**Computer Science and Application** 计算机科学与应用, 2023, 13(5), 1135-1147
Published Online May 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.135111

Hans汉斯

# 基于SMOTE-Tomek与AdaBoost相结合的不平衡分类算法在金融信贷领域的研究

**马　宁，刘　硕，王乐秀**

中国石油大学(北京)理学院，北京

收稿日期：2023年4月27日；录用日期：2023年5月24日；发布日期：2023年5月31日

## 摘　要

**在互联网金融快速发展的时代，信贷风险成为目前金融领域急需解决的问题之一。而信贷风险评估模型作为一种有效的工具，可以利用客户信息和客户活动数据识别潜在的风险，在金融机构中发挥着至关重要的作用。本文针对Kaggle数据集中因逾期还款用户实例远少于正常还款用户实例而造成的样本高度不平衡问题，以信贷风险预测为切入点，提出一种面向不平衡样本的风险识别方法。该方法选定以决策树为基分类器的AdaBoost分类器来训练SMOTE-Tomek平衡过后的数据集，它通过一种迭代机制让原本性能不强的分类器组合起来，形成一个强分类器。并选用精确率、召回率、ROC曲线及AUC值来评价所选定分类器的分类效果。实验结果表明，AdaBoost分类器相对于决策树、支持向量机和朴素贝叶斯分类器在信贷客户的风险评估中表现最优。**

## 关键词

**信贷风险评估模型，样本不平衡，SMOTE-Tomek，AdaBoost**

# Research on Imbalanced Classification Algorithm Based on the Combination of SMOTE-Tomek and AdaBoost in the Field of Financial Credit

**Ning Ma, Shuo Liu, Lexiu Wang**

College of Science, China University of Petroleum (Beijing), Beijing

Received: Apr. 27th, 2023; accepted: May 24th, 2023; published: May 31st, 2023

文章引用: 马宁, 刘硕, 王乐秀. 基于 SMOTE-Tomek 与 AdaBoost 相结合的不平衡分类算法在金融信贷领域的研究[J]. 计算机科学与应用, 2023, 13(5): 1135-1147. DOI: 10.12677/csa.2023.135111

## Abstract

**In the era of rapid development of internet finance, credit risk has become one of the most urgent problems to be solved in the financial field. As an effective tool, credit risk assessment model can identify potential risks by using customer information and customer activity data, and play a vital role in financial institutions. In this paper, we take credit risk prediction as a starting point to solve the problem of high sample imbalance in Kaggle data set, which is caused by the fact that the number of overdue users is far less than that of normal users, a risk identification method for unbalanced samples is proposed. This method selects the AdaBoost classifier based on the decision tree to train the SMOTE-Tomek balanced data set. It uses an iterative mechanism to combine the weak classifiers to form a strong classifier. The accuracy rate, recall rate, ROC curve and AUC value are selected to evaluate the classification effect of the selected classifier. Experimental results show that the AdaBoost classifier performs best in the risk assessment of credit customers compared with decision trees, support vector machine and Naive Bayes classifier.**

## Keywords

**Credit Risk Assessment Model, Sample Imbalance, SMOTE-Tomek, AdaBoost**



## 1. 引言

伴随着中国经济和金融行业的快速发展，互联网技术被广泛应用到金融领域中，互联网金融成为了金融业的重要组成部分，信贷就是一项重要的创新成果。由于其灵活、便捷的融资方式，信贷[1]成为了越来越多人选择的融资渠道，但同时信用风险问题也一直制约着信贷平台的发展，较高的违约率带来了极大的负面影响。因此做好贷款违约预测有助于银行相关业务的顺利展开，通过数据挖掘技术对数据库中的全体借贷人以及信贷申请人的相关信息数据进行挖掘分析和分类，能够预测申请人是否会违约，从而决定是否放贷。

近年来，许多基于机器学习的方法被广泛应用于信用风险预测模型中，其中包括逻辑回归[2]、神经网络[3]和支持向量机[4] [5]等方法。国内外众多学者对这些方法的可行性进行了验证，但是在研究信用风险评估模型的实际问题中，因为逾期用户数量相对于正常还款用户的数量很少而造成了正负样本比例极不平衡。利用不平衡的数据集进行模型训练，将严重影响模型的分类性能，得到较差的预测效果。对于解决不平衡数据集的分类问题，一般有数据不平衡处理方法和分类器两方面内容。

### 1.1. 数据不平衡处理方法

比较常见的处理数据集不平衡的方法有欠采样和过采样方法[6] [7]，通过随机移除多数类样本或者复制少数类样本达到平衡类别分布的目的。陈启伟[8]等利用欠采样方法平衡数据集并与引入参数扰动的集成学习方法相结合建立信用评分模型。但是该方法对于正负样本比例失衡比较严重的数据集来说，分类效果仍有待提高。Niu [9]等利用 SMOTE 方法处理不平衡数据集，验证了该方法在信用风险评估模型中的有效性。但是 SMOTE 方法在生成新样本的过程中没有对少数类样本进行区别选择，并且容易出现样

本重叠的问题。对此 Han [10]等提出了边界合成少数类过采样 Borderline-SMOTE 算法，改善了样本重叠的问题，但该方法只对处于边界的少数类样本进行过采样，容易造成正负类边界模糊的问题。

