一种面向非均衡样本的企业金融风险预测方法

仇 实[1] 高影繁[1] 姚长青[1] 刘志辉[1] 李佳星[2]
（1.中国科学技术信息研究所，北京 100038；2.河北银行，河北石家庄 050011）

摘要：在企业所面临的众多风险中，企业金融风险表现尤为突出，而在大数据环境下，严重的数据不均衡成为横亘在企业金融风险分析面前的一道鸿沟。本文针对企业竞争情报分析中的样本不均衡问题，以金融企业信贷风险预测为切入点，提出一种面向非平衡样本的企业风险识别方法。该方法采用人工智能分析领域中的特征选择、非均衡样本平衡处理和集成学习等智能分析手段，为大数据环境下企业竞争情报中的企业风险识别问题提供解决思路。

关键词：非平衡样本；金融企业；信贷风险；Catboost；EasyEnsemble
中图分类号：TP391 **文献标识码：**A **DOI：**10.3772/j.issn.1674-1544.2021.05.002

Enterprise Financial Risk Forecasting Method Based on Unbalanced Samples

QIU Shi[1], GAO Yingfan[1], YAO Changqing[1], LIU Zhihui[1], LI Jiaxing[2]
(1.Institute of Scientific and Technical Information of China, Beijing 100038; 2.Bank of Hebei, Shijiazhuang 050011)

Abstract: Among the many risks faced by enterprises, corporate financial risks bear the brunt. In the context of big data, serious data imbalance has become a chasm in front of the analysis of corporate financial risks. Aiming at the problem of sample imbalance in competitive intelligence analysis of enterprises, this paper takes credit risk prediction of financial enterprises as the breakthrough point, and puts forward an enterprise risk identification method for unbalanced samples. This method adopts a variety of intelligent analysis methods in the field of artificial intelligence analysis, such as feature selection, unbalanced sample balance processing and ensemble learning, to provide solutions to the problem of enterprise risk identification in enterprise competitive intelligence under the environment of big data.

Keywords: unbalanced samples, financial enterprises, credit risk, Catboost, EasyEnsemble

1 背景与研究现状

随着大数据时代的到来，企业的竞争环境发生了巨大的改变。竞争环境的变化不断对企业产生威胁，也不断对企业产生机会。对企业来说，如何检测竞争环境的变化、规避威胁、抓住机会，已经成为企业竞争情报研究中的重要课题。在企业所面临的众多风险中，企业金融风险是关键，而数据的不均衡性严重影响了企业金融风险分析。以金融企业的信贷风险分析为例，由于违

作者简介：仇实（1997—），男，中国科学技术信息研究所硕士研究生，主要研究方向为科技金融数据挖掘（通信作者）；**高影繁**（1974—），女，博士，中国科学技术信息研究所研究员，硕士生导师，主要研究方向为文本挖掘、知识组织；**姚长青**（1974—），男，博士，中国科学技术信息研究所研究员，硕士生导师，主要研究方向为情报理论与方法；**刘志辉**（1979—），男，博士，中国科学技术信息研究所研究员，主要研究方向为情报分析方法、竞争情报、学科情报与战略情报研究；**李佳星**（1992—），女，硕士，河北银行工程师，主要研究方向为金融科技。

基金项目：中国科学技术信息研究所重点工作项目“上市公司年报数据库建设及服务系统研发”（ZD2021-10）。
收稿时间：2021 年 6 月 1 日。

约的客户数量要远少于履约的客户数量，这种数据的类别不平衡问题会使得训练出的模型更容易将风险客户划为履约客户，从而加大了预测金融企业信贷违约风险的难度。为解决大数据环境下企业竞争情报中的企业风险识别问题，本文将针对企业竞争情报分析中的样本不均衡问题，以金融企业信贷风险预测为切入点，提出一种面向非平衡样本的企业风险识别方法。

