



文章编号:1001-9081(2019)06-1707-06

DOI:10.11772/j.issn.1001-9081.2018102180

## 基于带多数类权重的少数类过采样技术和随机森林的信用评估方法

田 臣\*, 周丽娟

(首都师范大学 信息工程学院, 北京 100048)

(\* 通信作者电子邮箱 1561422137@qq.com)

**摘要:**针对信用评估中最为常见的不均衡数据集问题以及单个分类器在不平衡数据上分类效果有限的问题,提出了一种基于带多数类权重的少数类过采样技术和随机森林(MWMOTE-RF)结合的信用评估方法。首先,在数据预处理过程中利用MWMOTE技术增加少数类别样本的样本数;然后,在预处理后的较平衡的新数据集上利用监督式机器学习算法中的随机森林算法对数据进行分类预测。使用受测者工作特征曲线下面积(AUC)作为分类评价指标,在UCI机器学习数据库中的德国信用卡数据集和某公司的汽车违约贷款数据集上的仿真实验表明,在相同数据集上,MWMOTE-RF方法与随机森林方法和朴素贝叶斯方法相比,AUC值分别提高了18%和20%。与此同时,随机森林方法分别与合成少数类过采样技术(SMOTE)方法和自适应综合过采样(ADASYN)方法结合,MWMOTE-RF方法与它们相比,AUC值分别提高了1.47%和2.34%,从而验证了所提方法的有效性及其对分类器性能的优化。

**关键词:**不均衡数据集;机器学习;带多数类权重的少数类过采样技术;随机森林;信用评估

中图分类号: TP18; TP399 文献标志码:A

### Credit assessment method based on majority weight minority oversampling technique and random forest

TIAN Chen\*, ZHOU Lijuan

(Information Engineering College, Capital Normal University, Beijing 100048, China)

**Abstract:** In order to solve the problem of unbalanced dataset in credit assessment and the limited classification effect of single classifier on unbalanced data, a Majority Weighted Minority Oversampling TEchnique-Random Forest ( MWMOTE-RF) credit assessment method was proposed. Firstly, MWMOTE technology was applied to increase the samples of minority classes in the preprocessing stage. Then, on the preprocessed balanced dataset, random forest algorithm, one of supervised machine learning algorithms, was used to classify and predict the data. With Area Under the Curve ( AUC) used to evaluate the performance of classifier, experiments were conducted on German credit card dataset from UCI database and a company's car default loan dataset. The results show that the AUC value of MWMOTE-RF method increases by 18% and 20% respectively compared with random forest method and Naive Bayes method on the same data set. At the same time, random forest method was combined with Synthetic Minority Over-sampling TEchnique ( SMOTE) and ADAdaptive SYNthetic over-sampling ( ADASYN), respectively, and the AUC value of MWMOTE-RF method increases by 1. 47% and 2. 34% respectively compared with them. The results prove the effectiveness and the optimization of classifier performance of the proposed method.

**Key words:** umbalanced dataset; machine learning; Majority Weight Minority Oversampling TEchnique ( MWMOTE); random forest; credit assessment

### 0 引言

伴随着互联网金融的日渐兴起,数据挖掘和机器学习等新兴技术在企业经营和科学决策中的普遍应用,在线信贷作为一种更高效的借贷服务早已颠覆了传统银行相关部门的地位,传统的信用评分模型已经不能高效准确地处理信贷客户数据。因此,构建并应用精确、客观和可靠的信用风险评估方法,对于银行业和有信贷业务的公司,在不同的商业周期和环境下减轻信贷业务危机和损失<sup>[1]</sup>有着十分重要的现实意义。

迄今为止,大量数据分析技术和建模技术被应用到风险评估领域,从而出现了四大类风险评估方法:统计学方法、运筹学方法、非参数分析法和人工智能方法。基于统计学方法中最具代表性的就是逻辑回归分析,其是当前理论体系中最为成熟的一种分类模型,最早由 Wiginton 等<sup>[2]</sup>于 1980 年应用于信用风险评估 中。人工智能方法中包括专家系统、神经网络评估系统、支持向量机、遗传算法和随机森林方法。Desai 等<sup>[3]</sup>于 20 世纪 90 年代将神经网络应用于信用风险分析,同期 Baesens 等<sup>[4]</sup>将支持向量机方法运用于信用评分领域。

