[文章编号] 1003－4684(2026)02-0139-06

面向非平衡数据的信贷违约风险预测模型

郑 列，刘学丽

（湖北工业大学理学院，湖北 武汉 430068）

[摘 要] 针对信贷非平衡数据的风险预测问题，提出一种基于 RENN 和 Borderline-SMOTE 的综合采样算法和基于 Focal Loss 改进与阈值移动策略的 LightGBM 算法相结合的信用风险预测模型。首先，对多数类样本应用 RENN 算法删除噪声样本，对少数类样本应用 Borderline-SMOTE 算法生成新边界样本，使样本数据集实现相对均衡。然后，使用焦点损失函数优化 LightGBM 算法，改变 Focal Loss 损失函数的类别权重和聚焦参数来提高对少数类样本和困难样本的关注。进一步使用阈值移动策略调整 LightGBM 算法的决策边界，得到最终预测模型。最后，将模型应用于 Lending Club 信贷数据集，结果表明，提出的采样方法和预测模型在 G-mean、AUC、KS 值三个指标上均优于其他模型。

[关键词] 非平衡数据；违约风险预测；混合采样；LightGBM 算法；Focal Loss

[中图分类号] TP391；F830.589 [文献标识码] A

互联网在各行各业加速渗透，带动了金融科技的进步和消费者消费理念的升级，针对个人的网络贷款业务应运而生。然而，对于金融机构和网络贷款平台，个人信贷业务往往收益与风险并存。在大数据时代，利用机器学习技术构建精准识别信贷违约用户的预测模型，并基于预测结果进行全流程的用户审查，能够为金融机构降风险增效率[1]。由于真实数据集中，违约用户的数量远小于正常还款用户数量，即通常是非平衡的。对于此类问题，违约用户的误判易导致高额的损失，所以非平衡数据的处理和难例挖掘工作在机器学习的应用领域愈发重要[2]。

非平衡数据的分类问题，通常指标签不同的目标变量比例失衡的现象[3]。当前国内外学者主要从数据和算法两个层面进行研究。在数据层面，主要对数据进行分布的调整，通过数据过采样、欠采样或数据分组等方法解决数据集分布不平衡的问题。Chawla 等[4]提出经典的过采样 SMOTE 算法，在少数类样本及其邻近同类样本间随机线性插值，而产生新采样以平衡数据集。Han 等[5]在 SMOTE 算法的基础上提出了 Borderline-SMOTE 算法，改善了样本重叠的问题。ADASYN 算法[6]是将 SMOTE 算法与适应性采样进行结合。在算法层面，多采用代价敏感算法[7]、集成算法[8]、关联分类等方式提升少数类样本的拟合性能[9]。Wozniak 等[10]提出了一种有效的代价敏感决策树算法，能够根据给定的代价矩阵构造基分类器，并在随机特征子空间上进行训练。Zong 等[11]引入加权的极限学习机以解决非平衡数据分类问题。

近年来，国内外研究人员针对信用风险评估领域的违约预测问题做了许多研究。现有的信用风险评估模型主要分为三类：第一类是数理统计方法，其中包含判别分析和 Logistic 回归等。林存浩等[1]利用基于分布式 Logistic 模型进行违约用户识别，并处理分布式异质数据以保护用户隐私。第二类是基于机器学习的信用风险预测模型，常用的有支持向量机、随机森林、梯度提升决策树等。张在美等[12]运用优化的辅助分类器生成对抗网络和梯度提升决策树。第三类是基于深度学习神经网络的违约用户预测模型。梁龙跃等[13]利用 GAMI-net 筛选违约样本后，使用基于图卷积神经网络的信用风险预测模型提升了预测精度。牛晓健等[14]将深度神经网络的隐藏层作为“特征提取器”，将深度学习作为研究个人信用风险评估的方法。

综上所述，在现有的数据采样技术和算法改进解决违约用户预测的研究中，数据过采样时未综合考虑噪声样本的处理和样本重叠的问题，以及忽略了不平衡数据建模时样本分类难易程度不同的问

[收稿日期] 2024－01－16
[基金项目] 教育部人文社会科学研究规划基金项目(17YJA790098)
[第一作者] 郑 列(1963－)，男，湖北英山人，湖北工业大学教授，研究方向为应用数学。
[通信作者] 刘学丽(1998－)，女，山东泰安人，湖北工业大学硕士研究生，研究方向为数据挖掘与机器学习。