1.1　情报信息分析中的样本不均衡问题研究现状

在情报信息分析的应用场景中，不平衡样本问题始终存在。王刚等[1]基于情感知识和机器学习两种方法，根据电子商务意见挖掘数据不均衡的语言特征，提出基于非均衡数据分析的意见挖掘方法；翁梦娟等[2]采用卷积神经网络作为融合分类器的分类方法，提高不均衡数据集下中图分类法的标引精度。在医学情报信息分析领域，非均衡样本问题更是广泛存在。陈钊志等[3]面向心衰医疗数据中的样本不均衡现象，在欠采样后的数据上训练局部敏感判别矩阵型分类器并得到了较好的预测结果；郭玉萱等[4]提出了一种基于权值的过采样方法，并将其与Bagging和Boosting两种集成学习模型结合，解决了鉴别草药肝毒性场景下的不均衡样本问题。

1.2　机器学习在信贷风险预测中的应用研究

现实场景下的数据往往具有维度大、无关信息多、信息冗余等特点，学者们也因此积极探索与应用特征遴选方法，如在文本特征选择中使用二进制烟花算法[5]和基于信息增益的特征筛选方法[6]。在信贷风险预测的应用场景中，学者们也尝试了多种特征选择方法，如凌健等[7]提出一种基于Relief评估和SVM交叉验证的特征选择方法；张雷等[8]在训练信用评估模型前使用随机森林进行特征提取和特征重要性的评估。由于相关的研究中多数都没有针对信贷数据的样本不均衡现象对方法进行改进，这可能会导致对分类有效的特征无法被成功筛选。

当下在构建金融企业信贷风险预测模型时，学者常使用两类方法：一是以逻辑回归法和判别分析法为代表的传统方法，二是机器学习方法。由于前者对数据集的分布有着严格的假设和前提，预测效果也因此受到限制，实际表现往往弱于机器学习方法。在具体应用中，有学者使用SVM算法对信贷风险进行预测，并尝试将SVM与不同的算法进行组合以提升预测效果[9]；王思宇等[10]基于集成学习的LightGBM算法进行信用风险评估，与朴素贝叶斯、决策树、随机森林、XGBoost等机器学习算法的比较结果表明，在使用AUC作为模型评估指标时LightGBM算法的得分最高；马晓君等[11]运用CatBoost 算法构建 P2P违约预测模型，与LightGBM和XGBoost算法进行对比，全面分析了CatBoost算法的性能优势。

1.3　基于采样法的非平衡样本处理方法研究

如果不进行任何处理，直接使用样本不平衡的数据进行模型训练，得到的分类器将倾向于把待预测的数据判断为多数类。因此，为了得到更合理的预测模型，一些学者在构建信贷风险预测模型前，会利用采样法对不平衡数据进行处理。采样法较为常见的两种形式是过采样和欠采样。

过采样方法通过某种策略增加少数类样本，使少数类样本数量与多数类样本数量达到平衡。最简单的过采样方法即为随机选择少数类样本数据进行复制，直至少数类样本的数量等于多数类样本的数量。但这种方式会使训练集中的少数类存在大量相同样本，极大地增加了模型过拟合的风险。基于此，有学者提出了过采样算法SMOTE[12]。该算法通过对少数类样本和其附近的同类别样本之间随机进行线性插值来合成新的少数类样本，这在一定程度上降低了过拟合。Borderline−SMOTE是一种SMOTE的改进算法，其主要思想是使用少数类样本边界附近的样本合成新样本，从而让新合成的样本更有利于模型的训练[13]。但与同期其他的SMOTE改进算法一样，Borderline−SMOTE很容易受到数据噪声的影响。2018 年，提出了一种基于k−Means和SMOTE的过采样方法KMeans SMOTE算法[14]。在 71 个数据集上的测试结果表明，与其他过采样算法相比，该算法具有性能上的优势。

与过采样不同，欠采样通过减少多数类中

的样本使少数类样本数量与多数类样本数量达到平衡，因此这种采样策略可以避免合成样本带来的噪声。最简单的欠采样方法是随机欠采样，即随机从多数类样本中抽取与少数类数量相同的样本。由于通过随机欠采样方法得到的训练数据可能会缺失多数类样本的某些关键信息，为尽可能充分地利用多数类的数据，有学者提出了基于集成学习思想的EasyEnsemble算法[15]。原始的EasyEnsemble算法使用AdaBoost作为基分类器，该算法与当下流行的几种梯度提升算法相比已经不具有性能优势，但EasyEnsemble算法的采样思想依然在当下的预测模型构建中具有重要的参考意义。