收稿日期:2018-10-30;修回日期:2019-01-21;录用日期:2019-01-23。

基金项目:国家重点研发计划项目(2017YFB1400803);国家自然科学基金资助项目(31571563,61601310)。

作者简介:田臣(1994—),男,北京人,硕士研究生,主要研究方向:数据挖掘; 周丽娟(1969—),女,辽宁辽阳人,教授,博士,主要研究方向:数据挖掘、机器学习、大数据处理、云计算、数据库系统。



Davis 将<sup>[5]</sup>遗传算法应用在了信用评分领域。国内的诸多学者也在信用评估领域中有所研究,李志辉等<sup>[6]</sup>采用主成分分析法和 Fisher 线性方法、Logit 模型、BP 神经网络技术构造我国商业银行信用风险识别模型,通过实证分析得出相对于其他两类模型,Logit 模型具有更强的信用风险识别和预测能力。王春峰等<sup>[7]</sup>改进了蚁群算法并将其应用在了商业银行信用风险评估中,分析结果相较于判别分析、回归分类算法更好。随机森林方法是一种既可用于分类也能用于回归任务的数据挖掘方法,预测准确率高、不容易出现过拟合、训练速度快等优点使其在很多领域都有广泛的应用<sup>[8-10]</sup>。

就我国银行业的个人信贷业务而言,发展较晚,信贷风险控制方面还存在着明显的不足<sup>[11]</sup>,而最为核心的问题,仍然是如何有效地对不对称信息进行处理,如何高效解决数据类别不平衡问题。所谓类别不平衡数据就是在数据集中,各类别样本数目差别很大,样本分布不均,其中类别数量多的为多数类,类别数量少的为少数类,又称为稀有类。在多数情况下,诸如如网络入侵检测<sup>[12]</sup>、欺诈检测、垃圾邮件识别,信用评估领域等少数类往往是研究的重点。目前处理不平衡问题的主要数据层面方法是过采样或者欠采样,重新分配类别分布,例如:合成少数类过采样技术(Synthetic Minority Over-sampling TEchnique, SMOTE)方法<sup>[13]</sup>、自适应综合过采样(ADAdptive SYNthetic over-sampling, ADASYN)方法<sup>[14]</sup>和 Borderline-SMOTE 方法<sup>[15]</sup>等。

基于以上分析与认识,考虑到单一方法难以在不平衡数据集上达到良好预测效果,本文提出了一种基于带多数类权重的少数类过采样技术和随机森林(Majority Weighted Minority Oversampling Technique-Random Forest, MWMOTE-RF)结合的信用评估方法。本文方法的基本思想是将 MWMOTE 数据处理作为随机森林算法的前置预处理系统,通过 MWMOTE 对信用样本数据进行少数类样本数量增加,从而改善随机森林向多数类类别样本的倾向性问题。最后结合 UCI 数据集和汽车违约贷款数据集与传统的随机森林方法和朴素贝叶斯方法进行实验分析对比。除此之外,分别通过 SMOTE 方法、ADASYN 方法和 Borderline-SMOTE 方法产生平衡数据集训练随机森林模型作为实验对比模型。

## 1 相关方法及模型的构建

### 1.1 MWMOTE

在信用评估领域,客户评估数据中履约的客户占绝大多数,而违约的客户作为少数类样本是我们重点研究的对象。随机森林算法在处理不平衡数据的问题上存在着缺陷,主要是由于少数类样本占比少,在此数据集上训练出来的决策树不能很好地体现少数类的特点,只有增大少数类占有量或是平衡多数类样本数量才能使随机森林算法更加健壮。针对不平衡数据的处理方法有三大类<sup>[16]</sup>:抽样法、代价敏感方法和集成方法。其中抽样方法分为欠抽样和过抽样,在处理不平衡数据集问题上目前应用最广的是 SMOTE 方法,作为过抽样方法的一种,其主要是结合少数类样本按照一定规则合成少数类样本,最终达到平衡数据集的目的<sup>[17]</sup>。但其存在着几点不

足<sup>[18]</sup>:不能精确控制合成新样本数量;不能对少数类样本进行区别性选择;样本混叠现象严重。

鉴于 SMOTE 方法存在的不足,本文采用了带多数类权重的少数类过采样法<sup>[19]</sup>,相较于应用广泛的 SMOTE 方法,可以有效避免新合成样本混叠问题。该方法的核心思路是首先识别难以学习的信息丰富的少数类样本,并根据它们与最近的多数类样本之间的欧氏距离给它们赋值;然后,使用聚类方法从加权信息量大的少数类样本中合成新样本。通过这种方式,所有生成的新样本都位于某个少数类簇中。

