集中召回率高于次优方法约 16 个百分点;

3)以基尼指数作为特征重要性度量的指标,并使用波达计数法(Borda Count)对不同森林的特征重要性进行融合。

2 相关工作

2.1 P2P 网贷信用风险评估方法

随着机器学习和深度学习的兴起,许多学者将其应用于 P2P 网贷借款人信用风险评估领域,并且已取得十分可观的成果。

近年来,基于机器学习方法的 P2P 网贷借款人信用风险评估受到广泛关注。例如,Lu[5]使用支持向量机(Support Vector Machine, SVM)算法对 P2P 网贷借款人进行违约预测。Tan 等[6]通过 Logistic 筛选特征,并利用梯度提升树(Gradient Boosting Decision Tree, GBDT)构建 P2P 网贷借款人信用风险评估模型,能够对借款人行为做出准确预判。Xu[7]提出基于随机森林(Random Forest, RF)的 P2P 网贷借款人信用风险评估模型,并引入代价敏感学习方法以增强模型的实用性。另外,Ma 等[8]通过改进的代价敏感决策树对网贷借款人进行信用风险评估;Zhang 等[9]使用模糊 SVM 减小数据类别大小差异对 P2P 网贷借款人信用评估模型的影响,从而提高预判精度。

随着深度学习的广泛应用,部分学者开始将其用于 P2P 网贷借款人信用风险评估的建模工作中。Wang 等[10]提出一种基于注意力机制长短期记忆网络(Long Short-Term Memory, LSTM)的 P2P 网贷借款人信用风险评估方法,将借款人的网站行为当作事件,利用 Word2vec 模型将其转换成向量,然后基于注意力机制 LSTM 方法对借款人的违约行为进行预测。Yang 等[11]提出一种名为 DeepCredit 的深度架构并将其用于 P2P 网贷借款人信用风险评估中,该模型在预测贷款拖欠率和违约率方面取得较高的精度;同时首次对用户详细点击行为进行分析,发现用户在贷款站点中的还款历史和金融活动顺序对还款行为具有较高价值。Wide & Deep 模型是由 wide 线性部分和 deep 神经网络部分共同组成,前者参数和特征的加权与后者最后一个隐层的输出相加,通过激活函数得到该模型的输出。总体来看,本文采用的深度森林模型和 Wide & Deep 模型均为深度模型,但前者是基于随机森林的方法,而后者是基于神经网络的方法。Bastani 等[12]基于广而深的网络(wide and deep networks)建立了一个两阶段的 P2P 网络贷款评分方法,第一阶段识别出非违约贷款,然后将这些贷款移至第二阶段预测获利能力,最后综合两阶段结果输出最优贷款。

2.2 深度森林的应用

本文使用深度森林构建 P2P 网贷借款人信用风险评估模型,深度森林提出之后,被广泛应用于数据挖掘、推荐系统、自然语言处理和图像分类等领域。例如,Tong 等[13]提出基于深度森林的量表数据挖掘方法,分别对老年健康综合评估数据库中的两个量表进行对比分析。实验表明,在保证分类性能基本不变的情况下,可进一步减少提取到的关键属性数量。Lev 等[14]基于深度森林模型提出了 Siamese Deep Forest (SDF)模型,实现了相似性度量的学习。Ge 等[15]基于阿里平台提供的移动推荐大赛数据集,使用深度森林构建用户购买行为预测模型,实验结果表明,该模型在降低时间开销的同时,也提高了预测准确率。Lu 等[16]通过提取恶意代码图像的 HOG 特征来对恶意代码进行分类,使用深度森林的多粒度扫描将每个 HOG 特征向量分割为多个片段,从而提高表征学习能力和获得更长的上下文,提高了分类准确率。据我们所知,这是首次将深度森林用于 P2P 网贷借款人信用风险评估问题当中。

3 P2P 网贷借款人信用风险评估模型的构建

3.1 问题描述

为了后续更好地描述算法,我们定义了训练数据集 $T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_i,y_i),\cdots,(\boldsymbol{x}_N,y_N)\}$,其中 N 是借款人的数目;$\boldsymbol{x}_i\in R^n$ 表示第 i 个借款人的基本信息和历史借款信息。$y_i\in\{1,0\}$ 表示第 i 个借款人的履约情况,取值为 1 和 0 时,分别表示借款人"违约"和"正常履约"的情况。本文的目标是构建一个模型,寻找函数 $f(\boldsymbol{x}_i\rightarrow\hat{y}_i)$,使其根据借款人的基本信息和历史借款人信息,预测其未来的履约状态。