### 1.2. 分类器

传统分类算法包括朴素贝叶斯算法[11]、支持向量机算法和决策树算法[12]等。朴素贝叶斯算法是基于条件独立性假设的一种算法，当条件独立性假设成立时，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。支持向量机算法是建立在统计学习理论基础上的机器学习方法，SVM 可以自动寻找出对分类有较好区分能力的支持向量，由此构造出的分类器可以最大化类与类的间隔，因而有较好的适应能力和较高的分准率。决策树算法是一种逼近离散函数值的方法。它是一种典型的分类方法，首先对数据进行处理，利用归纳算法生成可读的规则和决策树，然后使用决策对新数据进行分析。

但由于传统分类算法在解决不平衡数据的分类问题时存在局限性，为此可以在传统分类算法上做出改进，主要方法有代价敏感学习[13]和集成学习方法[8] [14] [15]。代价敏感学习通过优化目标函数使分类模型更关注少数类样本的分类准确率解，从而增加少数类样本错分的惩罚代价来解决数据不平衡的问题。集成学习算法是一种组合算法，它通过组合多个基分类器输出最终分类效果较好的强分类器。经典的集成学习算法包括 Boosting 算法[12]，随机森林算法[16]和 Bagging 算法[17] [18]。Bagging 中典型的一种算法是 AdaBoost 算法[14] [19] [20] [21]，该算法以每一次迭代训练出的基分类器为基础，对错分样本加大权重，之后再进行迭代训练。在不平衡的数据集中，少数类样本相对而言被错分的代价较大，可以在很大程度上增强少数类样本的影响程度而有效的提高少数类样本的分类效果，从而训练出更加偏向于少数类分类正确率的模型。

根据上述分析，本文提出一种基于 SMOTE (Synthetic Minority Oversampling Technique)的改进算法 SMOTE-Tomek 算法和 AdaBoost 分类算法相结合的信用风险预测模型来改善数据不平衡问题对分类效果的影响。该模型从数据不平衡处理方法和分类器两个方面进行改进，来解决信用风险预测中数据不平衡的问题。对于数据不平衡处理方法，利用改进的过采样方法生成新样本来平衡数据集；在分类器方面，利用 AdaBoost 分类算法训练新的数据集得到最终的预测模型。

## 2. 不平衡数据处理方法

### 2.1. SMOTE 算法

SMOTE 的全称是“合成少数类过采样技术”，非直接对少数类进行重采样，而是设计算法来人工合成一些新的少数样本。通过创造合成的少数类样本来实现对少数类的过采样的方法是对随机过采样技术的改进，可以在一定程度上避免随机过采样模型过度拟合的问题。但是 SMOTE 算法在生成新的少数类样本时，只是单一地在同类近邻样本间插值，并没有考虑到少数类样本附近的多数类样本分布情况。若新生成的少数类样本周围有多数类样本，则很容易发生重叠的现象，使样本分类时发生错误。

### 2.2. SMOTE-Tomek

SMOTE-Tomek 算法使用 SMOTE 对样本中少数类样本进行上采样，然后使用 Tomek Links 方法对多数类样本进行下采样。SMOTE-Tomek 方法可以一定程度上缓解 SMOTE 方法容易产生样本重叠的问题，Tomek Links 流程如下：

1) 将不平衡数据集分为多类样本数据集 $D_{max}$ 和少数类样本数据集 $D_{min}$。

2) 对多类样本数据集 $D_{max}$ 中的每一个多类样本求其最近的少数类样本，对少数类样本数据集 $D_{min}$ 中

的每一个少数类样本求其最近的多数类样本。

3) 比较其最近距离，并判断样本是否相同，若相同则为 Tomek Links 对。

4) 将 Tomek Links 对中的多数类删除，从而得到新的数据集，然后选择分类算法对其学习。

## 3. AdaBoost 分类器

AdaBoost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器(基分类器)，然后把这些基分类器集合起来，构成一个更强的最终分类器(强分类器)。其算法本身是通过改变数据分布来实现的，它根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练，最后将每次训练得到的分类器最后融合起来，作为最后的决策分类器。AdaBoost 能对不平衡数据集获得更好的通用性能，AdaBoost 的提升机制更专注于被误分的少样本，以提高少类样本的分类精度。

关于 AdaBoost 的基分类器，目前有多种选择，如以径向基核支持向量机(RBFSVM)和决策树为基分类器。本文使用决策树模型作为基分类器，以决策树为基函数的提升方法称为提升树，在分类问题中此决策树为二分类决策树。

以决策树为基分类器的 AdaBoost 算法流程如下：

1) 初始化训练数据(每个样本)的权值分布。每一个训练样本，初始化时赋予同样的权值 $w=\frac{1}{N}$，$N$ 为样本总数。

$$D_1=\left(w_{11},w_{12},\cdots,w_{1i},\cdots,w_{1N}\right),w_{1i}=\frac{1}{N},i=1,2,\cdots,N$$

$D_1$ 表示第一次迭代每个样本的权值。$w_{11}$ 表示第 1 次迭代时的第一个样本的权值。

2) 进行多次迭代，$m=1,2,\cdots,M$，$m$ 表示迭代次数。

a) 模型采用前向分布算法，首先确定初始提升树 $G_0\left(x\right)=0$。

b) 使用具有权值分布 $D_m\left(m=1,2,3,\cdots,N\right)$ 的训练样本集进行学习，得到基分类器：

$$G_m\left(x\right)=\sum_{m=1}^{M}T\left(x;\Theta_m\right)$$

$G_m\left(x\right):\chi\to\left\{-1,+1\right\}$，该式子表示第 $m$ 次迭代时的基分类器，将样本 $x$ 要么分类成−1，要么分类成 1。其中 $T\left(x;\Theta_m\right)$ 表示决策树，$\Theta_m$ 表示决策树参数，$M$ 为树的个数。