带多数类权重的少数类过采样技术(MWMOTE)生成新样本过程如算法 1 所示。

#### 算法 1 MWMOTE 生成新样本。

**输入** 多数类样本集  $S_{\text{maj}}$ , 少数类样本集  $S_{\text{min}}$ , 预计合成的样本的数量  $N$ , 用来测噪声样本的少数类样本邻居样本数  $k_1$ , 用于构造信息量大的少数类样本集的多数类邻居数  $k_2$ , 用于构造信息量大的少数类样本集的少数类邻居数  $k_3$ ;

**输出** 过采样生成的少数类样本集  $S_{\text{omin}}$ 。

算法开始:

1) 对所有少数类样本  $x_i \in S_{\text{min}}$ , 计算得出最近邻数据集  $NN(x_i)$ ,  $NN(x_i)$  是由  $k_1$  个通过与  $x_i$  进行欧氏距离计算的邻居样本组成。

2) 构建过滤后的少数类样本集  $S_{\text{minf}}$ , 若  $NN(x_i) = 0$  表示第  $i$  个少数类样本附近的  $k_1$  个邻居没有少数类样本, 该样本为噪声样本。

$$x_i \in S_{\text{minf}} \quad (1)$$

3) 对所有样本  $x_i \in S_{\text{minf}}$  计算得出最近邻数据集  $N_{\text{maj}}(x_i)$ ,  $N_{\text{maj}}(x_i)$  是由  $k_2$  个通过与  $x_i$  进行欧氏距离计算的多数类邻居样本组成。

4) 获取多数类边界数据集,记为  $S_{\text{bmaj}}$ :

$$S_{\text{bmaj}} = \bigcup_{x_i \in S_{\text{minf}}} N_{\text{maj}}(x_i) \quad (2)$$

5) 对所有样本  $y_i \in S_{\text{bmaj}}$  计算得出最近邻少数类数据集  $NN(y_i)$ ,  $NN(y_i)$  是由  $k_3$  个通过与  $y_i$  进行欧氏距离计算的少数类邻居样本组成。

6) 获取少数类数据信息集,记为  $S_{\text{imin}}$ :

$$S_{\text{imin}} = \bigcup_{y_i \in S_{\text{bmaj}}} N_{\text{min}}(y_i) \quad (3)$$

7) 对所有样本  $y_i \in S_{\text{bmaj}}, x_i \in S_{\text{imin}}$  计算信息权重  $I_w(y_i, x_i)$ 。

8) 对所有样本  $x_i \in S_{\text{imin}}$ , 计算其选择权重  $S_w(x_i)$ :

$$S_w(x_i) = \sum_{y_i \in S_{\text{bmaj}}} I_w(y_i, x_i) \quad (4)$$

9) 根据选择权重  $S_w(x_i)$  计算其选择概率  $S_p(x_i)$ :

$$S_p(x_i) = S_w(x_i) / \sum_{z_i \in S_{\text{imin}}} S_w(z_i) \quad (5)$$

10) 对  $S_{\text{min}}$  进行聚类分析,得到  $M$  个类簇  $L_1, L_2, \dots, L_M$ 。

11) 初始化  $S_{\text{omin}} = S_{\text{min}}$ 。

12) For  $j = 1, 2, \dots, N$

a) 根据选择概率  $S_p(x_i)$ , 从  $S_{\text{imin}}$  中选取  $x$  将其划分到  $L_k$  类簇中,  $1 \leq k \leq M$ 。

b) 随机从  $L_k$  中选取样本  $y$ 。

c) 合成新的样本  $s$ :

$$s = x + \alpha \times (y - x) \quad (6)$$

其中:系数  $\alpha$  是一个随机数,其取值范围为  $[0, 1]$ 。

d) 将  $s$  加入到  $S_{\text{omin}}$ :  $S_{\text{omin}} = S_{\text{omin}} \cup \{s\}$ 。

结束

### 1.2 随机森林

随机森林是一种统计学理论,是 bagging 算法和分类回归



树(Classification And Regression Tree, CART)的结合。通过组合多个CART进行预测,最终通过投票得到预测结果。

Bagging算法又称自举汇聚法,是一种基于数据随机重抽样的分类器构建方法,在原始数据集上进行有放回的抽样 $N$ 次,得到 $N$ 个新数据集。新数据集与原始数据集大小相等。在这 $N$ 个数据集上分别对学习算法进行训练,得到了 $N$ 个弱分类器,由此方法集成为一个强分类器并最终选择分类器投票结果中最多的类别作为分类结果。此处的学习算法为CART,一种改进的决策树。与ID3和C4.5两种影响较大的决策树方法相比,CART算法是基于基尼系数的决策树算法。CART包括分类树和回归树两部分,其中分类树根据基尼系数进行特征空间的划分,回归树通过最小化平方误差进行特征选择和特征值选择。