纵观现有的研究成果，在金融企业信贷风险预测的相关研究中，学者们总是将不平衡数据下的特征选择和模型预测的相关问题分开讨论，少有对该问题的解决提出整套方案，也少有对多种不平衡数据采样方法在信贷风险中的应用效果进行探究。基于此，本文使用性能优秀的CatBoost分类器，搭配改进后的AUCRF和EasyEnsemble算法构建了一种组合优化算法，为存在高维特征和不平衡样本的科技情报分析场景提供一种新的算法解决方案。

2　企业金融风险预测方法

本文提出的算法主要有3个步骤：基于AUC的模型评估指标使用随机森林算法进行特征选择；基于EasyEnsemble算法的采样思想，使用Bagging与欠采样结合的采样策略得到若干均衡训练集；基于均衡训练集训练并集成CatBoost分类器。图1为算法的流程示意图。

2.1　基于AUCRF的特征选择算法

随机森林是一种以决策树为基模型的集成学习算法，其中每棵决策树$h(x)$按照以下规则生成：①假设总数据集的样本数量是N，训练集的样本数量为n，则从N随机有放回地抽取n个样本作为该决策树的训练集；②假设数据的特征数量是M，则从M中随机选择m个特征用于决策树的训练。

在得到若干个决策树模型后，通过投票法聚合分类器$H(x)$，如式（1）所示：

$$H(x)=\underset{y\in Y}{argmax}\sum_{t=1}^{T}P(h_t(x)=y) \qquad (1)$$

其中，Y表示数据集中类别标签的集合，$P(\cdot)$为指示函数。

本文算法在文献[16]的基础上进行了改进。首先，根据随机森林在节点分裂时的Gini指数变化量计算出特征的重要性。然后，设定阈值变量，并使阈值从0开始不断增大。每次增大阈值后选出特征重要性大于该阈值的特征组合进行模型训练，计算并记录此时的模型AUC。最后，比