如图 1 所示,本文首先对借款人的基本信息和历史借款信息进行数据清洗和特征工程等特征预处理步骤,然后构建以深度森林为核心的借款人信用风险评估模型,最后使用基尼指数求解特征重要性评分。

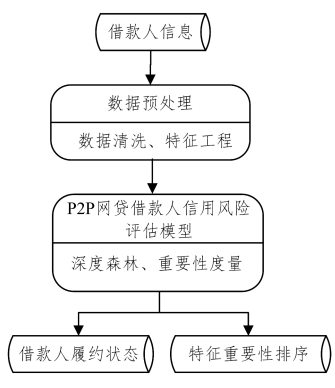

图 1 P2P 网贷借款人信用风险评估模型的框架

Fig. 1 Framework of credit risk assessment of P2P online loan borrowers

3.2 深度森林

本文 P2P 网贷借款人信用风险评估模型的核心是深度森林方法，它包括多粒度扫描和级联森林两个模块。

在多粒度扫描步骤中，设输入变量为 m 维，待解决问题为二分类。首先，分别经过不同大小的滑动窗口进行滑动采样，生成不同粒度的样本向量。例如，m 维样本向量经过 4 * 1 的滑动窗口，会生成($m-3$)个样本向量，每个向量是 4 维。然后，使用相同粒度的样本向量训练两个随机森林，并将森林的预测结果连接成特征向量，用于训练级联森林模块。使用不同大小的滑动窗口进行采样，可以生成更多的样本，从而增强级联森林的表征能力。

级联森林是类似神经网络的层状结构，每一层由若干随机森林构成。每层级联森林会预测生成新的类别标签向量，向量中的每一维分别表示样本属于对应类别的概率值，将类别标签向量与多粒度扫描得到的特征向量连接后作为下一层级联森林的输入。这样层层传递，直至验证集的性能没有显著提升。然后对最后一层所有森林生成的各个类别的概率求平均，此时具有最大概率值的类别为最终的预测结果。

经过预处理后，设训练集为 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_i,y_i),\cdots,(\boldsymbol{x}_H,y_H)\}$。假设问题为二分类，则深度森林方法的训练过程如算法 1 所示。

算法 1 深度森林训练过程

输入：训练数据集 D；

输出：深度森林模型。

1. 设输入特征 $\mathbf{x}_i$ 的维度为 m，某个滑动窗口的扫描维度为 v，扫描步长为 b，则经过该滑动窗口扫描之后的样本向量数目为 n=(m−v)/b+1。
2. 利用前一步骤得到的样本向量，训练两个随机森林，并生成 2n 个二维概率向量，每一维分别表示该样本违约和正常履约的概率。拼接上述概率向量，可得长度为 2n 的一维特征向量 $\mathbf{w}$。
3. 假设设置 u 个滑动窗口，重复调用步骤 1 和步骤 2 u 次，可得 u 个一维特征向量$\{\mathbf{w}_1,\mathbf{w}_2,\cdots,\mathbf{w}_u\}$。
4. $\mathbf{w}_1$ 为整个级联森林的输入。在内部的每一层森林中，将$\{\mathbf{w}_1,\mathbf{w}_2,\cdots,\mathbf{w}_u\}$中的特征向量依次与该层的输出进行拼接，作为下一层森林的输入。每层森林由 4 个随机森林构成，分别输出 4 个二维类别标签向量。
5. 设第 l 层和第 l−1 层森林的预测准确率分别是 θ_l 和 θ_{l-1}，如果 $\theta_l-\theta_{l-1}<0$ 则调用步骤 4)；否则，结束训练。

当滑动窗口的数目 u 为 3，级联森林的层数 l 为 5 时，深度森林的训练过程如图 2 所示。为减弱过拟合现象，每个森林均采用 k-fold 交叉验证，即每个样本都会被用作 $k-1$ 次训练以及产生 $k-1$ 个类别向量，对训练结果取平均并作为下一级的增强特征。如果验证集性能没有显著提升，则停止训练。该操作自动决定了级联层数，换句话说，深度森林可以自动调节模型的复杂度。这使得深度森林能够训练不同规模的数据集，从而避免了传统神经网络因模型复杂度高而不能应用于小规模数据集的问题。