随机森林的构建过程如下:

1)假设原样本集有 $N$ 个样例,则每轮从原始样本集中有放回地抽取 $n$ 个样例,得到一个与原始样本集相同大小的样本集。经过 $K$ 轮的抽取获得的训练集分别为 $T_1, T_2, \dots, T_K$ 。

2)每个训练集训练一个决策树模型。共得到 $K$ 个CART模型。

3)假设原始样本的特征个数为 $D$ ,从 $D$ 个特征中随机选择其中的 $d$ 个特征( $d < D$ )组成一个新的特征集。对于单个的CART模型,每次分裂时根据基尼系数对对应的新特征集选择最好的特征进行分裂。

4)每棵树不断分裂,直到该节点的所有训练样本都属于同一类。这期间不需要剪枝处理。

5) $K$ 个CART相互独立,其被赋予的权重均相等。对于分类问题,最终的分类结果使用所有的CART投票来确定最终分类结果;对于回归问题,使用所有决策时输出的均值来作为最终的输出结果。

选择随机森林方法主要基于以下考虑:随机森林方法作为一种集成学习方法相较于单一学习器有着优越的泛化性能。文献[9]中,通过实验分析对比可知,随机森林方法的准确率和稳定性要优于支持向量机方法、 $k$ -近邻方法、CART方法、基于径向基的神经网络方法和梯度提升决策树(Gradient Boosting Decision Tree, GBDT)方法等。

本文所用的随机森林算法是python的sklearn库中封装好的。随机森林在sklearn的分类库中所属类是RandomForestClassifier,重要的调节参数如表1所示。

### 1.3 模型融合过程

在MWMOTE的实现过程中,构建了一个用来合成新样本的少数类信息集 $S_{\min}$ 。然而,这个集合的所有样本可能并不同等重要。一些样本可能比其他样本为数据提供更有用的信息,因此,有必要根据样本的重要性为其分配权重。权重越大的样本意味着需要从它附近产生许多合成样品。MWMOTE所用到的选择权重计算公式是鉴于三点观察:接近决策边界的样本包含的信息比距离远的样本多;稀疏簇中的少数类样本比稠密簇中的样本更重要;在密集多数类群附近的少数类样本比在稠密多数类群附近的样本更重要。

在MWMOTE中,信息权重 $I_w(y_i, x_i)$ 为贴近度因子 $C_f(y_i, x_i)$ 与密度因子 $D_f(y_i, x_i)$ 的乘积:

$$I_w(y_i, x_i) = C_f(y_i, x_i) * D_f(y_i, x_i) \quad (7)$$

贴近度因子 $C_f(y_i, x_i)$ 的计算非常直观,如果 $x_i \notin$

$N_{\min}(y_i)$ ,则 $C_f(y_i, x_i) = 0$ ;否则将按照以下步骤计算 $C_f(y_i, x_i)$ :

$$d_n(y_i, x_i) = dist(y_i, x_i) / l \quad (8)$$

其中 $l$ 为特征空间的维数。按照下述方式计算 $C_f(y_i, x_i)$ :

$$C_f(y_i, x_i) = \frac{f(1/d_n(y_i, x_i))}{C_f(th)} * C_{\max} \quad (9)$$

其中: $C_f(th)$ 和 $C_{\max}$ 为自定义的值; $f$ 为截止函数。

$$f(x) = \begin{cases} x, & x \leq C_f(th) \\ C_f(th), & \text{其他} \end{cases} \quad (10)$$

密度因子 $D_f(y_i, x_i)$ 在MWMOTE中计算方式如下:

$$D_f(y_i, x_i) = \frac{C_f(y_i, x_i)}{\sum_{q \in S_{\min}} C_f(y_i, q)} \quad (11)$$

