计算机科学
COMPUTER SCIENCE

CTGANBoost:基于CTGAN与Boosting的信贷欺诈检测研究

卓佩妍, 张瑶娜, 刘炜, 刘自金, 宋友

引用本文

卓佩妍, 张瑶娜, 刘炜, 刘自金, 宋友. CTGANBoost:基于CTGAN与Boosting的信贷欺诈检测研究[J]. 计算机科学, 2024, 51(6A): 230600199-7.

ZHUO Peiyan, ZHANG Yaona, LIU Wei, LIU Zijin, SONG You. CTGANBoost:Credit Fraud Detection Based on CTGAN and Boosting [J]. Computer Science, 2024, 51(6A): 230600199-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230600199

CTGANBoost:基于 CTGAN 与 Boosting 的信贷欺诈检测研究

卓佩妍　张瑶娜　刘　炜　刘自金　宋　友
北京航空航天大学软件学院　北京 100191
(zhuopy@buaa.edu.cn)

摘　要　在金融行业中,信贷欺诈检测是一项重要的工作,能够为银行和消金机构减少大量的经济损失。然而,信贷数据中存在类别不平衡和正负样本特征重叠的问题,导致少数类识别灵敏度低且不同类别数据区分度低。针对这些问题,提出一种面向信贷欺诈检测的 CTGANBoost 方法。首先,在 AdaBoost(Adaptive Boosting)方法的每一轮 Boosting 迭代中,引入基于类别标签信息约束的 CTGAN(Conditional Tabular Generative Adversarial Network)方法学习特征分布,进行少数类数据增强工作;其次,基于 CTGAN 合成的增强数据集,设计了权重归一化方法,确保在样本加权过程中保持原始数据集的分布特征和相对权重。在 3 个开源数据集上的实验结果表明,CTGANBoost 方法的表现均优于其他主流的信贷欺诈检测方法,AUC 值提升了 0.5%～2.0%,F1 值提升了 0.6%～1.8%,验证了 CTGANBoost 方法的有效性和泛化能力。

关键词:信贷欺诈;数据类别不平衡;集成学习;生成对抗网络;自适应增强
中图分类号　TP391

CTGANBoost:Credit Fraud Detection Based on CTGAN and Boosting

ZHUO Peiyan,ZHANG Yaona,LIU Wei,LIU Zijin and SONG You
School of Software,Beihang University,Beijing 100191,China

Abstract　In the financial industry,credit fraud detection is an important task,which can reduce a lot of economic losses for banks and consumer institutions. However,there are problems of class imbalance and overlapping features of positive and negative samples in credit data,which lead to low sensitivity of minority class recognition and low data discrimination. To address these problems,a CTGANBoost method is proposed for credit fraud detection. First,in each Boosting iteration of AdaBoost,the conditional tabular generative adversarial network(CTGAN) method based on class label information constraint is introduced to learn feature distribution for minority class data augmentation. Secondly,based on the enhanced data set synthesized by CTGAN,a weight normalization method is designed to ensure that the distribution characteristics and relative weights of the original data set are maintained during the sample weighting process. Experimental results on three open source datasets show that CTGANBoost outperforms other mainstream credit fraud detection methods,with AUC values increase by 0.5%～2.0% and F1 values increase by 0.6%～1.8%,which verifies the effectiveness and generalization ability of CTGANBoost method.

Keywords　Credit fraud,Imbalance data,Ensemble learning,Generative adversarial network,AdaBoost

1　引言

信贷欺诈[1]指借款人通过虚构个人信息、伪造证件、中介包装等方式骗取贷款,且借款人在主观上没有还款意愿,或客观上不具备偿还能力,可能造成出借人(银行、消金机构)资金损失的行为。随着信贷业务规模的不断扩大,银行和金融信贷机构面临越来越严重的信贷欺诈风险。信贷欺诈的检测和预防是一项昂贵、耗时和劳动密集型的任务。信贷欺诈检测通过识别恶意骗贷用户与真正借款用户的差异,预测和评估信贷欺诈风险,从而确保金融机构的资产安全,维护金融市场的稳定。

信贷欺诈检测是一项复杂的任务,其中的重点和难点主要有:

(1)数据类别不平衡:在实际工作中,存在欺诈行为的客户数量远少于正常交易的客户,这导致数据集中的欺诈交易样本与正常交易样本的数量极度不平衡,模型的训练和测试都面临很大的困难。现有的平衡数据类别的方法以随机欠采样和插值过采样[2]为主。随机欠采样方法容易丢失多数类信息;插值过采样方法容易改变数据分布情况,引入噪声,容易在高维、大数据集上表现不佳。针对信贷欺诈场景下的类别不平衡问题,需要采取更加有效的方法。

(2)特征重叠:欺诈者竭力将非法交易掩饰成正常交易,使得正常样本与欺诈样本之间有许多特征重叠[3],某些特征在正常样本和欺诈样本中具有相似的分布和统计特性,难以有效地区分,这使得欺诈检测模型必须具备强大的灵敏度和适应性。现有的欺诈检测方法多为基于专家规则的筛选方法,或者基于数据挖掘的单一分类器,分类效果有限,并不能完全满足要求[4]。信贷欺诈问题依然是银行和消金机构的重点关注难题。