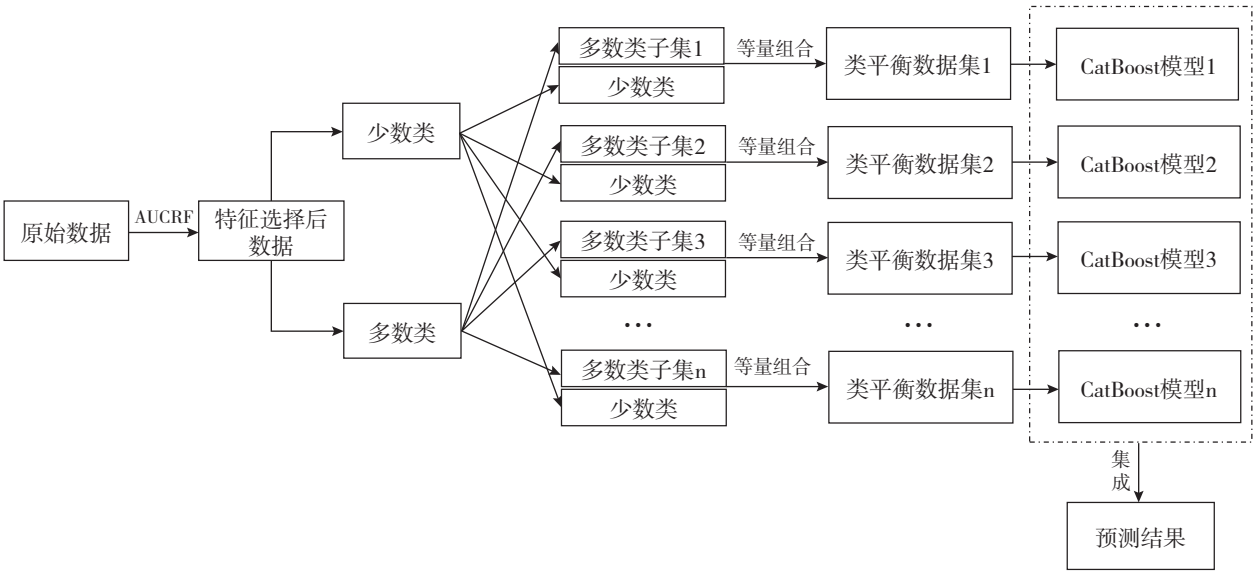

图1　本文提出的算法流程示意

较所有模型的AUC，AUC最大的模型对应的特征组合即为最优特征组合。

2.2　基于Bagging与欠采样组合的采样策略构建均衡训练集

本文算法的采样方法参考了EasyEnsemble算法。在本文算法的采样方法中，对于一个给定的训练数据集，假设少数类样本的数量为M，多数类样本的数量为N，则对N反复进行有放回的随机欠采样，得到T个多数类子集N_1, N_2, N_3, N_4, …, N_t, 其中每个子集的样本数量等于M。再将得到的T个多数类子集分别与少数类样本合并，便可得到T个样本数量为$2M$的样本均衡训练集。这种采样方法尽可能地保留了多数类样本的信息，又避免了合成样本所带来的噪音。

2.3　训练并集成CatBoost基分类器

CatBoost是由俄罗斯公司Yandex在2017年提出的，属于Boosting算法的一种。CatBoost为加法结构，即该强分类器由一系列的弱分类器线性相加组合而成。其结构表达式如式（2）所示：

$$\hat{y}_m = \alpha_0 f_0(x) + \alpha_1 f_1(x) + \ldots + \alpha_m f_m(x) \qquad (2)$$

其中，$f_m(x)$为第m个基分类器，α_m为$f_m(x)$的权重。

目标函数如式（3）所示：

$$L = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (3)$$

其中，$l(\cdot)$是损失函数，代表预测值与真实值的差距；Ω是惩罚项，用于避免过拟合。

相较于Boosting族的其他算法，CatBoost不仅拥有卓越的性能，还可以科学高效地处理类别型特征。CatBoost对类别特征的处理方法名为Target−based Statistics，即通过计算出的数值代替类别特征中的元素。具体做法是在Greedy TS算法中加入先验项优化后，再依靠排序原则（ordering principle）对训练样本进行随机排序并编号，在每次计算时只对排序编号小于该样本的类别标签值进行计算。除此之外，CatBoost算法还会依据贪婪方法（greedy method）将不同类别里的特征进行组合，从而创造出新的特征。

本文算法使用步骤2中得到的T个样本均衡训练集训练CatBoost基分类器，并对CatBoost分类器进行集成，最终的集成结果$H(x)$如式（4）所示：

$$H(x) = \mathrm{sign}\left(\sum_{i=1}^{T} h_i(x) - \sum_{i=1}^{T} \theta_i \right) \qquad (4)$$

其中，$h_i(x)$为第i个CatBoost基分类器，θ_i为阈值。

3　实验及结果分析

3.1　数据来源

本文实验采用了国内某金融机构的客户违约数据集作为模型训练和测试数据集。取该机构2018年全年作为观察期，2019年全年为表现期，剔除缺失值和异常值后共有124 880条样本，其中正样本有12 984个，占总样本的10.40%。数据共有变量51个，其中50个为客户在观察期内的表现和属性，1个为标注了客户在表现期内违约情况的标签。客户是否违约的判断标准按照贷款五级分类制定，除“正常”之外都算作违约。部分变量的中文名称与其含义如表1所示。

3.2　数据变量预处理

对于数值型变量，本文对其进行标准化处理。具体做法是将变量按均值μ进行中心化后，再按标准差σ进行缩放，此时数据就会服从均值为0，方差为1的标准正态分布。标准化公式如式（5）所示：

$$x_i^{-} = \frac{x - \mu}{\sigma} \qquad (5)$$

对于类别型变量，CatBoost可以通过设置自行对其进行处理，但为了便于与其他算法进行对比，本文把二值类变量和类别型变量的各个类别编码为数值。比如将客户贷款产品的按揭情况：按揭、非按揭分别转换为1、0；将客户婚姻状况：未婚、已婚、丧偶、离婚分别转换为1、2、3、4。

3.3　模型评估指标

在不平衡分类问题中，虽然多数类样本的数量要远多于少数类样本，但在进行预测时我们往往更关心的是少数类样本的分类情况。

表 1　金融数据部分变量名称与含义

变量序号	变量名称	变量类型	变量含义
1	性别	类别型	客户性别：男、女、未知
2	年龄	数值型	客户年龄
3	贷款总余额	数值型	观察期末客户的未还清贷款金额
4	出账金额	数值型	客户未还清贷款额度
5	逾期次数	数值型	观察期内客户出现逾期的次数
6	逾期天数	数值型	观察期末时的逾期天数
7	产品按揭分类	二值类	客户贷款产品的按揭情况：按揭、非按揭
…	…	…	…
51	y	标签	是否违约：是 1 否 0