c) 计算基分类器 $G_m$ 的训练误差。

$$e_m=\sum_{i=1}^{N}w_i^m$$

d) 计算基分类器 $G_m\left(x\right)$ 的话语权，话语权 $\alpha_m$ 表示 $G_m\left(x\right)$ 在最终分类器中的重要程度。

$$\alpha_m=\frac{1}{2}\log\frac{1-e_m}{e_m}$$

该式是随 $e_m$ 减小而增大，即误差率小的分类器，在最终分类器的重要程度大。

e) 通过经验风险极小化确定下一颗决策树参数 $\Theta_m$：

$$\hat{\Theta}=\arg\min_{\Theta_m}\sum_{i=1}^{N}L\left(y_i,G_{m-1}\left(x\right)+T\left(x_i;\Theta_m\right)\right)$$

f) 更新训练样本集的权值分布用于下一轮迭代，其中被误分的样本的权值会增大，被正确分的权值

减小。

$$D_{m+1}=\left(w_{m+1,1},w_{m+1,2},\cdots,w_{m+1,i},\cdots,w_{m+1,N}\right)$$

$$w_{m+1,i}=\frac{w_{mi}}{Z_m}\exp\left(-\alpha_m y_i G_m\left(x_i\right)\right),i=1,2,\cdots,N$$

$D_{m+1}$是用于下次迭代时样本的权值，$w_{m+1,i}$是下一次迭代时第 $i$ 个样本的权值。其中 $y_i$ 代表第 $i$ 个样本对应的类别(1 或−1)，$G_m\left(x_i\right)$ 表示基分类器对样本 $x_i$ 的分类(1 或−1)。若分类正确，$y_i G_m\left(x_i\right)$ 的值为 1，反之为−1。其中 $Z_m$ 是归一化因子，使得所有样本对应的权值之和为 1。

$$Z_m=\sum_{i=1}^{N}w_{mi}\exp\left(-\alpha_m y_i G_m\left(x_i\right)\right)$$

3) 迭代完成后，组合基分类器，形成强分类器。

$$G\left(x\right)=sign\left(\sum_{m=1}^{M}\alpha_m G_m\left(x\right)\right)$$

## 4. 模型性能评估指标

为了正确评估是否有效分类样本，引入模型性能评估指标。本文所构建的为二分类模型，采用的评估指标有混淆矩阵、精确度、召回率、F1-score 以及 ROC 曲线和 AUC 值。

### 4.1. 混淆矩阵

在本文二分类的情况下，表 1 混淆矩阵中定义了四个类别：真阴性(TN)代表借款人实际未违约，预测结果也未违约；伪阳性(FP)代表借款人实际未违约，预测却违约了；真阳性(TP)代表借款人实际违约，预测结果也违约；伪阴性(FN)代表借款人实际违约，但预测结果却为未违约。

**Table 1.** Confusion matrix
**表 1.** 混淆矩阵

| | 预测正类 | 预测负类 |
|---|---|---|
| 真实正类 | TP | FN |
| 真实负类 | FP | TN |

### 4.2. 精确度、召回率、F1-Score

精确度是指真阳性(TP)的数量在模型所预测的观测值为阳性的样本中(TP + FP)所占的比例。公式定义为：

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}}$$

混淆矩阵中的借款人违约的人数中所占的比例。公式定义为：

$$R=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

F-measure 也称为 F-score，是召回率 $R$ 和精度 $P$ 的加权调和平均值，是为调和召回率的增减和精度之间的矛盾。该综合评价指标 $F$ 引入一个系数 $\alpha$，进行召回率和精度的加权调和，其表达式如下：

$$F=\left(\alpha^2+1\right)P\cdot R\Big/R\alpha^2\left(P+R\right)$$

最常用的 $F1$ 指标是上述式中的系数 $\alpha$ 为 1 的情况，即：

$$F1 = 2PR/(P+R)$$

$F1$ 的最大值为 1，最小值为 0。

### 4.3. ROC 曲线及 AUC 值

ROC 曲线全称为受试者工作特征曲线(Receiver Operating Characteristic Curve)，它是根据一系列不同的二分类方式(分界值或决定阈)，它是以假正率 FPR 为横坐标，真正率 TPR 为纵坐标而得出的 ROC 曲线图。

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

$$\mathrm{FPR} = \frac{\mathrm{TP}}{\mathrm{TN}+\mathrm{FP}}$$

AUC 就是 ROC 曲线下面积，在比较不同的分类模型时，可以将每个模型的 ROC 曲线都画出来，比较曲线下面积做为模型优劣的指标。

## 5. 数据介绍与模型建立

### 5.1. 数据介绍

银行等金融机出于客户隐私和信息安全等方面的考虑，公开提供的贷款违约或信用评分有关的数据集非常少，较大规模的数据集就更少了。本文使用的贷款违约数据集来自 Kaggle 网站发布的 2007~2010 年贷款违约竞赛数据，竞赛组织者要求参赛者对借款人是否全额偿还贷款进行分类和预测。

因此本文仅使用 Kaggle 网站提供的训练集中 9578 个样本作为本文的数据集(以下简称 Kaggle 数据集)，用于训练本文所提及算法模型并测试该模型的预测性能。Kaggle 数据集包括了借款人的借款目的、贷款利率、自我报告年收入、借款人 FICO 信用评分数等 12 个变量，表 2 列出了变量名及各变量描述：