题。基于此,本文提出一种基于RENN和Borderline-SMOTE算法的综合采样算法RENN-BSMOTE和基于Focal Loss改进与阈值移动策略的LightGBM算法相结合的信用风险预测模型,以提高信用风险预测模型的整体性能。将本文提出的采样方法和预测模型与经典模型相对比,结果表明本文提出的采样方法和预测模型的G-mean、AUC、KS值均优于其他模型。

1 基于RENN-BSMOTE和FLLightGBM难例挖掘的分类模型

1.1 基于梯度提升决策树的特征提取方法

为了在高维特征中提取关键信用特征,本文使用基于梯度提升决策树(GBDT)的特征提取方法进行特征选择。GBDT是基于Boosting的方法,每一步迭代都会选择一个延梯度下降速度最快的弱学习器,逐步优化损失函数。在衡量特征重要性时,所有的非叶子节点在分裂时加权不纯度的减少即为该节点此次分裂的收益。收益越大,其对应特征的重要度越高。特征提取的具体过程如下:

1)遍历训练样本的所有特征,对于特征j,遍历特征j所有的切分点c。

2)使用平方误差最小判断特征集合和最优切分点,公式如式(1)所示:

$$\min_{j,s}\left[\min_{c1}\sum_{xi\in R1(j,s)}(y_i-c_1)^2+\min_{c2}\sum_{xi\in R2(j,s)}(y_i-c_2)^2\right] \tag{1}$$

1.2 基于RENN-BSMOTE的样本均衡方法

因信贷数据集中存在非平衡样本和噪声样本,本文采用RENN算法减少噪声样本对分类效果的影响。RENN算法利用最近邻规则,对于多数类样本,如果此样本的多数K近邻样本与自身类别相反,就将其视为噪声样本并删除;在此基础上,重复进行上述的删除操作,直到样本数据无法再被删除。

信贷数据集的样本数据非平衡比例较高,仅使用欠采样算法无法使样本完全平衡,故使用Borderline-SMOTE算法生成少数类样本,以平衡数据集。Borderline-SMOTE算法是对SMOTE算法的改进,只选择少数类中的边界样本作为根样本进行样本生成,能够改善样本重叠的问题。其核心思想是计算少数类样本的K最邻近样本,若超过一半的K近邻样本属于多数类,则定义为边界样本。仅对于边界样本利用SMOTE算法生成新样本。SMOTE算法的插值方式如式(2)所示,其中$rand(0,1)$表示区间(0,1)内任意一个随机数。

$$S=X_i+rand(0,1)*(Y_j-X_i) \tag{2}$$

1.3 基于Focal Loss改进与阈值移动的LightGBM算法

LightGBM[15]是基于决策树的梯度提升框架,高效、可扩展,能够处理大型数据集和高维特征。因为LightGBM算法支持自定义损失函数,本文引入He Kaiming[16]提出的焦点损失函数Focal Loss对传统的交叉熵损失函数进行改进,以此对违约用户进行难例挖掘。焦点损失函数最初在目标检测领域被提出,用来解决样本分类难易程度不同的问题。由Focal Loss改进的交叉熵损失函数如公式(3)所示,函数引入权重参数α,调节数据不平衡情况;引入聚焦参数γ,增大难分样本的权重,使模型更加关注困难样本的学习。使用Focal Loss改进LightGBM算法,能够同时解决样本不平衡和难分问题,提升对违约样本的识别度。

$$FL=-\frac{1}{n}\sum_{i=1}^{n}(\alpha y_i(1-p_i)^{\gamma}\log(p_i)+(1-\alpha)(1-y_i)p_i{}^{\gamma}\log(1-p_i)) \tag{3}$$

为进一步提升对少数类的识别能力,将FL-LightGBM算法与阈值移动策略相结合。阈值移动(Threshold-moving)是在预测阶段加入再缩放的思想,以缓解类别不平衡问题。设训练集中有m^+个正例和m^-个反例,分类器输出预测正例的概率为y。假设训练集是真实样本空间的采样,则对于任意测试样本,令

$$\frac{1-y'}{y'}=\frac{y}{1-y}\times\frac{m^-}{m^+} \tag{4}$$

若$y'>0.5$,则为正例,若$y'<0.5$,则为反例。

1.4 建立模型

本文从信贷违约风险的特征选择、类别不平衡数据的处理方法、分类算法的优化三个方面出发,以面向非平衡数据的二分类问题的评价指标为优化方向,对信贷数据集的用户违约行为进行预测,构建了基于RENN-BSMOTE和FLLightGBM难例挖掘的信贷违约风险预测模型。建模流程如图1所示,具体算法步骤如下:

1)输入非平衡数据集,对数据进行数据清洗和预处理。

2)利用基于梯度提升决策树的特征提取法,根据节点分裂时的收益确定关键信用的特征组合。

3)设置非平衡尺度a_0,对多数类样本应用RENN算法删除噪声样本,对少数类样本应用Borderline-SMOTE算法生成新样本,直至非平衡尺度$a=a_0$。

4)使用Focal Loss损失函数改进的LightGBM算法进行难例挖掘分类,训练过程中使用网格搜索法确定最佳参数,得到最终分类模型。

5)对测试集进行预测，并使用阈值移动策略判定正例和反例。

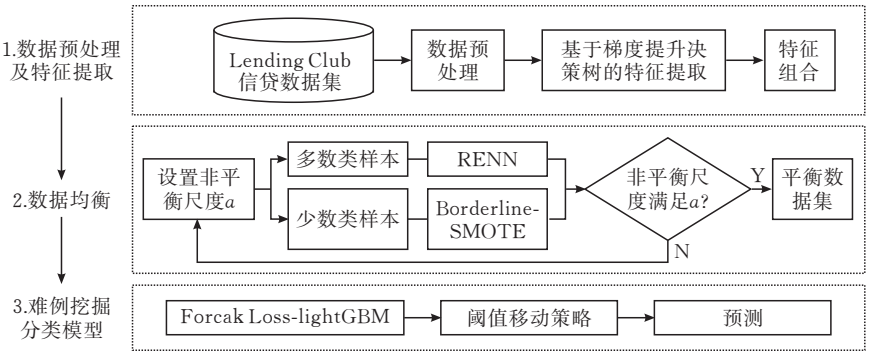

图 1　建模流程

2　实证分析

2.1　数据集

本文使用美国 P2P 网贷平台 Lending Club 公布的 2018 年第一季度的贷款交易数据进行信贷违约风险预测研究。选择前一万条用户样本进行分析，每一例样本由 145 项用户特征指标及对应的贷款状态标签组成。将贷款状态(loan_status)作为目标特征，对目标变量作数值化处理，定义贷款状态中的 Current 和 Fully Paid 为未违约用户，作为标签“0”，定义其他状态为违约用户，作为标签“1”。经统计，该数据集有 9786 例履约样本，214 例违约样本，不平衡比达到 45.7，是典型的高维非平衡信贷数据集。

在特征工程方面，为免引起信息泄露，删除支付总额(total_pymnt)等 15 个贷后特征。接着剔除缺失值比例超过 60%以上的变量，工作职位级别(emp_title)等离散程度较大的无效变量也予以删除。随后利用数据集中的有效特征构造未结清贷款占比等 6 个比例特征。共计删除 72 个特征变量，构造 6 个新特征变量，数据集剩余 79 个特征。在缺失值的处理方面，将缺失值填充为“Unknown”，并对离散变量和连续变量进行变量分箱和 WOE 编码处理。由于用户填写信息时，缺失的部分可能是用户想要隐瞒的信息，所以本文将缺失值作为一个特殊类别进行处理。

2.2　评价指标

对于非平衡数据，若使用传统的以精度最大化为优化方向的评价准则，易产生有偏的分类器[17]。本文从非平衡信贷违约用户预测的实际需求出发，选取 AUC、G-mean 和 KS 值对模型的预测效果进行评价。二分类的混淆矩阵如表 1 所示。

AUC(Area Under ROC Curve)被定义为 ROC 曲线下的面积，即以真阳性率(TPR)为纵轴、假阳性率(FPR)为横轴所作曲线下面积[18]。AUC 值能够衡量学习器的分类能力，取值在 0.5 至 1 之间，越接近于 1，说明模型的分类能力越强。其中，TPR 和 FPR 的计算方法如式(5)和式(6)。

表 1　混淆矩阵

类别	预测正类	预测负类
实际正类	TP	FN
实际负类	FP	TN

$$TPR=\frac{TP}{TP+FN} \tag{5}$$

$$FPR=\frac{FP}{FP+TN} \tag{6}$$

G-mean 定义为 TPR 和 TNR 的几何平均值。当数据集中正类和负类的样本数量相差较大时，G-mean 值更能准确反映分类器在处理不平衡数据集时的性能，反映分类的偏差。G-mean 的计算方法如式(7)。

$$G-mean=\sqrt{\frac{TP}{TP+FN}\times\frac{TN}{TN+FP}} \tag{7}$$

KS(Kolmogorov-Smirnov)值能够衡量对履约用户和违约用户的区分能力。KS 值的取值范围是[0,1]。KS 值在 0.3 至 0.6 之间，表明模型具有较强的区分能力。