本文针对信贷欺诈不平衡数据的分类问题提出一种

基金项目:河北省重点研发计划(21310101D)
This work was supported by the Key Research and Development Program of Hebei Province ,China(21310101D).
通信作者:宋友(songyou@buaa.edu.cn)

算法，称为 CTGANBoost(CTGAN with Boosting)。在 AdaBoost 算法的每一轮提升迭代中，使用标签信息引导 CTGAN 生成和原始少数类样本分布相似的合成样本，挖掘出更多的少数类样本信息，从而降低训练数据的不平衡程度；基于 CTGAN 合成的增强数据集，设计了权重归一化方法，以平衡样本对模型的影响力，防止模型对合成样本过拟合，确保在样本加权过程中保持原始数据集的分布特征和相对权重；同时，在每一轮 Boosting 迭代中，使用弱分类器进行预测，计算误分类损失以调整样本权重，使模型更加重视误分类样本，提高检测欺诈行为特征的能力；根据每个弱分类器的预测分类损失，对弱分类器进行加权组合，最终得到一个强分类器。CTGANBoost 方法解决了类别不平衡导致的少数类识别灵敏度低和特征重叠导致的不同类别数据区分度低的问题。

在 3 个开源数据集上的实验证明，本文提出的 CTGANBoost 方法在信贷欺诈检测问题上有较好的预测效果，在性能上相比其他模型有较大的提升。

2 相关工作

2.1 数据类别不平衡问题解决方法

传统分类器在处理类别不平衡数据时，会优先保证整体的准确率，重视多数类、忽视少数类，这使得模型存在偏向性，导致其容易陷入对多数类样本的过拟合以及对整体任务的欠拟合，缺乏泛化能力。

针对数据类别不平衡问题，解决方法分为数据层面和算法层面。在数据层面的方法分为过采样和欠采样：过采样方法主要有随机过采样[5]和以 SMOTE[6]，ADASYN[7]为代表的启发式过采样；欠采样方法主要有随机欠采样[5]、改良欠采样[8]。在算法层面的方法主要有：代价敏感学习[9]、One-Class 学习[10]、集成学习[11]等。本文针对数据层面的过采样方法展开研究。

在传统过采样方法的研究中，Douzas 等[12]结合 K-means 和 SMOTE 方法合成结构化数据，能够避免噪声产生。Maldonado 等[13]提出了一种新的距离度量，用于计算每个少数样本的邻域，获得了性能提升。Lu 等[14]结合 ADASYN(Adaptive Synthetic Sampling)方法与随机森林方法，在不平衡数据集上取得了较好的表现。

近年来，基于生成式对抗网络(Generative Adversarial Networks，GANs)[15]的过采样方法被用于解决类别不平衡问题。针对 GANs 的研究集中在非结构化的连续数据(如图像)上，但信贷欺诈的数据集是表格式的结构化数据，并不能完全适用。在当前的研究中，使用 GANs 对带有分类变量的表格数据建模的方法已经出现。Xu 等[16]提出 CTGAN 方法，拟合表格中的离散数据列的条件概率分布，采用高斯混合模型归一化，能够对离散型数据和连续型数据进行建模及生成。Zhao 等[17]提出了 CTAB-GAN 方法，引入信息损失、分类损失和生成器损失。Choi 等[18]提出了 MedGAN 方法，通过自动编码器和生成对抗网络的组合生成高维离散变量，用于生成电子病历。Rajabi 等[19]提出的 Tabfairgan 方法增加了公平性约束以生成准确且公平的数据。

2.2 特征重叠问题解决方法

特征重叠通常指不同类别的样本出现在同一数据空间区域的问题，这增加了分类器区分重叠区域中不同类别样本的难度。由于欺诈者竭力模仿真实持卡人的交易行为，欺诈交易和正常交易将在某些数据空间区域交织在一起，造成重叠问题。

特征重叠问题的解决方法有三大类：(1)数据层面；(2)特征层面；(3)模型层面。数据层面的研究主要对重叠样本进行欠采样，Vuttipittayamongkol 等[20]提出 4 种基于 KNN 的方法，用于识别和删除重叠区域中的多数类样本。Bunkhumpornpat 等[21]提出了一种基于聚类的欠采样技术，用于处理特征重叠的类不平衡问题，并应用基于密度的聚类模型 DBSCAN 来发现并移除重叠区域中的多数样本。特征层面的研究主要对重叠特征做特征选择，Fu 等[22]基于稀疏正则化提出了 MOSNS 和 MOSS 方法进行特征选择，有效提升了不平衡学习性能。Omar 等[23]通过稀疏特征选择来最小化重叠并进行二值分类。在模型层面的研究中，Li 等[24]提出了一种基于加权投影聚类分组和一致模糊样本变换的不平衡集成学习算法，同时解决了类不平衡和类重叠问题。Li 等[3]提出一种基于动态加权熵的混合方法，使用无监督异常检测模型获取重叠子集，能够提高模型效率、避免信息丢失。

2.3 信贷欺诈检测方法

信贷欺诈检测方法分为两大类：(1)基于专家规则；(2)基于数据挖掘。其中，基于专家规则的方法高度依赖领域专家的从业经验，无法识别出新涌现的欺诈模式。实践证明，数据挖掘是更加有效的欺诈检测方法[25]，其可以从历史数据中自动捕获欺诈模式，挖掘出更多的欺诈交易风险。