表1 随机森林参数

Tab. 1 Random forest parameters

参数	含义
n_estimators	弱学习器的最大迭代次数,默认是10
bootstrap	是否有放回的采样,默认是True
criterion	CART作划分时对特征的评价标注。分类RF对应的CART分类树默认是基尼系数gini,另一个可选择的标准是信息增益entropy,是用来选择节点的最优特征和切分点的两个准则
max_features	RF划分时考虑的最大特征数,可以使用多种类型的值,默认是“None”
max_depth	决策树最大深度,默认为“None”,常用的取值在10~100
min_samples_split	内部节点再划分所需最小样本数,是限制子树继续划分的条件
min_samples_leaf	叶子节点最少样本数
max_leaf_nodes	最大叶子节点数。通过限制最大叶子节点,可以防止过拟合,默认为“None”
min_impurity_split	节点划分最小不纯度(基于基尼系数,均方差),此值限制决策树增长

本文将MWMOTE和随机森林进行组合,通过MWMOTE对原始不平衡数据集进行处理,生成新的少数类样本并和原数据集组合以达到平衡数据集的目的。随机森林模型在新的样本数据的基础上进行训练,则MWMOTE和随机森林融合模型的训练过程如下:

1)设置参数 $k_1, k_2, k_3$ 和 $N$ ,在原始数据的基础上通过MWMOTE方法产生新的少数类样本数据,降低数据集的不平衡度,最终和原始数据结合并经过标准化处理后成为均衡样本数据集。

2)进行特征工程,删除冗余特征,选择重要性高的特征做为特征集对随机森林模型进行训练。

3)对随机森林模型的相关参数进行初始化,诸如:最大迭代次数、决策树个数、最大特征数和决策树深度等,并结合步骤1)和步骤2)中得到的新的数据样本和特征集进行训练,得到一个最终的预测模型。

## 2 实验数据和评价指标

### 2.1 数据集

本文所用的是UCI KDD Archive提供的德国信用卡数据



以及某公司提供的汽车违约贷款数据作为有信用记录的样本。汽车违约贷款数据含有 5 845 个样本,每一个样本有 19 个连续变量、1 个离散性变量。通过类别标签划分用户,其中,4 648 个信用好的用户、1 197 个信用差的用户。按照 4:1 的比例,本文选取 793 个信用好的用户,207 个信用差的用户,共计 1 000 个用户样本作为最终的实验使用数据集。汽车违约贷款数据集的基本特征如表 2 所示。德国信用卡数据有 1 000 个样本,每一个样本有 7 个连续型变量、13 个离散型变量。类别标签将样本用户进行区分,其中,700 个信用好的样本数据作为多数类,300 个信用差的样本数据作为少数类,是一个非平衡数据集。德国信用数据集的基础特征如表 3 所示。

表 2 汽车违约贷款数据集特征

Tab. 2 Features of car default loan dataset

特征名称	符号	特征名称	符号
application_id	A1	purch_price	A11
account_number	A2	msrp	A12
bankruptcy_ind	A3	down_pyt	A13
tot_derog	A4	loan_term	A14
tot_tr	A5	loan_amt	A15
age_oldest_tr	A6	ltv	A16
tot_open_tr	A7	tot_income	A17
tot_rev_tr	A8	veh_mileage	A18
tot_rev_debt	A9	fico_score	A19
rev_util	A10	bad_ind	A20

表 3 德国信用卡数据集特征

Tab. 3 Features of German credit card dataset

特征名称	符号	特征名称	符号
status	A1	residence	A11
duration	A2	property	A12
history	A3	age	A13
purpose	A4	plans	A14
amount	A5	housing	A15
bonds	A6	existing	A16
employment	A7	job	A17
installmentrate	A8	provide	A18
personal	A9	telephone	A19
guarantors	A10	foreignworker	A20

参考文献[20]数据预处理的特征选择,消除不相关和冗余的特征,最终实验用的德国信用数据训练集中,只选取 {status, amount, duration, age, purpose, history, employment, bonds, property, installmentrate} 作为最后的特征集,以达到提高分类精度和缩短训练时间的目的。

## 2.2 评价指标

受测工作者特征曲线(Receiver Operating Characteristic Curve, ROC)作为公认的不平衡数据集分类器的评价标准,并不能定量评价分类器<sup>[21]</sup>,因此本文采用 AUC(Area Under Curve)值作为性能度量标准。AUC 值被定义为 ROC 曲线下的面积。对于二分类问题,文献[22]给出了计算式如下:

$$\hat{A} = \frac{S_0 - n_+ (n_+ - 1)/2}{n_+ n_-} \quad (12)$$

式中: $n_+$  和  $n_-$  分别表示测试集中正例个数和负例个数; $S_0 = \sum r_i, r_i$  表示第  $i$  个正例在排序表中的序号。常用到的分类器评估标准包含 Gains、lift、AUC 值、ROC 曲线和准确率。相较于常用到的准确率评估,AUC 值作为性能度量标准有四点优秀性质<sup>[23]</sup>:1)与所选决策阈值相互独立;2)不随类概率分布的变化而改变;3)提高了变量分析测试中的敏感度;4)避免了确定不同种类错误分类的代价。