**Table 2.** The Kaggle dataset variable case
**表 2.** Kaggle 数据集变量情况

| 变量名 | 变量描述 |
|---|---|
| Credit.Policy | 客户符合 LendingClub.com 的信用承保标准则为 1，否则为 0 |
| purpose | 贷款目的 |
| int.rate | 贷款的利率 |
| installment | 如果贷款已到位，借款人所欠的每月分期付款 |
| log.annual.inc | 借款人自我报告年收入的自然对数 |
| dti | 借款人的债务收入比率 |
| FICO | 借款人的 FICO 信用评分 |
| days.with.cr.line | 借款人拥有信用额度的天数 |
| revol.bal | 信用卡结算周期结束时未支付的金额) |
| revol.util | 借款人的循环额度利用率 |
| inq.last.6mths | 借款人在过去 6 个月内债权人的查询次数 |
| delinq.2yrs | 借款人在过去 30 年中逾期 2+天的次数 |
| pub.rec | 借款人的贬损性公共记录 |

根据对上表中变量的可视化分析可知以下几个规律：

一是信用评分特别低和特别高的人占比都较少，大多数信用评分中等，大体呈现为右偏态的正态分布，并且信用评分分值越高，违约率越低，如图 1。这是信用评分的核心价值所在，可以根据信用评分的高低进行诸如是否发放、贷款额度、是否需要抵押等重要决策。

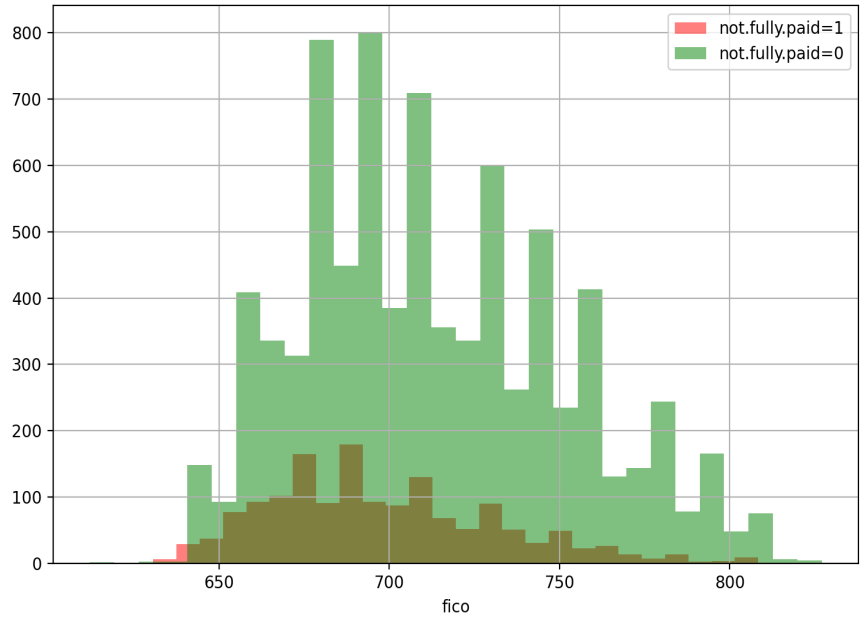


**Figure 1.** Histograms of loan repayment in full under different FICO scores
**图 1.** 不同 FICO 分数下是否全额偿还贷款分布直方图

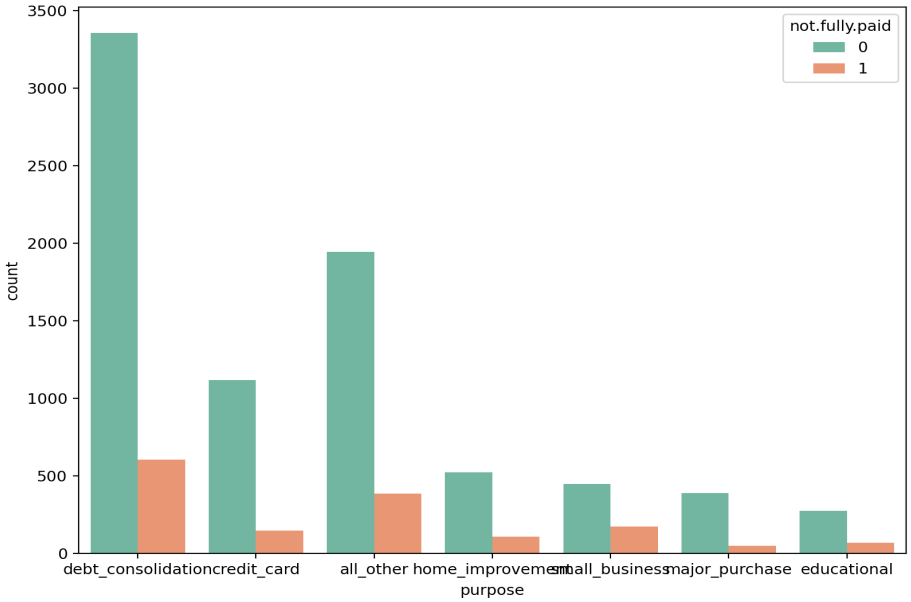


**Figure 2.** The histogram of the distribution of whether or not to repay loans in full for different purposes
**图 2.** 不同贷款目的时是否全额偿还贷款分布直方图

二是用户未能全额偿还贷款的原因有众多，其中债务合并贷款在众多贷款目的中占主要因素，如图 2。

债务合并贷款(Debt Consolidation Loans)是为那些被众多债务纠缠而找不到解决办法的人准备的，它可以将债务人多项债捆绑起来负，组合为一笔新贷款统一偿还。