Xuan 等[26]使用两种随机森林分别训练正常交易和异常交易的行为特征，取得了较好的分类效果。Meng 等[27]结合 SMOTE 方法和 XGBoost 方法以提升性能。Fu 等[28]提出了基于 CNN 的欺诈检测框架，以捕获异常交易样本中的欺诈行为的潜在模式。Bahnsen 等[29]对交易聚合策略进行了扩展，在分析事务时间的周期性行为的基础上，基于 von Mises 分布创建一组新的特性，能够节省财务成本。Chen 等[30]提出了一种基于深度卷积神经网络(Deep Convolution Neural Network，DCNN)的金融欺诈检测方案，当涉及大量数据时，使用该技术可以提高检测精度。Carcillo 等[31]提出了一种结合有监督学习和无监督学习的混合方法，在不同粒度级别上计算无监督异常值分数，提高了欺诈检测的准确性。

3 关键技术

3.1 CTGAN 算法

CTGAN(Conditional Tabular Generative Adversarial Network)[16]是用于生成表格数据的条件生成对抗网络，相比传统的生成对抗网络，其能够更好地处理结构化数据。CTGAN 使用一个生成器网络来生成与原始数据集相似的合成数据集，并通过一个判别器网络来评估生成数据的真实度。生成器和判别器通过对抗学习的方式不断调整参数，以最大化合成数据的真实度。在生成数据时，CTGAN 会将生成器和条件向量(即原始数据集中的类别变量)结合起来，以保证生成数据与原始数据集的条件分布相似。这使得 CTGAN 能够生成符合实际的且具有一定规律性的合成数据集。

因此，这种数据合成方法与以 SMOTE 为代表的利用特征插值进行数据合成的传统方法有本质的不同。CTGAN 在保持数据集的整体概率分布的同时，对离散数据进行有针对性的训练。为了捕捉列与列之间所有可能的相关性，在生成器和判别器中使用全连接网络，并且这两个神经网络都包含

两个全连接隐藏层。CTGAN 的损失函数使用的是 WGAN[32] 的损失函数,网络结构则是采用了 PacGAN[33] 的框架,生成器网络结构为 $G(z,cond)$。如式(1)所示:

$$\begin{cases} h_0 = z \oplus cond \\ h_1 = h_0 \oplus \mathrm{ReLU}(BN(FC_{|cond|+|z|\to 256}(h_0))) \\ h_2 = h_1 \oplus \mathrm{ReLU}(BN(FC_{|cond|+|z|+256\to 256}(h_1))) \\ \hat{\alpha}_i = \tanh(FC_{|cond|+|z|+512\to 1}(h_2)), & 1 \leqslant i \leqslant N_c \\ \hat{\beta}_i = \mathrm{gumbel}_{0.2}(FC_{|cond|+|z|+512\to m_i}(h_2)) & 1 \leqslant i \leqslant N_c \\ \hat{d}_i = \mathrm{gumbel}_{0.2}(FC_{|cond|+|z|+512\to |D_i|}(h_2)) & 1 \leqslant i \leqslant N_d \end{cases} \tag{1}$$

其中,z 表示随机噪声,$cond$ 表示条件概率。在生成器中,使用了批标准化[34] 和 ReLU 激活函数生成合成行数据。标量 $\hat{\alpha}_i$ 是由 tanh 激活函数生成的,状态指示器 $\hat{\beta}_i$ 和离散值 $\hat{d}_i$ 则由 gumbel[35] 函数生成。在判别器中的每个隐藏层使用 leaky ReLU 激活函数和 dropout[36] 函数。

CTGAN 的判别器网络结构为 $C(r_1,\cdots,r_{10},cond_1,\cdots,cond_{10})$,如式(2)所示。

$$\begin{cases} h_0 = r_1 \oplus \cdots \oplus r_{10} \oplus cond_1 \oplus \cdots \oplus cond_{10} \\ h_1 = drop(leaky_{0.2}(FC_{10|r|+10|cond|\to 256}(h_0))) \\ h_2 = drop(leaky_{0.2}(FC_{256\to 256}(h_1))) \\ C(\cdot) = FC_{256\to 1}(h_2) \end{cases} \tag{2}$$

其中,r_i 表示一个样本。

将 CTGAN 应用于处理不平衡数据集,可以在满足原始数据集概率分布的前提下,获得样本更多、分布更均匀、信息更丰富的数据集。

3.2 AdaBoost 算法

AdaBoost(Adaptive boosting)是自适应增强算法,针对同一个训练集训练多个分类器(弱分类器),然后把这些弱分类器集合起来,构成一个更强的最终分类器(强分类器)。假设给定一个数据量为 n 的二分类数据集:

$$S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\} \tag{3}$$

数据集中的每个样本由特征与标签组成,特征 $x_i \subseteq R^n$,n 表示特征维度,$y_i \in \{-1,+1\}$。

AdaBoost 将多个弱分类器组合起来,通过迭代,构成一个预测效果更好的强分类器。在第一轮迭代中,每个样本的权重都是 $1/n$,其中 n 为训练数据集中的样本数量。在接下来的每一轮迭代中,错误分类的样本的权重会增加,而正确分类的样本的权重会减少。这使得下一个弱分类器更有可能专注于先前分类器无法正确分类的样本,从而提高模型的整体准确性。