## 3 实验和结果分析

为了提高实验分析的准确性,本文采用多次随机实验进行验证,将原始数据集划分为训练集和测试集,共进行 100 次实验验证。数据划分情况如表 4 所示,模型相关参数如表 5 所示。

表 4 数据集划分

Tab. 4 Division of datasets

样本总量	训练集样本数	测试集样本数	比例	实验次数
1 000	700	300	7:3	50
1 000	800	200	8:2	50

表 5 实验参数

Tab. 5 Experimental parameters

参数	值	参数	值	参数	值
$n_{estimators}$	50	$k_1$	5	$C_f(th)$	5
$max\_features$	5	$k_2$	3	$C_{max}$	2

将实验数据划分出的测试集作为最终模型实验分析对比所用的测试集。在实验用的德国信用数据集的基础上通过 MWMOTE 方法扩充 200 个少数类合成新样本,在实验用的汽车违约贷款数据集的基础上通过 MWMOTE 方法扩充 300 个少数类合成新样本以达到平衡数据集的目的,作为新的样本数据集,其数据划分标准和原始数据集一样。在平衡数据集上训练出的模型称之为 MWMOTE-RF,在原始数据集上训练出的随机森林模型称为 RF,朴素贝叶斯模型称为 NB。

除此之外,本文分别通过 SMOTE 方法,自适应综合过采样方法和 Borderline-SMOTE 方法对实验数据进行处理,在各自产生的平衡数据集上训练随机森林模型,对应生成的模型分别称之为 SMOTE-RF、ADA-RF 和 BSMOTE-RF。

图 1 为不同德国信用数据划分比例下的 MWMOTE-RF 模型、RF 模型和 NB 模型的 ROC 曲线对比。图 2 为不同汽车违约贷款数据划分比例下的 MWMOTE-RF 模型、RF 模型和 NB 模型的 ROC 曲线对比。

图 1~2 均为不同数据划分比下实验测试中某一次的 ROC 曲线对比。由于在信用风险评估中主要关注的是有违约行为的样本,所以 AUC 值计算涉及到的正例为违约样本。结合表 6 给出的实验结果,在训练数据集和测试数据集划分比例为 7:3 时,在德国信用数据集下的 AUC 值,MWMOTE-RF 模型为 0.9403,比 RF 模型提高了 18.04%,比 NB 模型提高了 20.82%,比 SMOTE-RF 模型的 AUC 值提高了 1.47%,比 ADA-RF 模型的 AUC 值提高了 2.34%;在汽车违约贷款数据集下的 AUC 值,MWMOTE-RF 模型为 0.9357,比 RF 模型提



高了 17.7%，比 NB 模型提高了 20.45%，比 SMOTE-RF 模型提高了 0.76%，比 BSMOTE-RF 模型提高了 0.81%。在德国信用数据集和汽车违约信贷数据划分比为 8:2 的条件下，也可以很明显地看出 MWMOTE-RF 模型优于 RF 模型和 NB 模型。经过 MWMOTE 技术处理过的数据集增加了少数类样本的数量，从而使得随机森林模型在后续训练的过程中更好地学习了少数类样本的特征，提高了信用评估的精准度。综上可知，本文提出的 MWMOTE-RF 模型基本上可以满足信贷公司或银行对客户信用评估的实际需要。

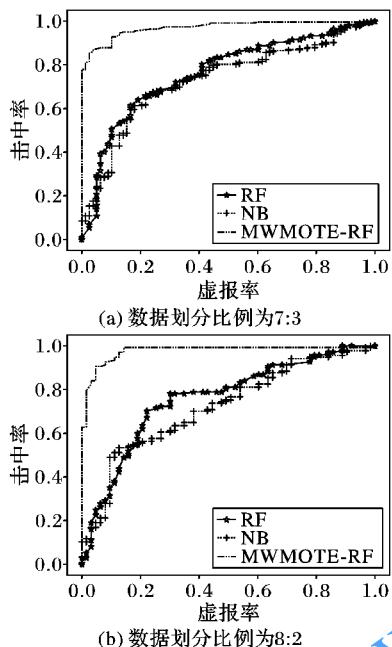


图 1 不同德国信用数据划分比例下各模型 ROC 曲线对比

Fig. 1 Comparison of ROC curves for different models with different division ratio of German credit data