三是根据图 3 所示，not.fully.pay 和 credit.policy 的趋势基本一致，无论是否能够全额偿还贷款，客户不符合 LendingClub.com 信用承保标准的贷款利率基本高于符合 LendingClub.com 信用承保标准的客户贷款利率，且不符合信用承保标准客户的 FICO 得分低于符合标准的客户的得分。

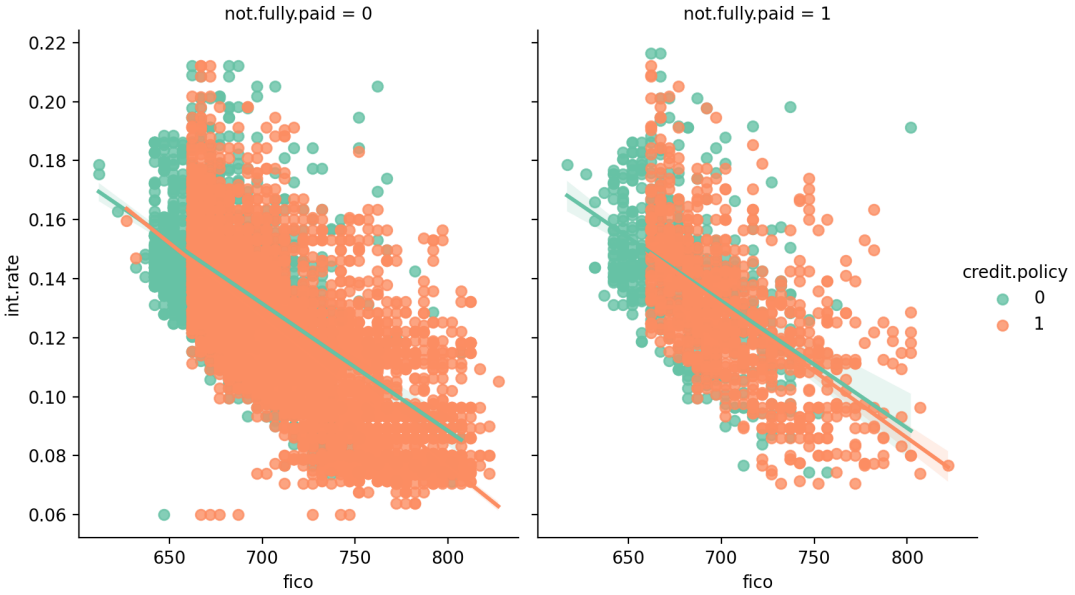


**Figure 3.** Trend charts not.fully.pay and credit. Policy
**图 3.** not.fully.pay 和 credit.policy 的趋势图

## 5.2. 模型建立

本文建立模型首先对数据进行预处理，建立训练集与测试集，然后通过 SMOTE 及其改进方法平衡原训练集，再利用改进的 AdaBoost 分类算法在新的训练集上根据筛选得到的变量特征进行训练。建模流程如图 4 所示，具体实现过程如下：

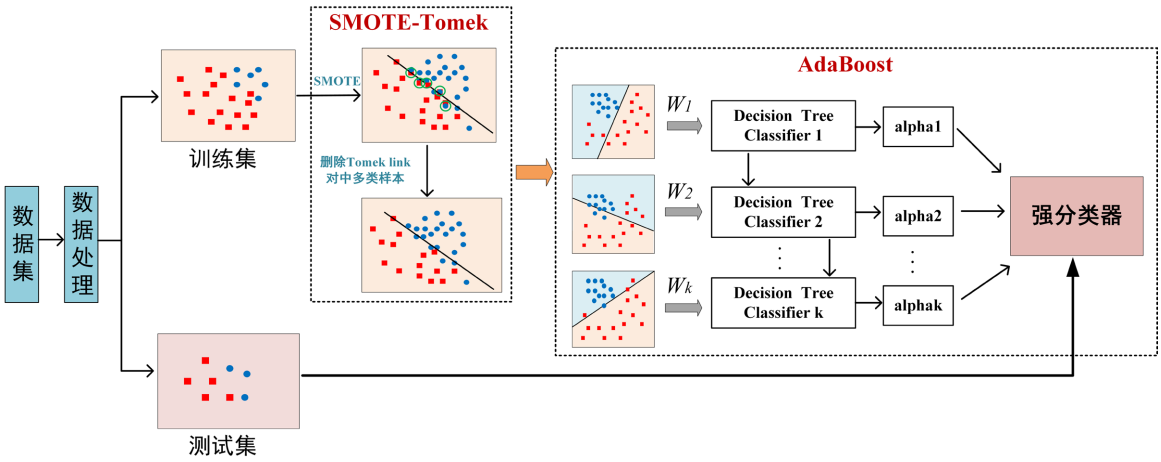


**Figure4.** Algorithm flowchart
**图 4.** 算法流程图

1) 本文首先对数据进行预处理，查看数据集中是否存在异常值与缺失值，并对“purpose”这一变量进行 One-hot 编码，将此变量变成稀疏变量，不仅仅解决了分类器对属性数据不好处理的问题，也在一定程度上起到了扩充特征的作用。将稀疏之后的“purpose”变量和“int.rate”“installment”“log.annual.inc”“dti”“fico”等变量作为特征变量。