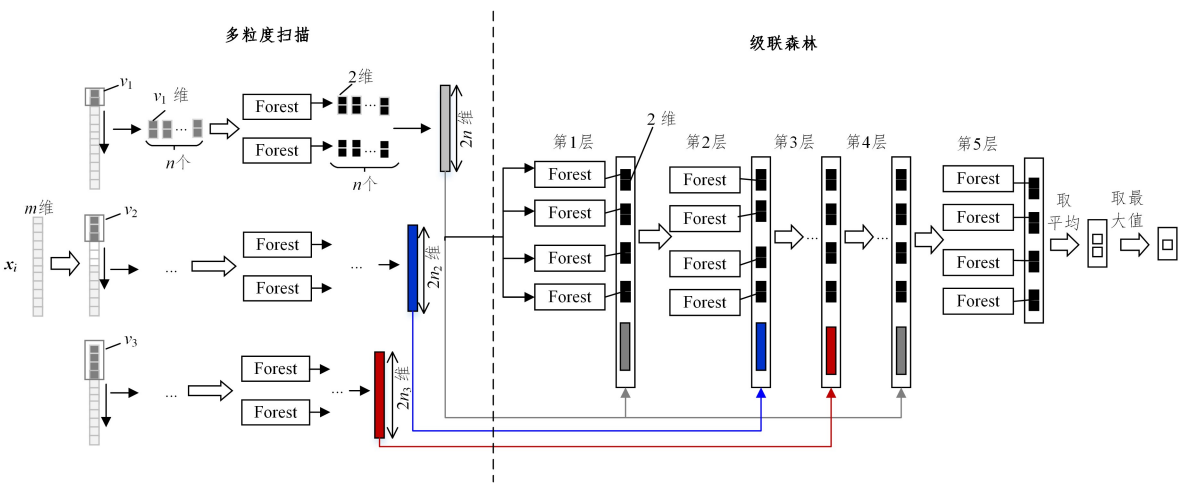

图 2 $u=2,l=5$ 时的深度森林训练过程示意图

Fig. 2 Schematic diagram of deep forest training process when $u=2,l=5$

3.3 特征重要性评分

本文使用基尼指数构建特征重要性度量，首先计算出单个随机森林的特征重要性评分，然后使用 Borda Count 对深度森林中不同随机森林的特征重要性评分进行融合。

基尼指数用来描述决策树分裂节点的不纯度，即总体内包含的类别越杂乱，基尼指数就越大。GI_A 表示特征 X_i 在节点 A 上的基尼指数，计算公式如式(1)所示。其中 C 为类别数目，p_{Ac} 是节点 A 的样本中 c 类别的样本占的比例。

$$GI_A=1-\sum_{c=1}^{C}p_{Ac}^2 \tag{1}$$

其中，S_{iA} 表示特征 X_i 在节点 A 的重要性评分，即节点 A 分支前后的基尼指数变化量；S_{iA} 越大，分支后的节点越不混乱，特征 X_i 重要性也就越强。其计算公式如式(2)所示：

$$S_{iA}=GI_A-GI_l-GI_r \tag{2}$$

其中，GI_l 和 GI_r 分别表示分支后 l 和 r 两个子节点的基尼指数。

假设随机森林 f 中有 T 棵树，则 X_i 在该森林中的重要性评分 S_i 及其归一化之后的评分分别如式(3)和式(4)所示。

$$S_i=\sum_{t=1}^{T}\sum_{A\in M}S_{iA} \tag{3}$$

$$S_i=\frac{S_i}{\sum_{j=1}^{n}S_j} \tag{4}$$

其中，M 表示特征 X_i 在决策树 t 中出现的节点的集合，n 为特征个数。

由式(2)−(4)可得 n 个特征在随机森林 f 中的重要性评分为 $L_f=\{S_1^f,S_2^f,\cdots,S_n^f\}$。假设深度森林中包含 N 个随机森林，则 n 个特征在不同随机森林中的重要性评分为 $L=\{L_1,L_2,\cdots,L_N\}$。