在所有弱分类器都训练完成之后,根据每个弱分类器的预测性能计算相应的权重,并通过加权组合,将所有的弱分类器集成为最终的分类器,其中每个弱分类器的权重与其分类准确性成正比。集成分类器的形式为:

$$\begin{aligned} H(x) &= sign[f(x)] \\ &= sign[k_1h_1(x)+k_2h_2(x)+\cdots+k_nh_n(x)] \end{aligned} \tag{4}$$

其中,$sign(x)$ 为符号函数,k_i 为第 i 个分类器的权重,$h_i(x)$ 表示弱分类器。

3.3 CTGANBoost 算法

CTGANBoost 算法的主要思路是:在每一轮 Boosting 迭代中引入 CTGAN,并通过样本标签信息引导 CTGAN 训练,生成少数类样本以降低训练数据集的类别不平衡程度,生成少数类样本的过程同时也是对少数类样本信息的进一步挖掘,这使得模型能够获取更多的少数类信息,提高模型对少数类样本的重视程度,提升泛化能力;设计一种新的权重分配算法,为每一轮迭代开始时生成的少数类样本赋予初始权重,并调整其他样本权重,在整体上保持样本分布的一致性;在每一轮迭代中,弱分类器从带有权重的训练数据集中学习,并根据该轮迭代中弱分类器的分类损失计算下一轮迭代的样本权重。经过以上训练,将每一轮训练得到的弱分类器进行线性组合,最终得到一个强分类器。CTGANBoost 的模型结构如图 1 所示。

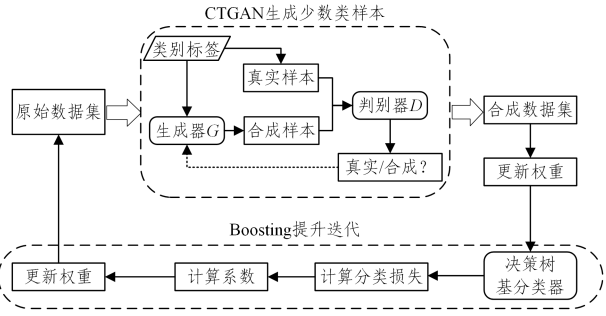

图 1 CTGANBoost 模型结构图

Fig. 1 Diagram of CTGANBoost model structure

本算法使用的弱分类器为决策树(Decision Tree,DT),以决策树为基函数的提升方法称为提升树(Boosting Tree)。提升方法实际采用加法模型与前向分步算法。加法模型是一种将多个基函数通过线性组合的方式构建新模型的方法。前向分步算法是一种迭代算法,每一步都通过拟合当前模型的残差来确定下一个基函数,用于逐步构建加法模型。

对于信贷欺诈检测这类二分类问题,使用的决策树是二叉分类树,提升树模型可以表示为决策树的加法模型:

$$f_M(x)=\sum_{m=1}^{M} T(x;\Theta_m) \tag{5}$$

其中,$T(x;\Theta_m)$ 表示决策树,Θ_m 为决策树的参数,M 为决策树的数目。提升树是一种高效的学习算法,在输入和输出之间的关系非常复杂时,它也可以通过树的线性组合来很好地拟合训练数据。CTGANBoost 算法的伪代码如算法 1 所示。

算法 1 CTGANBoost 算法

输入:训练集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,其中 $x_i \subseteq R^n$,$y_i \subseteq \{-1,+1\}$,Boosting 迭代次数 T,基本分类器 h

输出:强分类器 H

1. 初始化分布权重 $W_1=(w_{1,1},\cdots,w_{1,i},\cdots,w_{1,n})$,$w_{1,i}=1/n$,$i=1,2,\cdots,n$
2. For t=1 to T:
3. 使用 CTGAN 生成 m 个少数类样本
4. 更新权重:$W_t'=(w'_{t,1},\cdots,w'_{t,i},\cdots,w'_{t,n+m})$,其中

$$w'_{t,i}=\begin{cases} \dfrac{w_{t,i}*n}{m+n}, & i=1,2,\cdots,n \\ \dfrac{1}{m+n}, & i=n+1,n+2,\cdots,n+m \end{cases}$$

5. 利用更新后的权重 W_t' 训练一个弱分类器:$h_t: x \to \{-1,+1\}$
6. 根据弱分类器的预测结果计算分类损失:

$e_t=P(h_t(x_i)\neq y_i)=\sum_{i=1}^{n+m} w'_{t,i} I(h_t(x_i)\neq y_i)$

当 $h_t(x_i)\neq y_i$ 时,$I(h_t(x_i)\neq y_i)$ 取值为 1,否则为 0

7. 根据分类损失计算 $h_t(x)$ 的系数:

$k_t=1/2(\log((1-e_t)/e_t))$

8. 更新训练集的权重:$W_{t+1}=(w_{t+1,1},\cdots,w_{t+1,i},\cdots,w_{t+1,n+m})$,

$w_{t+1,i}=w_{t,i}\exp(-k_t * y_i * h_t(x_i))/Z_t, i=1,2,\cdots,n+m$

其中，Z_t是规范化因子：

$$Z_t=\sum_{i=1}^{n+m} w_{t,i}\exp(-k_t * y_i * h_t(x_i))$$

9. 丢弃本轮迭代生成的少数类样本，将原始样本的权重归一化

10. End for

11. 构建基分类器线性组合：$f(x)=\sum_{t=1}^{T} k_t * h_t(x)$

12. 得到最终分类器：$H(x)=\text{sign}(f(x))=\text{sign}(\sum_{t=1}^{T} k_t * h_t(x))$

对算法 1 做如下解释：

对于生成的少数类数量 m，根据不同的数据集，由经验值确定。

在计算基分类器的分类损失时：

$$e_t=P(h_t(X_i)\neq Y_i)=\sum_{i=1}^{n+m} w_{t,i}I(h_t(X_i)\neq Y_i) \tag{6}$$

其中，$w_{t,i}$表示第 t 轮迭代中的第 i 个样本的权重，且$\sum_{i=1}^{n+m} w_{t,i}=1$，这表明$h_t(x)$在加权的训练数据集上的分类损失等于被$h_t(x)$错误分类的样本的权重之和。