ROC曲线（receiver operating characteristic curve）又叫接收者操作特征曲线，其纵坐标为TPR，即把正例分对的概率；横坐标是FPR，即把负例错分为正例的概率。当正样本属于少数类且负样本属于多数类时，ROC曲线描述的是在模型不断增强找出少数类能力时多数类分错情况的变化趋势。由于不同的ROC曲线间难以直接进行比较，所以一般使用ROC曲线的面积即AUC值来度量模型分类性能，本文也使用AUC作为模型评估指标。

3.4　算法的特征选择结果

算法的特征选择曲线如图 2 所示。该曲线的横坐标为特征重要性阈值，纵坐标为特征选择模型的AUC分数。由图 2 可知，在阈值增大的过程中，模型首先通过去除一些重要性较低的特征提升了预测效果。但随着阈值的不断增大，对提高模型预测性能有帮助的特征逐渐被筛选掉，模型的预测能力也随之下降。根据特征选择曲线最高点对应的 17 个特征制作特征重要性评估图，如图 3 所示。

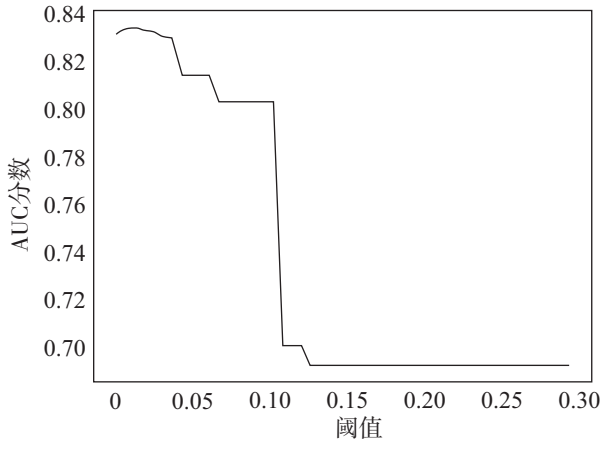

图 2　特征选择曲线

根据图 3 对算法的特征选择结果进行分析，可以发现算法筛选出的特征数量仅为原始特征数量的 34%。这在一定程度上提高了预测模型的泛化能力。筛选出的特征组合包含了客户的历史信用记录、客户的基本情况和产品的基本信息。进一步分析可知：①客户的历史信用记录是判断金融企业信贷风险的重要参考要素。观察期内有违约记录的客户，表现期内依然有较大的可能违约。在历史信用不佳的客户中，逾期天数较长和逾期次数较多的客户应当是重点关注的对象。②客户的基本情况是信贷违约的参考要素之一。一般而言，学历高、资产余额充足、与金融企业业务往来密切的客户不容易违约。③信贷产品的余额、利率、担保方式等因素也会对金融企业信贷业务的风险产生影响。

3.5　多种采样方式下的预测算法实验结果

为探究不同的不平衡数据处理方式与CatBoost算法的组合效果，本文基于最优特征组合，在不同的采样方法下进行对比实验，其结果如表 2 所示。

由表 2 可知：①在 5 种采样方法中，本文选用的采样方法与CatBoost算法的组合效果最好，而在相关领域中最为常用的SMOTE算法与CatBoost算法的组合效果最差。②过采样算法的