2.3　关键信用特征提取

利用基于梯度提升决策树的特征提取方法进行关键信用特征筛选，计算所有的非叶子节点在分裂时的收益，得到 21 项关键信用特征如表 2 所示。

表 2　关键信用特征

排名	特征名称	分裂收益
1	earliest_cr_line	0.235
2	sub_grade	0.0482
3	total_bal_ex_mort	0.0399
4	grade	0.0317
5	num_il_tl	0.0304
6	addr_state	0.0260
7	inq_fi	0.0216
8	tot_cur_bal	0.0216

续表2 关键信用特征

排名	特征名称	分裂收益
9	num_actv_rev_tl	0.0204
10	installment	0.020211
11	num_bc_tl	0.0198
12	int_rate	0.0186
13	total_acc	0.0182
14	application_type	0.0181
15	open_acc_6m	0.0157
16	mo_sin_old_rev_tl_op	0.0154
17	il_util	0.0147
18	total_bal_il	0.0132
19	good_bankcard_rate	0.0126
20	delinq_2yrs	0.0125
21	tot_coll_amt	0.0123

2.4 实验结果分析

为证明本文提出的RENN-BSMOTE数据均衡方法和基于Focal Loss改进与阈值移动策略相结合的难例挖掘技术的有效性和优势，设计如下算法对比实验。

2.4.1 数据均衡方法对比实验 为了验证本文提出的RENN-BSMOTE数据均衡方法的有效性，加入随机过采样算法、SMOTE算法、Borderline-SMOTE算法、ADASYN算法及三种混合采样算法在Lending Club信贷数据集上进行对比分析。为控制其他实验条件，分类器均使用本文提出的Focal Loss改进的LightGBM难例挖掘算法。实验过程中均使用五折交叉验证，降低实验效果的随机性。各个数据均衡方法对应的G-mean、AUC、KS如表3所示。

表3 数据均衡方法的对比实验结果

	G-mean	AUC	KS
ROS	0.3865	0.5671	0.1342
SMOTE	0.7363	0.7366	0.4732
Borderline-SMOTE	0.7187	0.7336	0.4672
ADASYN	0.733	0.738	0.475
SMOTETomek	0.7158	0.7252	0.4505
SMOTEENN	0.7105	0.7236	0.4472
RENN-SMOTE	0.753	0.7618	0.5236
RENN-Borderline-SMOTE	**0.756**	**0.764**	**0.529**

从表3可以看出，相对于广泛使用的四种过采样算法和三种混合采样算法，本文所提出的RENN-BSMOTE混合采样算法在G-mean、AUC、KS三项指标上都具有明显优势。结果表明，使用RENN去除多数类样本的噪声样本后，再结合过采样算法SMOTE和Borderline-SMOTE，能够有效提升用户违约预测模型的几何均值，同时兼顾多数类和少数类样本的区分，KS值达到0.5236和0.529，具有较强的分类能力。而Borderline-SMOTE更能关注到少数类的边界样本，因此RENN-BSMOTE混合采样算法更能够保证整体的分类准确率，性能最好。

2.4.2 分类器对比实验 为了验证本文提出的基于Focal Loss改进与阈值移动策略相结合的LightGBM难例挖掘技术的有效性，加入逻辑回归、随机森林、XGBoost、支持向量机及未改进损失函数的LightGBM算法在Lending Club信贷数据集上进行对比分析。为控制其他实验条件，使用RENN-BSMOTE混合采样算法使数据达到均衡，分类器均利用网格搜索法选择最优参数。实验过程中均使用五折交叉验证，降低实验效果的随机性。不同分类模型的预测结果对应的G-mean、AUC、KS如表4所示。

从表4的结果来看，本文提出的使用Focal Loss损失函数改进的LightGBM算法，结合阈值移动策略后的分类性能和样本区分程度最优，能够在履约样本和违约样本极度不平衡的情况下进行难例挖掘。改进损失函数后，G-mean、AUC和KS值比未改进原始交叉熵损失函数的LightGBM模型分别高9.2%、5.3%和10.7%。阈值移动策略实际上是在决策过程中考虑到样本比例，利用校准的后验概率放宽对占比小的违约样本的分类限制[19]，这可能会加大了履约样本被误分成违约样本的概率。反之，如果不考虑阈值的移动，FLLightGBM模型由于过度聚焦于少数类样本而导致违约用户未被识别，将会导致金融机构的大量损失。所以，在Focal Loss改进的LightGBM模型上使用阈值移动策略是非常必要的。