2) 对数据进行训练集与测试集的划分，并对选择的特征变量中数据进行标准化与归一化。

3) 对处理过后的训练集进行 SMOTE 及其改进方法采样，生成新的平衡训练集。

4) 利用 AdaBoost 分类算法在新的训练集上根据选择的特征进行训练，最终建立分类模型，其中 AdaBoost 算法中的基分类器限制为二分类决策树。

5) 用测试集数据进行测试，验证模型分类效果。

## 6. 结果分析

本部分主要分析 SMOTE 改进算法与 AdaBoost 算法相结合模型的分类性能。使用上述数据，将 SMOTE 改进算法与支持向量机分类器、决策树分类器以及朴素贝叶斯分类器和以决策树为基分类器的 AdaBoost 分类器分别相结合，并对性能评估指标进行对比，结果如表 3 所示。

**Table 3.** Evaluation mechanism and performance comparison
**表 3.** 评估机制与性能比较

| Method | | PPV | Recall | F1-score | AUC |
|---|---|---|---|---|---|
| SVM | SMOTE | 0.6934 | 0.7358 | 0.7935 | 0.66772 |
| | Border Line | 0.6132 | 0.5978 | 0.7122 | 0.70235 |
| | ADASYN | 0.8445 | 0.9472 | 0.9070 | 0.77380 |
| | SMOTEENN | 0.7044 | 0.7625 | 0.8051 | 0.60782 |
| | SMOTE-Tomek | 0.7072 | 0.7268 | 0.7989 | 0.73840 |
| DT | SMOTE | 0.9830 | 0.9856 | 0.9893 | 0.98576 |
| | Border Line | 0.9808 | 0.9828 | 0.9879 | 0.98677 |
| | ADASYN | 0.9786 | 0.9822 | 0.9866 | 0.98621 |
| | SMOTEENN | 0.9665 | 0.9643 | 0.9788 | 0.97599 |
| | SMOTE-Tomek | 0.9797 | 0.9842 | 0.9873 | 0.97653 |
| Bayes | SMOTE | 0.7418 | 0.8332 | 0.8378 | 0.69577 |
| | Border Line | 0.7451 | 0.8380 | 0.8403 | 0.69175 |
| | ADASYN | 0.7429 | 0.8380 | 0.8392 | 0.68488 |
| | SMOTEENN | 0.6868 | 0.7199 | 0.7864 | 0.70178 |
| | SMOTE-Tomek | 0.7407 | 0.8312 | 0.8369 | 0.69547 |
| AdaBoost | SMOTE | 0.9868 | 0.9931 | 0.9918 | 0.98714 |
| | Border Line | 0.9879 | 0.9931 | 0.9925 | 0.98726 |
| | ADASYN | 0.9885 | 0.9945 | 0.9928 | 0.98506 |
| | SMOTEENN | 0.9775 | 0.9808 | 0.9859 | 0.98658 |
| | SMOTE-Tomek | 0.9879 | 0.9945 | 0.9925 | 0.99304 |

由图 5 可知，以决策树为基分类器的 AdaBoost 分类器相比于支持向量机分类器、决策树分类器和朴素贝叶斯分类器的分类效果更好，并且由图 6 可知在模型参数一致的前提下，AdaBoost 与 SMOTE-Tomek

数据处理方法相结合的模型分类效果优于与数据处理方法相结合的模型分类效果。因此对于上述实例数据，本文采用 AdaBoost 分类器与 SMOTE-Tomek 数据处理方法来进行分类判断。