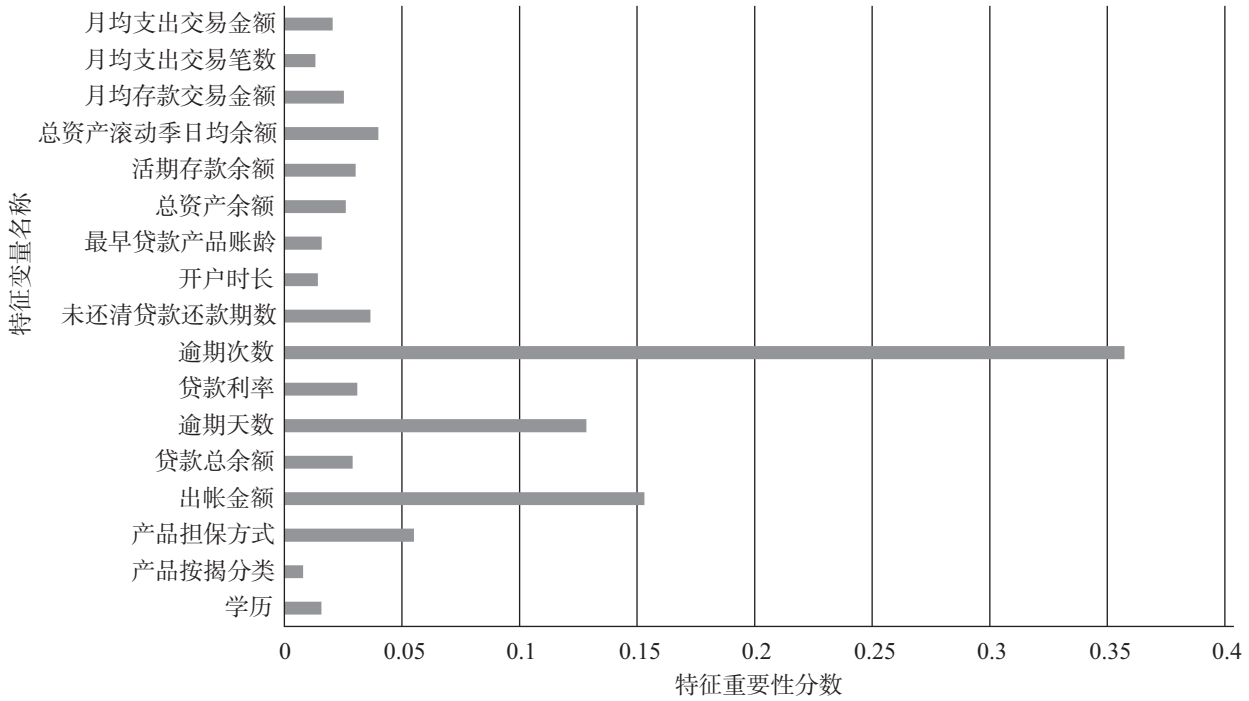

图 3　特征重要性评估

表 2　5 种采样方法的效果对比

算法	采样方式	AUC
SMOTE	过采样	0.823 5
Borderline SMOT	过采样	0.826 4
KMeans SMOTE	过采样	0.837 4
RandomUnderSampler	欠采样	0.839 1
EasyEnsemble	本文方法	0.841 0

整体表现不如欠采样算法，这可能是由于金融企业信贷数据具有一定的复杂性，使用过采样算法合成样本时会产生噪声数据，影响了模型的训练效果。③在欠采样算法中，本文算法比简单欠采样算法的分类效果更好，这是由于本文的采样策略可以通过集成的方式更充分地利用金融企业信贷数据中的信息。

3.6　多种分类算法下的预测算法实验结果

为验证本文算法的信贷风险预测效果，使用已得到的最优特征选择结果，以AUC作为模型评估指标，将CatBoost算法与当前几种较为常用的机器学习算法进行对比实验。同时为了观察EasyEnsemble与普通采样方法相比带来的效果提升，加入了SMOTE算法作为参照。实验结果如表 3 所示。

由表 3 可知，AdaBoost、LightGBM和CatBoost等 3 种集成学习算法的表现要优于单学习器的 3 种机器学习算法，且在 3 种集成学习算法中CatBoost的性能最佳。在所有的实验结果中，本文提出的组合优化算法AUC值最高，相比其他算法而言，在不平衡样本下的金融企业信贷风险预测中具有更好的预测效果。

4　结语

本文针对企业竞争情报分析中的样本不均衡问题，以金融企业信贷风险预测为切入点，针对金融企业信贷业务场景中的样本维数高且严重不均衡现象，提出了一种风险预测方法。该方法首先基于改进后的随机森林算法进行特征选择，再以Bagging与欠采样组合的采样策略构建多个均衡训练集，并在每个均衡训练集中以CatBoost作为基学习器进行训练和集成。为检验该算法的性