因为最终分类器是由 T 个弱分类器线性组合得到的，弱分类器系数k_t表示该弱分类器$h_t(x)$在最终分类器中的重要性。系数k_t的计算公式如下：

$$k_t=\frac{1}{2}\log\frac{1-e_t}{e_t} \tag{7}$$

可以看出，当$e_t\leqslant 1/2$ 时，$k_t\geqslant 0$，并且随着e_t的增大，系数k_t会相应地减小，因此分类损失越大的弱分类器在最终分类器中的作用越小，反之，分类损失越小的弱分类器，在最终分类器中的作用越大。

更新权重分布的公式可以改写为：

$$w_{t+1,i}=\begin{cases}\dfrac{w_{t,i}}{Z_t}\exp(-k_t),\cdots h_t(x_i)=y_i\\ \dfrac{w_{t,i}}{Z_t}\exp(k_t),\cdots h_t(x_i)\neq y_i\end{cases} \tag{8}$$

当一个样本被弱分类器错误分类时，它的权重会增大，而被正确分类的样本的权重会减小。错误分类的样本的权重将会被放大$e^{2k_t}=e_t/(1-e_t)$倍，因此，错误分类样本在下一轮迭代中将更受重视。

在每一轮迭代结束时，丢弃本轮迭代中由 CTGAN 生成的少数类样本，这些假样本不进入下一轮迭代，下一轮迭代将生成新的少数类样本。以此增加少数类样本的多样性，挖掘出更多的少数类信息，从而提升最终分类器辨别少数类样本的能力。

4 实验结果与分析

4.1 数据集概述

实验共使用 3 个数据集，均为开源数据集，数据集的基本信息如表 1 所列。数据集 1 为 UCI 提供的数据集(UCI credit)[1)]，来自台湾地区，包含 23 个特征，少数类占比为 22.12%；数据集 2 是 Kaggle 提供的数据集(Loan Default Prediction，简称 LDP)[2)]，是一个使用金融机构的实际数据创建的合成数据集，数据为脱敏数据，包含 3 个特征，少数类占比为3.33%；数据集 3 同样是来自 Kaggle 的数据集(Credit Card Fraud Detection，简称 CCFD)[3)]，数据来自 2013 年 9 月欧洲信用卡持卡人进行的交易，包含 30 个特征，除 Time(时间)和 Amount(金额)外，均经过 PCA 变换，少数类占比为 0.17%。

表 1 数据集基本信息统计

Table 1 Dataset basic information statistics

	样本数量	正常交易占比	异常交易占比	特征个数
数据集 1：UCI credit	30000	0.7788	0.2212	23
数据集 2：Kaggle LDP	10000	0.9667	0.0333	3
数据集 3：Kaggle CCFD	284807	0.9983	0.0017	30

图 2 中展示了 3 个数据集中正常样本和欺诈样本的部分特征分布图，由于 CCFD 数据集不含离散变量，因此仅绘制了 CCFD 的连续变量分布图。从图 2 可以看出，正常样本和欺诈样本的部分特征重叠程度较高，具有相似的分布和统计特性，区分难度较大。