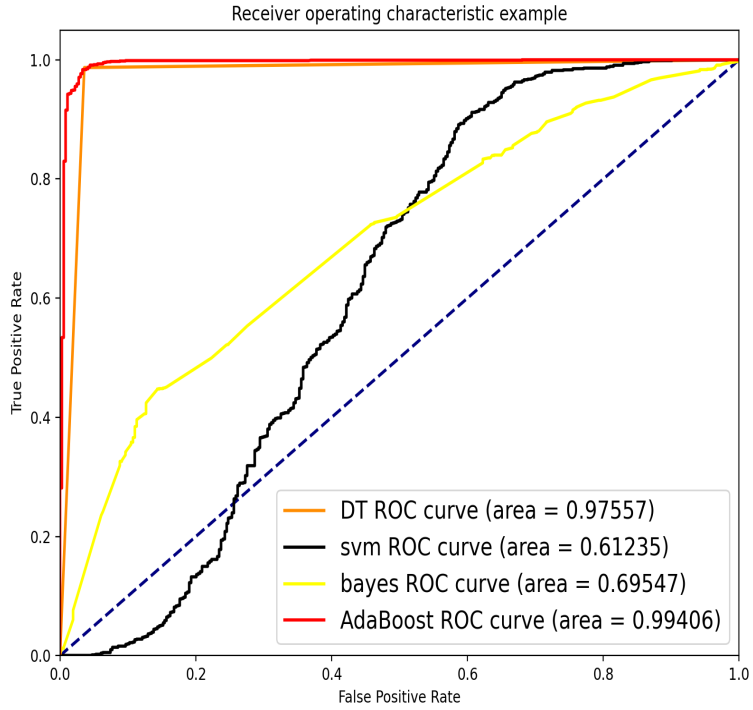


**Figure 5.** Roc curves of different classifiers
**图 5.** 不同分类器的 ROC 曲线图

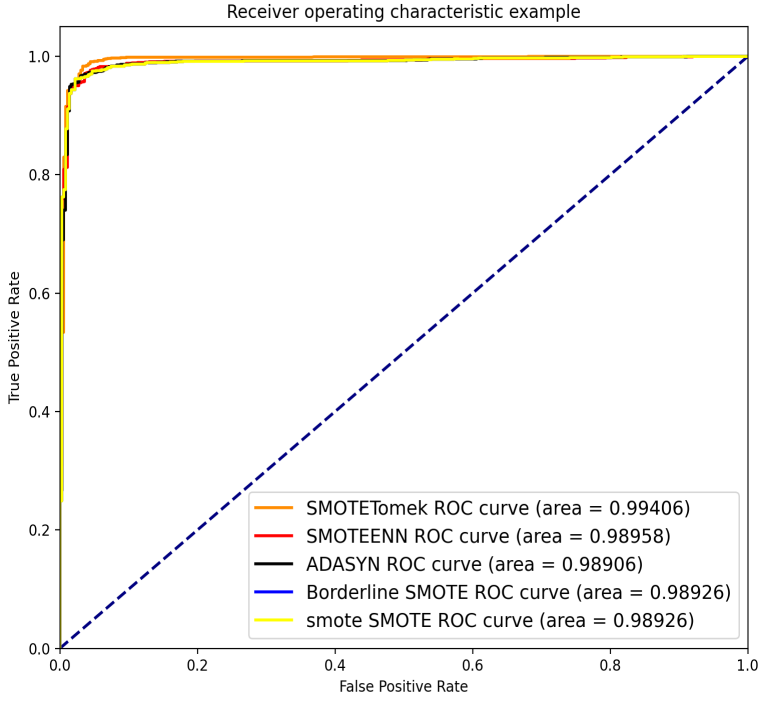


**Figure 6.** Different sampling methods combined ROC curves
**图 6.** 不同数据处理方法与 AdaBoost 结合 ROC 曲线图

对于以决策树为基分类器的 AdaBoost 模型中，基分类器的个数(即基分类器的最大迭代次数)对于模型的最终分类效果有重要影响，如果迭代次数太大，容易过拟合；迭代次数太小，容易欠拟合。根据对模型中迭代次数的参数调整，在本文中当基分类器的迭代次数达到 400 时，分类效果最好，而当次数达到 500 时会出现过拟合的情况，如图 7。

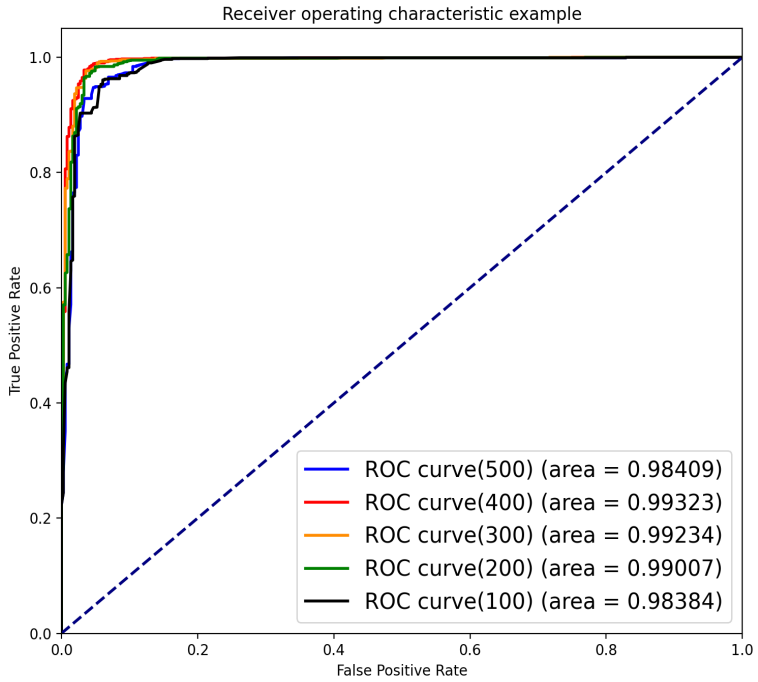


**Figure 7.** The base classifier has different iterations
**图 7.** 基分类器不同迭代次数

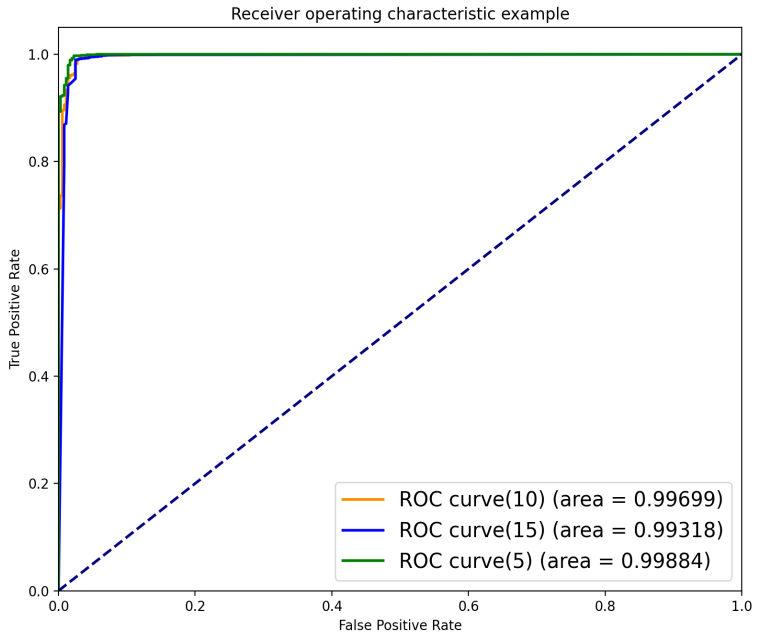


**Figure 8.** Different decision tree depths
**图 8.** 不同决策树深度

为了探索决策树的大小和模型分类效果之间的影响和联系，方便起见，调用 sklearn 中的决策树分类工具包，设置决策树最大深度这一参数(max_depth)分别为 5，10，15 进行实验，其他参数相同。由图 8 可知决策树最大深度为 5 时，其 AUC 值可达 0.99884，此时模型结果达到最优，随着决策树增大而效果变差，可能是出现了对训练集过拟合的情况。

