文章编号:1007－6735(2025)04－0461－10　DOI: 10.13255/j.cnki.jusst.20240626003

基于可解释不平衡数据分类方法的加密货币交易欺诈检测

尹　裴[1,2]，　蒋文龙[1]，　徐　岩[3]

（1. 上海理工大学 管理学院，上海 200093；2. 上海理工大学 智慧应急管理学院，上海 200093；
3. 北京市西城经济科学大学 基础系，北京 100120）

摘要：鉴于正常和欺诈交易数据在加密货币交易样本中分布极端不平衡，以及交易数据的高维特征和非线性关系，提出了一种用于加密货币交易欺诈检测的可解释不平衡数据分类方法。首先，使用 synthetic minority oversampling technique（SMOTE）过采样和对比学习的数据增强策略来平衡数据；接着，引入基于 transformer 的深度学习模型，学习样本相关性，并通过基于对比损失的预训练和基于贝叶斯优化的微调策略来优化模型，更好地区分正常和欺诈交易的特征分布，提取与欺诈相关的高阶、高维特征；最后，设计基于 shapley additive explanations（SHAP）的解释器，结合注意力分数对模型预测进行解释，揭示不同交易特征在欺诈检测中的作用。对比实验结果表明，该模型在召回率方面表现出色，能全面识别加密货币交易中的欺诈活动。同时，在 F_1 值上达到最佳，很好地平衡了准确率和召回率。消融实验验证了所提出的数据平衡和预训练–微调策略的必要性，说明其能有效处理不平衡数据的分类问题。研究不仅丰富了金融欺诈检测的研究体系，还增强了加密货币的交易安全性，促进市场健康发展，维护经济稳定与社会安全。

关键词：加密货币；欺诈检测；数据分布极端不平衡；可解释

中图分类号：TP 183　　**文献标志码：**A

Cryptocurrency transaction fraud detection based on imbalanced data classification method with interpretable analysis

YIN Pei[1,2]，　JIANG Wenlong[1]，　XU Yan[3]

(1. *Business School, University of Shanghai for Science and Technology, Shanghai 200093, China*; 2. *School of Intelligent Emergency Management, University of Shanghai for Science and Technology, Shanghai 200093, China*; 3. *Foundation Department, Beijing Xicheng College of Economic Science, Beijing 100120, China*)

Abstract: Given the extreme imbalance in the distribution of normal and fraudulent transactions data in

收稿日期: 2024−06−26

基金项目: 教育部人文社会科学研究项目（23YJCZH281）；国家自然科学基金资助项目（72301136）；信息网络安全公安部重点实验室开放课题资助项目（C23600）

通信作者: 尹　裴（1987—），女，博士，副教授．研究方向：机器学习、推荐系统．E-mail：pyin@usst.edu.cn

引文格式: 尹裴，蒋文龙，徐岩．基于可解释不平衡数据分类方法的加密货币交易欺诈检测[J]．上海理工大学学报，2025，47(4)：461-470.

Citation: YIN Pei, JIANG Wenlong, XU Yan. Cryptocurrency transaction fraud detection based on imbalanced data classification method with interpretable analysis[J]. Journal of University of Shanghai for Science and Technology, 2025, 47(4): 461-470.

cryptocurrency transaction samples, as well as the complex high-dimensional features and nonlinear relationships in transaction data, an imbalanced data classification method with interpretable analysis for cryptocurrency transaction fraud detection was proposed. Firstly, synthetic minority oversampling technique(SMOTE) and contrastive learning-based data enhancement strategies were used to balance the data. Secondly, a deep learning model based on transformer was introduced to learn sample relevance. Pre-training based on contrastive loss and fine-tuning strategies based on Bayesian were used to optimize the model. It distinguished the distribution of features between normal and fraudulent transactions bertter, extracted higher-order, high-dimensional features related to fraud. Finally, the interpreter was designed based on shapley additive explanations(SHAP), and combined with attention scores for model predictions interpretaion to reveal the role of different transaction features in fraud detection. Comparative results show that the model performs well in recall and can fully identify fraudulent activities in cryptocurrency transactions. Meanwhile, it achieves an excellent F_1 value, striking a balance between accuracy and recall. Ablation experiments verify the necessity of the proposed data balancing and pre-training-fine-tuning strategies, show that it can effectively deal with the classification problem of unbalanced data. The research not only enriches financial fraud detection but also enhances cryptocurrency transaction security, promotes market development, and contributes to economic stability and social security.