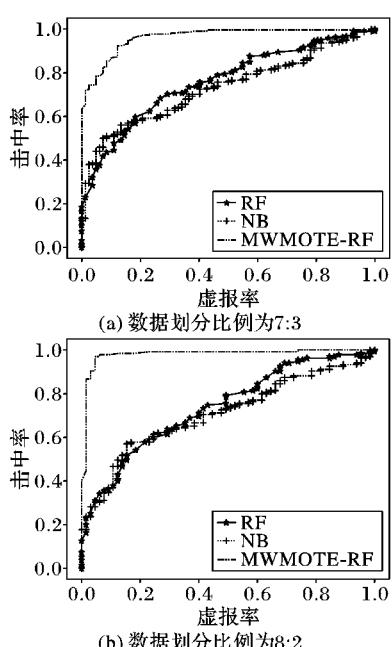


图 2 不同汽车违约贷数据划分比例下各模型 ROC 曲线对比

Fig. 2 Comparison of ROC curves for different models with different division ratios of car default loan data

表 6 不同模型实验结果对比

Tab. 6 Comparison of experimental results of different models

模型	数据划分	AUC 值	
		德国信用数据集	汽车违约贷款数据集
RF	7:3	0.7599	0.7587
NB	7:3	0.7321	0.7312
MWMOTE-RF	7:3	0.9403	0.9357
SMOTE-RF	7:3	0.9256	0.9281
BSMOTE-RF	7:3	0.9419	0.9276
ADA-RF	7:3	0.9169	0.9468
RF	8:2	0.7634	0.7677
NB	8:2	0.7342	0.7381
MWMOTE-RF	8:2	0.9538	0.9757
SMOTE-RF	8:2	0.9555	0.9795
BSMOTE-RF	8:2	0.9491	0.9771
ADA-RF	8:2	0.9539	0.9816

#### 4 结语

为了提高不平衡数据集中可能存在的少数类样本(违约客户)的预测准确率，本文提出了一种基于 MWMOTE 和随机森林结合的信用评估方法，改进了对违约客户的信用评估分析预测能力。经过 MWMOTE 技术处理后，该方法有效解决了信用评估中不平衡数据集的问题，一定程度上解决了分类器向多数类类别样本的倾向性问题。实验结果表明，在处理后的平衡数据集上训练的随机森林模型，其 AUC 值有很大程度提升。但随机森林和 MWMOTE 中的部分参数为人工设置，不一定是最优的模型参数，其次在高维和规模大的数据集上存在训练效率低的问题，因此如何选取合理参数并提升模型训练效率是下一步解决的问题。

#### 参考文献 (References)

- [1] WIN S. What are the possible future research directions for bank's credit risk assessment research? A systematic review of literature [J]. International Economics and Economic Policy, 2018, 15(4): 743–759.
- [2] WIGINTON J C. A note on the comparison of logit and discriminant models of consumer credit behavior [J]. Journal of Financial and Quantitative Analysis, 1980, 15(3): 757–771.
- [3] DESAI V S, CROOK J N, JR OVERSTREET G A. A comparison of neural networks and linear scoring models in the credit union environment [J]. European Journal of Operational Research, 1996, 95(1): 24–37.
- [4] BAESENS B, van GESTEL T, VIAENE S, et al. Benchmarking state-of-the-art classification algorithms for credit scoring [J]. Journal of the Operational Research Society, 2003, 54(6): 627–635.
- [5] DAVIS S, ALBRIGHT T. An investigation of the effect of Balanced Scorecard implementation on financial performance [J]. Management Accounting Research, 2004, 15(2): 135–153.
- [6] 李志辉,李萌.我国商业银行信用风险识别模型及其实证研究 [J].经济科学,2005(5):61–71.(LI Z H, LI M. Credit risk identification model of Chinese commercial banks and its empirical study [J]. Economic Science, 2005(5):61–71.)
- [7] 王春峰,赵欣,韩冬.基于改进蚁群算法的商业银行信用风险评估方法[J].天津大学学报(社会科学版),2005,7(2):81–85.