使用 Borda Count 对重要性评分序列 L 进行融合，具体步骤如下。首先对单个随机森林的评分序列 L 进行排序，排在第 j 位的特征 X_i 的投票 $vote_i^{(j)}=n-j$。然后将所有投票按照相同的特征相加，可得到特征 X_i 的投票结果为 $vote_i=\sum_{q=1}^{N}L_q(vote_i^{(j)})$，故所有特征的投票结果为 $V=\{vote_1, vote_2, \cdots, vote_n\}$。最后对 V 进行归一化操作，可得此深度森林的特征重要性评分，从而对模型结果做出一定的解释。

4 实验结果与分析

4.1 数据来源以及数据预处理

本实验使用两个数据集：1)从 LendingClub 网站上下载的 2007 年至 2015 年期间发放的 887979 条信用标，每支标为一条记录，共有 74 个字段，其中包括借款人基本信息和历史借款信息，具体信息如表 1 所列，其中违约样本占总样本的 7.6%；2)源于拍拍贷真实业务数据，该数据包含 2015 年 1 月 1 日至 2017 年 1 月 30 日的 53695 支不同借款人的信用标，该数据集主要用于实验验证，具体信息如表 2 所列。

表 1 LendingClub 数据集信息

Table 1 Information description of LendingClub dataset

基本信息	历史借款信息
房屋拥有情况、初始评级、收入的核实情况、工作年数等	借款金额、承诺还款金额、期数、借款利率，迄今为止的利息和、最近收到的总付款金额、发起贷款月份、未偿还的本金金额等

表 2 拍拍贷数据集信息

Table 2 Information description of Paipaidai dataset

基本信息	历史借款信息
初始评级、年龄、性别、手机认证、户口认证、征信认证、学历认证等	借款金额、借款期限、借款利率、借款成功日期、借款类型、是否首标、历史成功借款次数、已还本金等

数据预处理主要包括两个步骤：数据清洗和特征预处理。第一步对样本进行数据清洗，首先筛选出缺失值大于 95% 的特征，检验以上特征是否与违约有密切关系，然后从违约样本中统计特征的缺失值情况。使用“MISSING”填充类别型特征的空缺值；对于数值性特征，去除异常值后，再使用对应特征均值填充空缺值。第二步进行特征预处理，首先对原始的特征进行加工，生成衍生变量，对于日期特征，使用日期与对应特征最早日期的差替换原始值；对于类别较多的类别型特征，为降低计算量，减少其类别数目。然后对数值型和类别型特征分别进行归一化和 Onehot 编码。LendingClub 和拍拍贷数据集最终处理后的数据分别为 62 维和 58 维。

4.2 评价指标

本实验使用准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F_1 值、ROC 曲线(Receiver Operating Characteristic curve)和 AUC(Area Under Roc)作为实验的评价方法。TP, TN, FP 和 FN 的定义如表 3 所列，例如 TP 表示预测违约、实际也违约的样本数目。

表 3 TP, TN, FP 和 FN 的定义

Table 3 Definition of TP, TN, FP and FN

	预测违约	预测正常履约
实际违约	TP	FN
实际正常履约	FP	TN

准确率表示样本中被预测正确的比例，计算公式如式(5)所示。

$$Accuracy=\frac{TN+TP}{TP+FN+FP+TN} \tag{5}$$

精确率表示预测出的违约样本中预测正确的样本比例，计算公式如式(6)所示。

$$Precision=\frac{TP}{FP+TP} \tag{6}$$

召回率表示实际违约样本中被预测正确的比例，计算公式如式(7)所示。

$$Recall=\frac{TP}{FN+TP} \tag{7}$$

如式(8)所示，F_1 值是分类问题的衡量指标，也表示为精确率和召回率的调和平均数，最大值为 1，最小值为 0。

$$F_1=\frac{2\cdot Precision\cdot Recall}{Precision+Recall} \tag{8}$$

ROC 曲线，以 TPR 为纵轴，FPR 为横轴，用来评估模型的性能。其中，TPR 为召回率，FPR 表示负样本被判错的比例，计算公式如式(9)所示。相同 FPR 的 TPR 越高，ROC 曲线越凸、越接近左上角，表明其诊断价值越大。