## 7. 结论

信用风险问题一直制约着信贷平台的发展，一个有效的信用风险预测模型是研究的重点。在实际的研究中数据集不平衡问题严重影响着模型分类效果，为此本文从不平衡数据处理方法和分类器算法两个方面提出改进方法。在不平衡数据处理方面，通过 SMOTE 过采样及其改进方法平衡数据集；在分类器方面，利用决策树为基分类器，提出 AdaBoost 分类算法。通过与其他方法的对比实验，证实了 SMOTE-Tomek 与 AdaBoost 相结合的分类模型在信用风险预测中具有更好的预测效果。但是本文提出的模型仍然需要进一步改进，在未来可尝试寻找其他不平衡数据处理方法和分类器的结合，期望进一步提高分类效果。

## 基金项目

中国石油大学(北京)油气资源与探测国家实验室课题资助(PRP/DX-2208)。

## 参考文献

[1] 刘文雅. 基于数据挖掘的银行客户信贷违约的研究[D]: [硕士学位论文]. 大连: 大连理工大学, 2020.

[2] 逯瑶瑶. 基于机器学习分类算法的贷款违约预测研究[D]: [硕士学位论文]. 兰州: 兰州大学, 2021.

[3] 王浩, 唐桥虹, 唐娜, 等. 基于神经网络的心电分类算法抗扰性影响分析[J]. 中国医疗设备, 2023, 38(3): 61-65.

[4] 汪海燕, 黎建辉, 杨风雷. 支持向量机理论及算法研究综述[J]. 计算机应用研究, 2014, 31(5): 1281-1286.

[5] 沈翠华, 刘广利, 邓乃扬. 一种改进的支持向量分类方法及其应用[J]. 计算机工程, 2005, 31(8): 153-154.

[6] 王雅婷. 基于不平衡数据的多种采样方法的信用评分模型研究[D]: [硕士学位论文]. 南昌: 江西财经大学, 2022.

[7] 汪海涛, 余永奎, 段春雨. 基于大数据不平衡样本集的重采样方法及应用[J]. 现代计算机(专业版), 2018(22): 26-29.

[8] 陈启伟, 王伟, 马迪, 毛伟. 基于 Ext-GBDT 集成的类别不平衡信用评分模型[J]. 计算机应用研究, 2018, 35(2): 421-427.

[9] Niu, J., Liu, Z., Pan, Q., Yang, Y. and Li, Y. (2023) Conditional Self-Attention Generative Adversarial Network with Differential Evolution Algorithm for Imbalanced Data Classification. *Chinese Journal of Aeronautics*, **36**, 303-315. https://doi.org/10.1016/j.cja.2022.09.014

[10] HAN, H., WANG, W.Y., MAO, B.H. (2005) Border-line-SMOTE: A New Over-Sampling Method in Iimbalanced Data Sets Learning. In: Huang, D.S., Zhang, X.P. and Huang, G.B., eds., *ICIC* 2005: *Advances in Intelligent Computing*, Springer, Berlin, 878-887. https://doi.org/10.1007/11538059_91

[11] 郭秀娟, 李庆凯, 孟庆楠, 马玉贤. 基于朴素贝叶斯算法分析鸢尾花数据集分类[J]. 工业和信息化教育, 2022(6): 82-84.

[12] 王植, 张珏. 基于 Boosting 算法的 C5.0 决策树不平衡数据分类算法[J]. 河南科学, 2023, 41(1): 7-12.

[13] 孙艳歌, 邵罕, 杨艳聪. 基于代价敏感不平衡数据流分类算法[J]. 信阳师范学院学报(自然科学版), 2019, 32(4): 670-674.

[14] 韩慧敏, 沈润平, 黄安奇, 狄文丽. 基于集成学习方法的 CLDAS 土壤湿度降尺度研究[J]. 南京信息工程大学学报(自然科学版), 2021, 13(6): 693-706.

[15] 李小娟, 韩萌, 王乐, 等. 监督与半监督学习下的数据流集成分类综述[J]. 计算机应用研究, 2021, 38(7): 1921-1929.

[16] 沈智勇, 苏翀, 周扬, 沈智威. 一种面向非均衡分类的随机森林算法[J]. 计算机与现代化, 2018, 280(12): 56-60.

[17] 肖梁, 韩璐, 魏鹏飞, 等. 基于 Bagging 集成学习的多集类不平衡学习[J]. 计算机技术与发展, 2021, 31(10): 1-6.

[18] 王乐, 韩萌, 李小娟, 等. 不平衡数据集分类方法综述[J]. 计算机工程与应用, 2021, 57(22): 42-52.

[19] 王燕. 基于 Adaboost 算法的人脸图像情绪识别[J]. 杨凌职业技术学院学报, 2023, 22(1): 10-13.

[20] 左海超, 孙媛媛, 杨晓东, 徐涛. 基于迁移 AdaBoost 的航线节假日客流量预测[C]//中国公路学会. 2022 世界交通运输大会(WTC2022)论文集(交通工程与航空运输篇). 北京: 人民交通出版社, 2022: 8.

[21] 钱揖丽, 冯志茹. 利用 AdaBoost-SVM 集成算法和语块信息的韵律短语识别[J]. 计算机工程与科学, 2015, 37(12): 2324-2330.