表4 不同分类模型的对比实验结果

	G-mean	AUC	KS
LR	0.732	0.74	0.499
RF	0.579	0.657	0.313
XGB	0.565	0.653	0.306
SVM-liner	0.731	0.749	0.497
SVM-rbf	0.641	0.691	0.381
LightGBM	0.664	0.711	0.422
FLLightGBM(TM)	**0.756**	**0.764**	**0.529**

2.4.3 参数分析实验 在FLLightGBM难例挖掘模型中，Focal Loss焦点损失函数引入权重参数α和聚焦参数γ两个参数(α,γ)，分别调节样本类别权重和样本难度权重。为了分析两个参数的作用并找到最佳参数组合，本文经过测试选择了八组参数进行分析，分类结果如表5所示。为了直观展现各参数的变化情况，以参数组合为横坐标，评价指标值

为纵坐标作图，如图 2 所示。

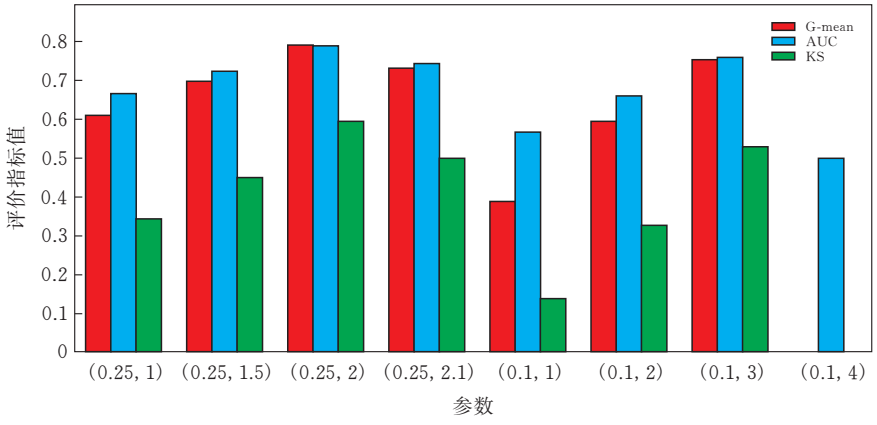

图 2　FLLightGBM 模型参数对比

分析图 2 可知，当权重参数 α 取 0.25 时，模型的三个评价指标都优于 α 取 0.1。取定 α 后，改变 γ 的取值，模型的预测能力和区分度会呈现先上升后下降的趋势。当 (α,γ) 的取值为(0.25，2)时，G-mean、AUC 和 KS 值高于其他参数组合，此时的参数组合适用于信贷违约数据集上的 FLLightGBM 难例挖掘模型。

表 5　FLLightGBM 模型参数对比分析

	G-mean	AUC	KS
(0.25,2.1)	0.738	0.751	0.502
(0.25,2)	**0.798**	**0.799**	**0.598**
(0.25,1.5)	0.704	0.727	0.454
(0.25,1)	0.611	0.672	0.344
(0.1,1)	0.388	0.57	0.14
(0.1,2)	0.596	0.665	0.33
(0.1,3)	0.758	0.766	0.533
(0.1,4)	0	0.5	0

3　结论

为解决非平衡信贷数据集导致的分类效果差的问题，本文提出了 RENN-BSMOTE 混合采样算法和 Focal Loss 难例挖掘技术，并在测试阶段使用阈值移动策略，共同提升模型对违约用户的预测能力。对 Lending Club 信贷违约数据集进行信用风险预测，实验结果表明，RENN-BSMOTE-FLLightGBM(TM)信贷风险预测模型相较于其他常用模型表现更优，能够在数据分布失衡的信贷违约数据集上有着良好的预测性能，同时兼顾了对少数类难例样本的挖掘能力。未来将考虑高维特征的提取方法，在数据处理方面更优，期望能使模型的预测更加准确。

[参　考　文　献]

[1]　林存洁，肖凤，乔楠.基于分布式 Logistic 模型的多源对公贷款违约预测方法[J/OL].数理统计与管理：1-18[2024-01-15]. https：// doi. org/10. 13860/j. cnki. sltj. 20231030-009.