- (WANG C F, ZHAO X, HAN D. A model on modified ants algorithm for credit risk assessment in commercial banks [J]. Journal of Tianjin University (Social Sciences), 2005, 7(2): 81–85.)
- [8] 方匡南, 吴见彬, 朱建平, 等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011, 26(3): 32–38. (FANG K N, WU J B, ZHU J P, et al. A review of technologies on random forests [J]. Statistic & Information Forum, 2011, 26(3): 32–38.)
- [9] 萧超武, 蔡文学, 黄晓宇, 等. 基于随机森林的个人信用评估模型研究及实证分析[J]. 管理现代化, 2014, 34(6): 111–113. (XIAO C W, CAI W X, HUANG X Y, et al. Research and empirical analysis of personal credit evaluation model based on random forest [J]. Modernization of Management, 2014, 34 (6): 111–113.)
- [10] 李进. 基于随机森林算法的绿色信贷信用风险评估研究[J]. 金融理论与实践, 2015(11): 14–18. (LI J. Study on green-credit risk assessment based on random forest algorithm [J]. Financial Theory & Practice, 2015 (11): 14–18.)
- [11] 杨爱香. 浅析我国商业银行信贷风险管理的现状及对策[J]. 时代金融, 2015 (30): 37, 39. (YANG A X. A brief analysis of China's commercial banks credit risk management status and countermeasures [J]. Times Finance, 2015(30): 37, 39.)
- [12] 封化民, 李明伟, 侯晓莲, 等. 基于SMOTE和GBDT的网络入侵检测方法研究[J]. 计算机应用研究, 2017, 34(12): 3745–3748. (FENG H M, LI M W, HOU X L, et al. Study of network intrusion detection method based on SMOTE and GBDT [J]. Application Research of Computers, 2017, 34(12): 3745–3748.)
- [13] 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(1): 321–357.
- [14] HE H B, BAI Y, GARCIA E A, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning [C] // Proceeding of the 2008 IEEE International Joint Conference on Neural Networks. Piscataway, NJ: IEEE, 2008: 1322–1328.
- [15] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning [C] // ICIC 2005: Proceedings of the 2005 International Conference on Advances in Intelligent Computing. Berlin: Springer, 2005: 878–887.
- [16] 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(6A): 22–27, 57. (ZHAO N, ZHANG X F, ZHANG L J. Overview of imbalanced data classification [J]. Computer Science, 2018, 45(6A): 22–27, 57.)
- [17] 沈学利, 覃淑娟. 基于SMOTE和深度信念网络的异常检测[J]. 计算机应用, 2018, 38(7): 1941–1945. (SHEN X L, QIN S J. Anomaly detection based on synthetic minority oversampling technique and deep belief network [J]. Journal of Computer Applications, 2018, 38(7): 1941–1945.)
- [18] 王超学, 张涛, 马春森. 面向不平衡数据集的改进型SMOTE算法[J]. 计算机科学与探索, 2014, 8(6): 727–734. (WANG C X, ZHANG T, MA C S. Improved SMOTE algorithm for imbalanced datasets [J]. Journal of Frontiers of Computer Science and Technology, 2014, 8(6): 727–734.)
- [19] BARUA S, ISLAM M M, YAO X, et al. MWMOTE — Majority weighted minority oversampling technique for imbalanced data set learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 405–425.
- [20] 叶晓枫, 鲁亚会. 基于随机森林融合朴素贝叶斯的信用评估模型[J]. 数学的实践与认识, 2017, 47(2): 68–73. (YE X F, LU Y H. Credit assessment model based on random forest and naive bayes [J]. Mathematics in Practice and Theory, 2017, 47(2): 68–73.)
- [21] 李治靖, 郭海湘, 李亚楠, 等. 一种基于Boosting的集成学习算法在不均衡数据中的分类[J]. 系统工程理论与实践, 2016, 36(1): 189–199. (LI Y J, GUO H X, LI Y N, et al. A boosting based ensemble learning algorithm in imbalanced data classification [J]. Systems Engineering — Theory & Practice, 2016, 36(1): 189–199)
- [22] HAND D J, TILL R J. A simple generalization of the area under the ROC curve for multiple class classification problems [J]. Machine Learning, 2001, 45(2): 171–186
- [23] 蒋帅. 基于AUC的分类器性能评估问题研究[D]. 长春: 吉林大学, 2016: 10–17. (JIANG S. Researches of performance evaluation of classifier based on AUC [D]. Changchun: Jilin University, 2016: 10–17.)

This work is partially supported by the National Key R&D Program (YFB1400803), the National Natural Science Foundation of China (31571563, 61601310).

**TIAN Chen**, born in 1994, M. S. candidate. His research interest includes data mining.

**ZHOU Lijuan**, born in 1969, Ph. D., professor. Her research interests include data mining, machine learning, big data processing, cloud computing, database system.