$$FPR=\frac{FP}{FP+TN} \tag{9}$$

任取一对正负样本，AUC 指正样本概率值大于负样本概率值的概率，也表示为 ROC 曲线下的面积，取值范围一般是 $[0.5,1]$。当不同模型的 ROC 曲线没有交点时，AUC 值越大的模型性能越好。AUC 的计算公式如式(10)所示。

$$AUC=\frac{\sum_{p\in P}Rank_p-\frac{n_1*(n_1+1)}{2}}{n_0*n_1} \tag{10}$$

其中，n_1 和 n_0 分别为正、负样本数目，P 为正样本的集合，将全部样本按照概率值从小到大排列并为其标注序号，$Rank_p$ 为正样本的序号。

4.3 参数调试及对比实验

为确保训练集和测试集中各类别样本的比例与原始数据集中相同，使用分层采样交叉切分。由于自然样本中正常履约样本和违约样本数量差距太大，本文定义违约样本为正样本。

多粒度扫描期间，滑动窗口大小的取值为 $[4,8,16]$，使用的随机森林的数目为 2，每个随机森林中决策树的数目为 500；构建级联森林过程中，节点中用于执行拆分的最小样本比例为 0.1，级联层中随机森林数目为 4 个，每个随机森林中决策树的数目为 500，节点用于执行拆分的最小样本比例设置为 0.06，允许级联层数上限数是 20。其他参数均为默认值。

将预处理之后的数据，在正负样本不平衡的情况下，使用模型进行违约预测，并选择逻辑回归、随机森林、SVM、Wide & Deep 模型作为对比方法进行实验，结果如表 4 所列。由于数据集正负样本不平衡，所有方法的召回率和 F1 值均比较低，但深度森林方法的召回率仍比次优方法 SVM 高 16.04%。对于准确率、精确率等其他指标，除了随机森林方法的精确率略高以外，深度森林方法均高于其他方法。另外，本文方法准确率的均值在所有方法中是最高的，同时标准差也相对较小，实验结果表明该方法具有更优的性能和较强的稳定性。

表 4 原始数据集上各方法的性能比较

Table 4 Performance comparison of methods on the original dataset

（单位：%）

Classfier	*Accuracy*	*Precision*	*Recall*	F_1	*AUC*
LR	97.20±0.06	98.65	64.82	78.23	92.67
Random Forest	97.29±0.31	99.99	66.59	79.94	95.88
SVM	93.45±0.22	98.49	68.50	80.80	92.98
Wide & Deep[12]	97.01±1.3	98.78	62.23	76.36	91.38
本文方法	**98.75±0.11**	**99.36**	**84.54**	**91.35**	**98.23**

另外，为平衡正负样本，对预处理之后的数据进行欠采样操作。实验结果如表 5 所列，除了 Wide & Deep 模型的精确度略高之外，深度森林方法的全部指标均高于其他方法。本文方法准确率的均值在所有方法中是最高的，同时标准差也相对较小。以上两个实验表明，在正负样本不平衡的情况下，深度森林方法仍具有较强的鲁棒性。

表 5 欠采样数据集上各方法的性能比较

Table 5 Performance comparison of methods on undersampled dataset

（单位：%）

Classfier	*Accuracy*	*Precision*	*Recall*	F_1	*AUC*
LR	86.17±0.05	94.55	72.24	81.91	93.02
Random Forest	92.21±0.22	98.14	84.07	90.56	96.41
SVM	82.54±0.27	80.96	78.80	79.87	90.57
Wide & Deep	92.11±0.98	97.98	83.66	90.26	96.69
本文方法	**95.72±0.09**	**96.72**	**93.48**	**95.07**	**99.22**

LendingClub 数据集上，对正常履约样本欠采样前后，不同方法的 ROC 曲线如图 3 和图 4 所示。在 ROC 曲线中，曲线越接近左上角(0，1)，性能越好。由图 3 和图 4 可看出，相同 *FPR* 下，深度森林方法的 *TPR* 均高于其他对比方法，说明深度森林方法具有更优的性能。