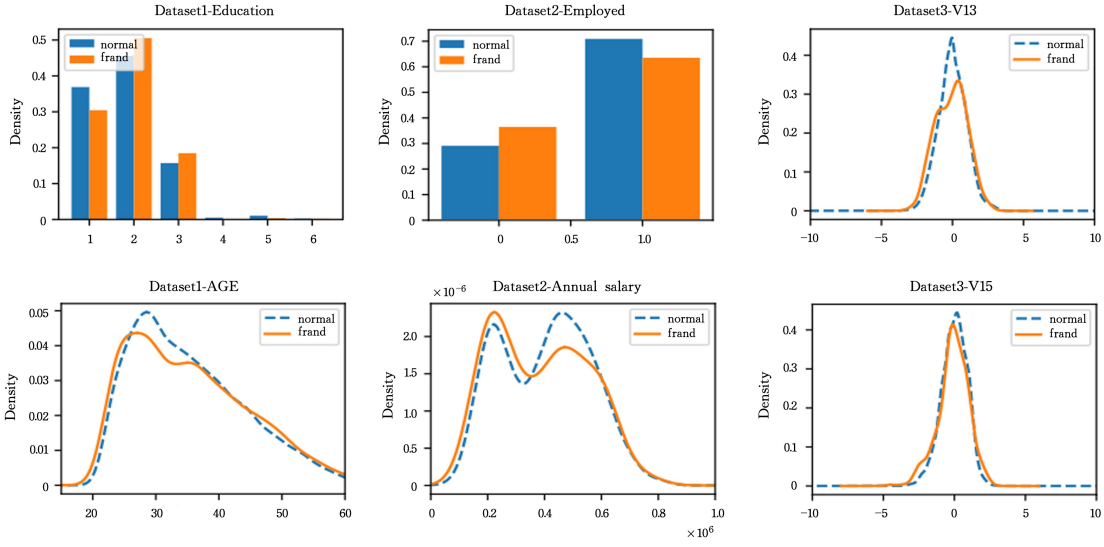

图 2 正常样本与欺诈样本的部分特征分布

Fig. 2 Partial feature distribution between normal and fraudulent samples

1) https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

2) https://www.kaggle.com/datasets/kmldas/loan-default-prediction

3) https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

4.2 评价指标

对于大部分机器学习二分类任务来说,采用准确率(Accuracy)作为评价指标能够最直观地反映出模型的性能。但是由于信贷欺诈检测的数据不平衡问题,使用准确率作为评价指标无法得到有意义的结果,如果模型将所有的数据都预测为多数类,最终能够得到一个较高的准确率,模型的性能看似优秀,但实际上缺乏对少数类的识别能力。因此,本文采用 AUC 值和 F1 值作为评价指标,这两个指标能够直观地体现模型对两种分类的综合预测能力。

AUC(Area Under Curve)被定义为 ROC 曲线下与坐标轴围成的面积,ROC 曲线指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),以真阳性率为纵坐标、假阳性率为横坐标绘制的曲线。AUC 的取值范围为 0.5～1,越接近 1,表示分类器性能越好,当 AUC 值为 0.5 时,表示分类器不具备分类性能。

F1 值是精确率(Precision)和召回率(Recall)的调和平均;精确率表示在被所有预测为正的样本中实际为正样本的概率;召回率表示在实际为正的样本中被预测为正样本的概率。F1 值的取值范围为 0～1,越接近 1,表示分类器性能越好,能够在识别少数类和多数类中达到较高的平衡。F1 值与精确率、准确率的计算关系为:

$$F1=\left(\frac{Recall^{-1}+Precision^{-1}}{2}\right)^{-1} \tag{9}$$

4.3 基线模型

本研究用于比较的基线模型有:

(1)Logistic Regression(LR):逻辑回归模型,是一种广义的线性回归分析模型,由于其极高的可解释性,在银行、消费金融等领域得到广泛应用。

(2)Naive Bayes:朴素贝叶斯法,基于贝叶斯定理与条件独立假设的分类方法。

(3)Decision Tree(DT):决策树,是一种树状结构,它的每一个叶子结点对应着一个分类。决策树模型是 AdaBoost 算法最常用的基分类器。

(4)SMOTE+DT:先采用 SMOTE 方法生成少数类样本,再采用决策树算法作为分类器。

(5)SMOTEBoost:在每一轮 Boosting 迭代中引入 SMOTE 方法解决数据集的类别不平衡问题。

4.4 实验结果及分析

为验证 CTGANBoost 方法在信贷欺诈领域的有效性,本文使用 5 个基线方法和 CTGANBoost 方法进行比较实验,设置评价指标为 AUC 值和 F1 值。出于对模型训练效率和训练效果的考虑,设置 Boost 迭代次数为 50,设置 CTGAN 的迭代次数为 10,batch_size 为 1 000,使用 Adam 优化算法,设置学习率为 0.000 2。随机选取数据集样本量的 75%作为训练集,其余 25%作为测试集。

各方法在 3 个不同的数据集上的表现如表 2、表 3 所列,其中加粗项表示性能最好的分类预测结果。从表 2 和表 3 可以看出,CTGANBoost 方法的表现总体上优于其他方法,有更好的预测性能和分类正确率。在 UCI credit 数据集中,CTGANBoost 方法的 *AUC* 值提高了 0.5%,*F*1 值提高了 1.8%;在 LDP 数据集中,CTGANBoost 方法的 *AUC* 值提高了 0.9%,*F*1 值提高了 0.6%;在 CCFD 数据集中,CTGANBoost 方法的 *AUC* 值提高了 2.0%,*F*1 值提高了 0.7%。相比其他方法,CTGANBoost 能够更好地捕捉到潜在的欺诈模式和行为特征,从而实现更准确的欺诈预测和分类。尽管 SMOTEBoost 方法在不平衡数据上表现优异,在大部分情况下优于其他基线模型,但 CTGANBoost 方法相比 SMOTEBoost 方法仍然有一定的提高。在类别不平衡程度最高的 CCFD 数据集中,CTGANBoost 方法在 *AUC* 值上提高了 2.0%,并且 *F*1 值也有一定提升。这在信贷欺诈检测中很有价值,每一个正确识别出的欺诈案例都可以避免潜在的损失和风险,从而节省大量的资金和资源。因此,即使是微小的性能提升也可能对业务决策和风险管理产生重大影响。

表 2 各分类算法在 3 个数据集上的 AUC 值

Table 2 AUC values of each classification algorithm on 3 datasets

	UCI credit	LDP	CCFD
Logistic Regression	71.4	92.4	88.4
Naive Bayes	74.3	92.6	95.8
Decision Tree(DT)	70.0	89.8	86.6
SMOTE+DT	71.4	91.8	92.9
SMOTEBoost	74.7	92.4	95.6
CTGANBoost	**75.2**	**93.5**	**97.8**