[2]　苗立志，白瑞思蒙，翟月昊，等.面向非平衡数据的癌症患者生存预测分析[J].计算机工程，2021，47(12)：316-320.

[3]　陈福康.不平衡数据下信贷违约风险控制研究[D].济南：山东大学，2021.

[4]　CHAWLA N V，BOWYER K W，HALL L O，et al. SMOTE：Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research，2002，16(01)：321-357.

[5]　HAN H，WANG W，MAO B. Borderline-SMOTE：A new over-sampling method in imbalanced data sets learning[C]. // Proceedings of the 2005 International Conference on Intelligent Computing.Cham：Springer，2005：878-887.

[6]　HE H，BAI Y，GARCIA E A，et al. ADASYN：Adaptive synthetic sampling approach for imbalanced learning[C]. // Proceedings of the 2008 International Joint Conference on Neural Networks. Piscataway：IEEE，2008：1322-1328.

[7]　鲁淑霞，张振莲，翟俊海.代价敏感惩罚 AdaBoost 算法的非平衡数据分类[J].南京航空航天大学学报，2023，55(02)：339-346.

[8]　曹玉茹，高洋洋.基于 SHAP 值惩罚特征的集成分类方法研究[J].统计与决策，2023，39(06)：21-26.

[9]　赵楠，张小芳，张利军.不平衡数据分类研究综述[J].计算机科学，2018，45(S1)：22-27.

[10]　KRAWCZYK B，WO Ź NIAK M，SCHAEFER G. Cost-sensitive decision tree ensembles for effective imbalanced classification[J]. Applied Soft Computing，2014，14：554-562.

[11]　ZONG W，HUANG G，CHEN Y. Weighted extreme learning machine for imbalance learning[J]. Neurocomputing，2013，101：229-242.

[12]　张在美，吕娟，刘彦.基于优化 ACGAN-GBDT 的个人

信用风险评估模型研究[J].财经理论与实践，2022,43(05)：84-89.
[13] 梁龙跃,王浩竹.基于图卷积神经网络的个人信用风险预测[J].计算机工程与应用，2023,59(17)：275-285.
[14] 牛晓健,凌飞.基于组合学习的个人信用风险评估模型研究[J].复旦学报(自然科学版)，2021,60(06)：703-719.
[15] KE G，MENG Q，FINLEY T，et al. LightGBM：A highly efficient gradient boosting decision tree[J]. Advances in Neural Information Processing Systems，2017,30：3146-3154.
[16] LIN T Y，GOYAL P，GIRSHICK R，et al.Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2017 (99)：2999-3007.
[17] 林智勇,郝志峰,杨晓伟.若干评价准则对不平衡数据学习的影响[J].华南理工大学学报(自然科学版)，2010,38(04)：147-155.
[18] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters，2006,27(08)：861-874.
[19] 李剑锋,柏雪,赵春财,等.融合生成对抗网络和难例挖掘的产品质量预测模型[J/OL].计算机集成制造系统：1-18[2024-01-15]. http：// kns. cnki. net/kcms/detail/11.5946.TP.20231101.1138.002.html.

Risk Prediction Model of Credit Default Oriented to Unbalanced Data

ZHENG Lie，LIU Xueli

(*School of Science*，*Hubei Univ. of Tech.*，*Wuhan* 430068，*China*)

Abstract：Aiming at the risk prediction problem of credit unbalanced data，a credit risk prediction model combining a comprehensive sampling algorithm based on RENN and Borderline-SMOTE and a LightGBM algorithm based on the Focal Loss improvement with threshold shift strategy is proposed. First，the noise samples in the majority class samples are reduced by the RENN algorithm and the boundary samples in the minority class are increased by using the Borderline-SMOTE algorithm to make the sample dataset relatively balanced. Then，the LightGBM algorithm is optimized using the Focal Loss function，and the category weights and focus parameters of the Focal Loss function are varied to increase the focus on minority class samples and difficult samples. Further，the decision boundary of the LightGBM algorithm is adjusted using the threshold shift strategy to obtain the final prediction model. Finally，the model is applied to the Lending Club credit dataset，and the results show that the sampling method and prediction model proposed in this paper outperform other models in terms of G-mean，AUC，and KS values.

Keywords：unbalanced data；default risk prediction；hybrid sampling；LightGBM algorithm；Focal Loss

[责任编校：闫 品]