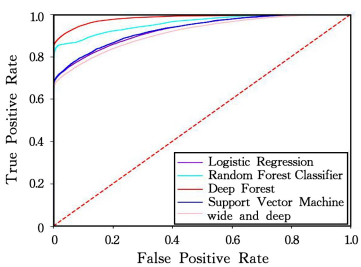

图 3 原始数据集上各方法的 ROC 曲线对比

Fig. 3 ROC curves comparison of different methods on original dataset

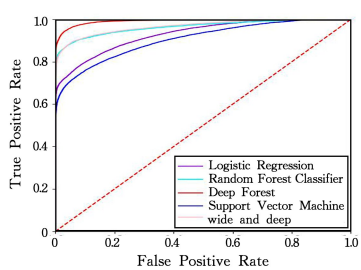

图 4 欠采样数据集上各方法的 ROC 曲线对比

Fig. 4 ROC curves comparison of different methods on undersampled dataset

为验证本文方法的稳定性和普遍性，在拍拍贷数据集上使用该模型对借款人进行信用风险评估。各方法的实验结果如表 6 所列，深度森林方法与对比方法的指标多半大于 95%，尤其是深度森林方法的准确率、精确率等指标均达到了 100%，高于其他对比方法。由于不同方法下的实验结果指标差距太小，故此处未列出对应的 ROC 曲线图。

表 6 拍拍贷数据集上各方法的性能比较

Table 6 Performance comparison of methods on Paipaidai dataset

（单位：%）

Classfier	*Accuracy*	*Precision*	*Recall*	F_1	*AUC*
LR	99.26±0.04	99.90	91.68	95.61	99.98
Random Forest	99.89±0.09	100	99.74	99.87	100
SVM	99.77±0.02	99.82	97.46	98.63	99.99
Wide & Deep	99.30±0.1	99.34	91.77	95.40	99.72
本文方法	**99.99±0.01**	**100**	**100**	**100**	**100**

4.4 特征重要性评分的展示与分析

在 LendingClub 数据集上，本文基于深度森林构建 P2P 网贷借款人信用风险评估模型，求解模型的特征重要性评分，从而对模型做出一定的解释。此处截取了特征重要性排在前十位的特征，其重要性归一化值如表 7 所列。其中第一位的“初始评级”是信用证评定的用户信用等级，分为 A，B，C 三级，每级分为 1，2，3 类，其中 A1 类借款人的信用最优，不同的信用等级反映了借款人的信用好坏。“认证状态”表示借款人的收入是否已由 LendingClub 核实，已被核实的说明借款人的收入是真实的，相对可靠。“所在州”是指借款人在申请贷款时所在的州，本文在处理该特征时，将美国 50 个州按照经济发展水平分为三大类，经济发展水平高的州贷款的人数多，同样违约的人数也多。贷款目的主要分为债务合并、信用卡还款、房屋装修和其他情况，不同借款目的的人群具有不同的违约率。最后对于借款利率、本月应还金额等特征，借款的利率、金额越多，借款人违约的概率就越大。

表 7 LendingClub 数据集上的特征重要性评分

Table 7 Feature importance scores on the LendingClub dataset

序号	特征	重要性归一化值
1	初始评级	0.1714
2	认证状态	0.0654
3	所在州	0.0542
4	借款目的	0.0375
5	借款利率	0.0287
6	本月应还金额	0.0274
7	所有贷款应还总金额	0.0257
8	此贷款发起的金额	0.0257
9	迄今为止收到的利息总共金额	0.0254
10	贷款人承诺的总金额	0.0252

同样地，本文在拍拍贷数据集上，使用基于深度森林的信用模型预测，特征重要性评分排名前十位的特征和重要性归一化值如表 8 所列。在两个数据集中，“初始评级”均位于首位，由此看来贷款人主观预测借款人是否违约时，可以将其作为重要的参考指标。“借款类型”分为应收安全标、电商、普通等，普通标则为最常见的一种标，应收安全标是指贷款人符合应收款安全标的额度大于某值以及借出信用分大于某值的标，电商表示借款人通过电商认证且店铺经营良好的标。可见，划分不同群体对模型的预测结果有一定的影响。另外，手机、户口等认证，侧面反映了借款人所填信息的真实性，对模型预测具有一定的重要性。

表 8 拍拍贷数据集上的特征重要性评分

Table 8 Feature importance scores on the Papadai dataset

序号	特征	重要性归一化值
1	初始评级	0.2106
2	借款类型	0.1456
3	是否首标	0.0473
4	手机认证	0.0446
5	上次还款日期	0.0434
6	性别	0.0403
7	户口认证	0.0313
8	学历认证	0.0303
9	借款金额	0.0286
10	淘宝认证	0.0283