表 3 各分类算法在 3 个数据集上的 F1 值

Table 3 F1 score of each classification algorithm on 3 datasets

	UCI credit	LDP	CCFD
Logistic Regression	14.9	15.5	61.5
Naive Bayes	50.5	33.8	23.0
Decision Tree(DT)	44.2	41.5	74.2
SMOTE+DT	46.3	31.9	62.7
SMOTEBoost	50.2	39.0	76.2
CTGANBoost	**52.3**	**42.1**	**76.9**

为了验证 CTGANBoost 方法中各个模块的有效性,本文进行了一系列消融实验。消融实验设置为:

(1)AdaBoost:仅使用 AdaBoost 方法,不做任何其他平衡数据类别的操作。

(2)CTGAN+DT:先使用 CTGAN 平衡数据类别,后使用单一决策树模型进行分类,不进行 Boost 迭代。

(3)CTGAN+Boost:先使用 CTGAN 平衡数据类别,后使用 AdaBoost 方法进行迭代建模,但 CTGAN 仅在最开始使用一次,不参与 Boost 迭代。

实验结果如表 4 所列,从表中可以看出,CTGANBoost 总体上表现优异,在 3 个数据集上的 AUC 值和 F1 值均高于其他方法。仅采用 Boosting 集成学习,模型欠缺对少数类样本的识别能力;仅采用 CTGAN 和单一分类器,模型无法从历史错误中学习到信息;仅在 Boosting 集成学习前使用一次 CTGAN,模型无法学习到更多少数类信息。这说明,模型在 Boosting 的每一轮迭代过程中使用 CTGAN 平衡数据类别,能够有效提升针对信贷欺诈不平衡数据的预测性能。无论是 Boosting 集成学习,还是 CTGAN 拟合数据分布生成少数类样本,均对提升模型性能起到了正向作用。

表 4 CTGANBoost 方法的系列消融实验

Table 4 Series of ablation experiments using CTGANBoost method

模型	数据集 1		数据集 2		数据集 3	
	AUC	*F*1	*AUC*	*F*1	*AUC*	*F*1
CTGANBoost	**75.2**	**52.3**	**93.5**	**42.1**	**97.8**	**76.9**
AdaBoost	74.2	44.7	89.1	36.2	95.6	72.0
CTGAN+DT	73.0	49.2	83.7	35.2	90.3	75.7
CTGAN+AdaBoost	73.7	48.4	90.4	38.7	96.3	75.5

结束语 为了解决信贷欺诈检测的数据类别不平衡问题和特征重叠问题，本文提出了基于 CTGAN 和 Boosting 的信贷欺诈检测方法，称为 CTGANBoost。该方法通过 CTGAN 挖掘更多少数类信息、给模型提供更多更优质的少数类样本；通过 Boosting 迭代累积误分类经验，并根据弱分类器的预测性能进行加权线性组合，最终得到一个性能较优的强分类器。该方法能够在更加重视欺诈样本的同时，保持对正常样本检测的准确度。实验结果表明，该方法在信贷欺诈检测问题上的性能优于主流方法。在未来的工作中，我们将结合信贷欺诈实时监测系统，展开进一步研究。

参 考 文 献

[1] AWOYEMI J O, ADETUNMBI A O, OLUWADARE S A. Credit card fraud detection using machine learning techniques: A comparative analysis[C]// 2017 International Conference on Computing Networking and Informatics (ICCNI). IEEE, 2017: 1-9.

[2] MISHRA S. Handling imbalanced data: SMOTE vs. random undersampling [J]. Int. Res. J. Eng. Technol, 2017, 4(8): 317-320.

[3] LI Z, HUANG M, LIU G, et al. A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection[J]. Expert Systems with Applications, 2021, 175: 114750.

[4] CARCILLO F, LE BORGNE Y A, CAELEN O, et al. Combining unsupervised and supervised learning in credit card fraud detection [J]. Information Sciences, 2021, 557: 317-331.

[5] MOHAMMED R, RAWASHDEH J, ABDULLAH M. Machine learning with oversampling and undersampling techniques: overview study and experimental results[C]// 2020 11th International Conference on Information and Communication Systems (ICICS). IEEE, 2020: 243-248.

[6] FERNÁNDEZ A, GARCIA S, HERRERA F, et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary[J]. Journal of Artificial Intelligence Research, 2018, 61: 863-905.

[7] BRANDT J, LANZÉN E. A comparative review of SMOTE and ADASYN in imbalanced data classification[J/OL]. 2021. https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1519153&dswid=-3893.

[8] LIN W C, TSAI C F, HU Y H, et al. Clustering-based undersampling in class-imbalanced data [J]. Information Sciences, 2017, 409: 17-26.

[9] FERNÁNDEZ A, GARCÍA S, GALAR M, et al. Cost-sensitive learning[M]// Learning from Imbalanced Data Sets, 2018: 63-78.

[10] SELIYA N, ABDOLLAH ZADEH A, KHOSHGOFTAAR T M. A literature review on one-class classification and its potential applications in big data[J]. Journal of Big Data, 2021, 8(1): 1-31.

[11] TANHA J, ABDI Y, SAMADI N, et al. Boosting methods for multi-class imbalanced data classification: an experimental review[J]. Journal of Big Data, 2020, 7: 1-47.

[12] DOUZAS G, BACAO F, LAST F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE [J]. Information Sciences, 2018, 465: 1-20.