表 3　几种组合算法的效果对比

算法	SMOTE	EasyEnsemble
逻辑回归	0.804 8	0.805 1
决策树	0.645 5	0.815 7
SVM	0.799 9	0.800 1
Adaboost	0.820 2	0.833 9
LightGBM	0.821 9	0.836 2
Catboost	0.823 5	0.841 0

能，本文首先进行特征选择，再将提出的算法与常见的几种分类算法进行对比实验。实验结果表明：本文算法筛选出的特征组合具有较强的可解释性，对不平衡样本中的高维特征筛选具有一定的参考价值；相比其他几种算法，本文提出的方法具有最高的AUC值，可以有效地解决金融企业信贷风险的预测问题。

但同时，本文也存在数据和方法两个方面的局限性。在数据的局限性方面，本文选取了金融企业信贷风险这一典型的类别不平衡场景验证本文的方法。至于本文方法在其他的应用场景下表现如何，还需要在后续的研究中做进一步的验证。在方法的局限性方面，本文方法在特征选择的步骤中使用不断增大阈值的方式遴选特征子集，但是这种方式会在一定程度上受到阈值设定的影响。在后续的研究中可以尝试将本文算法与智能优化算法结合，形成性能更佳、耗时更短、自动化程度更高的方法。

参考文献

[1] 王刚, 王珏, 杨善林.电子商务中心基于非均衡数据分类和词性分析的意见挖掘研究[J]. 情报学报, 2014, 33(3): 313−325.

[2] 翁梦娟, 姚长青, 韩红旗, 等.不均衡数据集下基于CNN的中图分类标引方法[J].数据分析与知识发现, 2020, 4(7): 87−95.

[3] 陈钊志, 李冬冬, 王喆, 等.基于下采样的局部判别矩阵型分类的心衰死亡率预测[J].华东理工大学学报(自然科学版), 2019, 45(1): 156−162.

[4] 郭玉萱, 阮春阳, 王晔, 等.基于不平衡数据分类的中药肝毒性检测[J].计算机应用与软件, 2018, 35(8): 226−230.

[5] 路永和, 陈泳珊.基于二进制烟花算法的特征选择方法[J].情报学报, 2017, 36(3): 249−259.

[6] 孙新, 欧阳童, 严西敏, 等.基于训练集裁剪的加权K近邻文本分类算法[J].情报工程, 2016, 2(6): 8−16.

[7] 凌健, 林成德.拆分特征选择及其在企业信用评估中应用[J].福建工程学院学报, 2006, 4(4): 436−439.

[8] 张雷, 王家琪, 费职友, 等. 基于RF−SMOTE−XGboost下的银行用户个人信用风险评估模型[J]. 现代电子技术, 2020, 43(16): 76−81.

[9] CAO J, LU H, WANG W W, et al. A novel five−category loan−risk evaluation model using multiclass ls−svm by pso[J]. International journal of information technology & decision making, 2012, 11(4): 857−874.

[10] 王思宇, 陈建平.基于LightGBM算法的信用风险评估模型研究[J].软件导刊, 2019, 18(10): 19−22.

[11] 马晓君, 宋嫣琦, 常百舒, 等.基于CatBoost算法的P2P违约预测模型应用研究[J].统计与信息论坛, 2020, 35(7): 9−17.

[12] CHAWL N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over−sampling technique [J]. Journal of artificial intelligence research, 2002, 16(1): 321−357.

[13] HAN H, WANG W Y, MAO B H. Borderline−SMOTE: a new over−sampling method in imbalanced data sets learning[C]//International Conference on Advances in Intelligent Computing. Berlin: Heidelberg, 2005: 878−887.

[14] GEORGIOS D, FERNANDO B, FELIX L. Improving imbalanced learning through a heuristic oversampling method based on k−means and SMOTE[J]. Information sciences, 2018, 465: 1−20.

[15] LIN X Y , WU J , ZHOU Z H . Exploratory under sampling for class−imbalance learning[J]. IEEE transactions on systems man & cybernetics part B, 2009, 39(2): 539−550.

[16] 刘忻梅, 唐俊, 段翀.AUCRF算法在信用风险评价中的特征选择研究[J]. 计算机应用与软件, 2018, 35(4): 293−295, 309.