综上所述，该模型可以筛选出对预测借款人是否违约影响比较大的特征，且特征重要性评分符合人们的客观认识和直观理解。

结束语 本文基于深度森林方法对 P2P 网贷借款人进行信用风险评估，在 LendingClub 和拍拍贷数据集的基础上，使用深度森林方法分类预测，同时将其与随机森林、Wide & Deep 模型等分类方法进行对比。实验表明，该方法在两个数据集上表现出的性能均优于其他对比方法，尤其是在数据集样本不平衡的情况下，其优越性更是明显。另外，本文通过特征重要性度量对模型的预测结果做出一定的解释，符合人们的直观理解和客观认识。

针对样本类别不平衡问题，本文简单地利用欠采样技术进行了平衡，后续工作中可以结合代价敏感学习或其他更有效的类别不平衡学习方法进一步提升模型性能。另外，为增强模型实用性和稳定性，可针对反作弊对抗场景进行更充分的评估。

参 考 文 献

[1] OHLSON J A. Financial Ratios and the Probabilistic Prediction of Bankruptcy [J]. Journal of Accounting Research, 1980, 18(1): 109-131.

[2] XIAO H M, HOU Y, CUI C N. Evaluation of P2P Lending Borrower's Credit on BP Artificial Neural Network [J]. Operations Research and Management, 2018, 27 (9): 112-118.

[3] ZHOU Z H, FENG J. Deep Forest: Towards An Alternative to Deep Neural Networks[C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. 2017: 3553-3559.

[4] BREIMAN L, FRIEDMAN J, OLSHEN R, et al. Classification and Regression Trees[M]. New York: Chapman & Hall, 1984.

[5] LU H Y. Construction of risk evaluation system of P2P online loan platform based on SVM [J]. Science and Technology Economics Market, 2018(2): 70-74.

[6] TAN Z M, XIE K, PENG Y P. Research on Credit Risk Evaluation of P2P Online Borrowers Based on Gradient Boosting Decision Tree Model [J]. Soft Science, 2018, 32(12): 136-140.

[7] XU T T. Application of random forest in credit risk assessment of P2P online loan borrowing [D]. Jinan: Shandong University, 2017.

[8] MA P J, WANG Y, YU L, et al. Risk assessment of P2P network lending based on cost-sensitive decision tree [J]. Computer Integrated Manufacturing System, 2018, 243 (7): 296-302.

[9] ZHANG Y C, SONG X P, LUO Y. Research on Customer Credit Evaluation Based on Fuzzy Support Vector Machine [J]. Statistics and Decision, 2008(7): 16-19.

[10] WANG C R, HAN D M, LIU Q G, et al. A Deep Learning Approach for Credit Scoring of Peer-to-Peer Lending Using Attention Mechanism LSTM[J]. IEEE Access, 2018(7): 2161-2168.

[11] YANG Z, ZHANG Y S, GUO B H, et al. DeepCredit: Exploiting User Cickstream for Loan Risk Prediction in P2P Lending[C]// International AAAI Conference on Web and Social Media Twelfth International AAAI Conference on Web and Social Media. Palo Alto, California USA: AAAI, 2018: 444-443.

[12] BASTANI K, ASGARI E, NAMAVARI H. Wide and deep learning for peer-to-peer lending[J]. Expert Systems With Applications, 2019, 134: 209-224.

[13] TONG T, LUO S L, PAN L M, Zhang Tiemei. Scale data mining method based on deep forest[J]. Electronic Design Engineering, 2020, 28(13): 88-91, 96.

[14] UTKIN L V, RYABININ M A. A Siamese Deep Forest [J]. Journal of Knowledge-Based Systems[J]. arXiv: 1704. 08715vl, 2017: 5-6.

[15] GE S L, YE J, HE M X. Prediction model of user purchase behavior based on deep forest[J]. Computer Science, 2019, 46(9): 190-194.

[16] LU X D, DUAN Z M, QIAN Y K, et al. A Malicious Code Classification Method Based on Deep Forest[J]. Journal of Software, 2020, 31(5): 1454-1464.

WANG Xiao-xiao, born in 1996, postgraduate. Her main research interest include data mining and so on.

CUI Chao-ran, born in 1987, professor, is a member of China Computer Federation. His main research interests include information retrieval, multimedia, recommender systems and machine learning.