[13] MALDONADO S, LÓPEZ J, VAIRETTI C. An alternative SMOTE oversampling strategy for high-dimensional datasets [J]. Applied Soft Computing, 2019, 76: 380-389.

[14] LU C, LIN S, LIU X, et al. Telecom fraud identification based on ADASYN and random forest[C]// 2020 5th International Conference on Computer and Communication Systems (ICCCS). IEEE, 2020: 447-452.

[15] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks [J]. Communications of the ACM, 2020, 63(11): 139-144.

[16] XU L, SKOULARIDOU M, CUESTA-INFANTE A, et al. Modeling tabular data using conditional gan[J/OL]. Advances in Neural Information Processing Systems, 2019, 32. https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-Abstract.html.

[17] ZHAO Z, KUNAR A, BIRKE R, et al. Ctab-gan: Effective table data synthesizing[C]// Asian Conference on Machine Learning. PMLR, 2021: 97-112.

[18] CHOI E, BISWAL S, MALIN B, et al. Generating multi-label discrete patient records using generative adversarial networks [C]// Machine Learning for Healthcare Conference. PMLR, 2017: 286-305.

[19] RAJABI A, GARIBAY O O. Tabfairgan: Fair tabular data generation with generative adversarial networks [J]. Machine Learning and Knowledge Extraction, 2022, 4(2): 488-501.

[20] VUTTIPITTAYAMONGKOL P, ELYAN E. Neighbourhood-based undersampling approach for handling imbalanced and overlapped data[J]. Information Sciences, 2020, 509: 47-70.

[21] BUNKHUMPORNPAT C, SINAPIROMSARAN K. DBMUTE: density-based majority under-sampling technique[J]. Knowledge and Information Systems, 2017, 50: 827-850.

[22] FU G H, WU Y J, ZONG M J, et al. Feature selection and classification by minimizing overlap degree for class-imbalanced data in metabolomics[J]. Chemometrics and Intelligent Laboratory Systems, 2020, 196: 103906.

[23] OMAR B, RUSTAM F, MEHMOOD A, et al. Minimizing the overlapping degree to improve class-imbalanced learning under sparse feature selection: application to fraud detection[J]. IEEE Access, 2021, 9: 28101-28110.

[24] LI F, WANG B, SHEN Y, et al. An overlapping oriented imbalanced ensemble learning algorithm withweighted projection clustering grouping and consistent fuzzy sample transformation [J]. Information Sciences, 2023, 637: 118955.

[25] JIANG H X, JIANG J Y, LIANG X. Review on Fraud Detection of Credit Card Transactions Based on Machine Learning[J/OL]. Computer Engineering and Applications: 1-29. [2023-06-03]. http://kns.cnki.net/kcms/detail/11.2127.tp.20230424.

1411.014. html.

[26] XUAN S,LIU G,LI Z,et al. Random forest for credit card fraud detection[C]// 2018 IEEE 15th International Conference on Networking,Sensing and Control(ICNSC). IEEE,2018:1-6.

[27] MENG C,ZHOU L,LIU B. A case study in credit fraud detection with SMOTE and XGboost[C]// Journal of Physics:Conference Series. IOP Publishing,2020:052016.

[28] FU K,CHENG D,TU Y,et al. Credit card fraud detection using convolutional neural networks[C]// 23rd International Conference Neural Information Processing:(ICONIP 2016)Kyoto,Japan,Part III 23. Springer International Publishing,2016:483-490.

[29] BAHNSEN A C,AOUADA D,STOJANOVIC A,et al. Feature engineering strategies for credit card fraud detection[J]. Expert Systems with Applications,2016,51:134-142.

[30] CHEN J I Z,LAI K L. Deep convolution neural network model for credit-card fraud detection and alert[J]. Journal of Artificial Intelligence,2021,3(2):101-112.

[31] CARCILLO F,LE BORGNE Y A,CAELEN O,et al. Combining unsupervised and supervised learning in credit card fraud detection [J]. Information Sciences,2021,557:317-331.

[32] ARJOVSKY M,CHINTALA S,BOTTOUL. Wasserstein GAN [OL]. https://proceedings. mlr. press/v70/arjovsky17a. html.

[33] LIN Z,KHETAN A,FANTI G,et al. Pacgan:The power of two samples in generative adversarial networks [J/OL]. Advances in Neural Information Processing Systems, 2018, 31. https://xplorestaging. ieee. org/document/9046238.

[34] SANTURKAR S,TSIPRAS D,ILYAS A,et al. How does batch normalization help optimization? [J/OL]. Advances in Neural Information Processing Systems,2018,31. https://proceedings. neurips. cc/paper/2018/hash/905056c1ac1dad141560467e0a99e1cf-Abstract. html.

[35] HUIJBEN I A M,KOOL W,PAULUS M B,et al. A review of the gumbel-max trick and its extensions for discrete stochasticity in machine learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2022,45(2):1353-1371.

[36] CHENG G,PEDDINTI V,POVEY D,et al. An Exploration of Dropout with LSTMs[C]// Interspeech. 2017:1586-1590.

ZHUO Peiyan, born in 1999, postgraduate. Her main research interests include data mining and financial technology.

SONG You, born in 1973, professor, Ph.D supervisor. His main research interests include data analysis techniques, financial technology, information processing, and knowledge graph.