Keywords: *cryptocurrency*; *fraud detection*; *extremely imbalanced data classification*; *interpretable analysis*

自2008年首次提出加密货币以来，该市场迅猛增长，数字货币如比特币和以太坊因其去中心化特性而备受推崇。但这些特性也为洗钱、网络钓鱼等欺诈活动提供了机会。相比于传统货币，加密货币交易规模庞大且监管不足，匿名性和智能合约的复杂性使欺诈检测变得困难。近年来，基于机器学习的金融欺诈研究取得进展[1]，深度学习模型，尤其是神经网络，因其高准确性受到关注[2]。尽管图神经网络提高了欺诈检测的准确率，但仍需更全面地考虑用户和交易间的相似性及特征间的相互作用[3]。

加密货币领域的欺诈检测研究仍处于初期，区块链的独特性质使得数据特征复杂，难以准确识别欺诈特征。区块链的去中心化导致正常交易与欺诈交易比例失衡，市场波动也可能使正常价格波动被误判为欺诈。此外，平衡精确度与召回率至关重要，优先识别欺诈交易比追求准确率更为重要。欧盟《General Data Protection Regulations》等法规也要求对检测结果进行解释，以提高模型透明度。

本文旨在解决上述问题，并探讨加密货币交易欺诈检测是否应视为分类任务，以及如何处理高度不平衡数据。本文创新性地结合了SMOTE（synthetic minority oversampling technique）过采样和对比学习，通过生成少数类的合成样本来改善类别不平衡，并通过引入正负样本对，使模型学习到更具区分度的特征表示。为了更好地捕捉欺诈交易的复杂特征，采用了基于对比损失的预训练策略，增强了模型对细微特征的敏感度。引入贝叶斯优化进行模型微调，提高了超参数选择的效率和精度。最后，通过结合注意力机制和SHAP（shapley additive explanations）方法提升了模型可解释性。注意力机制聚焦于与欺诈相关的重要特征，而SHAP则揭示了特征在欺诈识别中的影响力，提高了结果的可信度和透明度。

1 相关工作

1.1 金融欺诈检测

金融欺诈在金融领域和日常生活中可能造成严重影响，因此大量文献研究了多种欺诈类型，如财务报表欺诈、信用卡欺诈、保险欺诈和加密货币欺诈等。早期研究主要采用基于规则的方法进行欺诈检测，认为欺诈行为有明显模式，因

此，制定许多规则和静态阈值来过滤欺诈行为。例如，如果资金流动或盈利能力等重要指标异常高或低，系统将发出警报[4]。尽管这些方法简单易懂，但高度依赖专家知识，难以捕捉复杂、变化的模式，一旦攻击者了解规则，就容易被绕过[5]。随着欺诈者使用更复杂的策略，检测需要分析大量金融信息，人工总结的小规则集已不足以满足需求，而管理大型规则集变得困难且资源消耗大[6]。

鉴于基于规则方法的局限性，近年来的研究转向使用机器学习模型自动发现数据中的潜在欺诈模式。通常，这些方法提取用户的统计特征，并使用传统分类器如支持向量机（support vector machine，SVM）、树模型或神经网络进行欺诈判断。例如，Sehrawat 等[7] 使用长短期记忆网络（long short-term memory，LSTM）和门控循环单元（Gated Recurrent Unit，GRU）神经网络的自动编码器检测信用卡欺诈，自动编码器通过无监督学习提取数据表示，并与类标签结合，作为后续模型的输入。Ajitha 等[8] 将卷积神经网络（convolutional neural network，CNN）与其他机器学习算法的性能进行了比较，Mizher 等[9] 提出了基于卷积神经网络和机器学习算法的信用卡欺诈检测模型，结果显示，随机森林（random forest，RF）和 CNN 分别达到了 99.7% 和 93.5% 的准确率。

近年来，许多研究者认识到金融领域中实体之间存在的相互关联，开始利用图结构进行欺诈检测。例如，Wang 等[10] 提出的 FdGars 是用于在线应用审核的图卷积网络方法，Li 等[11] 提出的 GAS 集成了异构和同构图以捕获本地和全局上下文，Liu 等[12] 提出的 GraphConsis 研究了基于图的欺诈检测中的一致性问题。这些研究表明，图结构在特定欺诈检测任务中表现良好[13]。

1.2　针对加密货币的金融欺诈检测

在数字时代，加密货币和区块链技术正在改变全球经济。交易所越来越依赖数字货币而非传统货币，这主要得益于区块链的去中心化和无需第三方参与[14]。区块链作为公共账本，因去中心化、不变性和匿名性被广泛使用[15]。虽然这些特性提升了加密货币的可信度，但也为洗钱、网络钓鱼、庞氏骗局和首次代币发行（initial coin offering，ICO）骗局等提供了滋生土壤，推动了欺诈检测的研究。

加密货币与传统货币在欺诈检测上有显著差异[16]。加密交易的匿名性使欺诈者更难追踪，而传统金融体系要求身份验证和可追溯的交易历史。加密货币涉及复杂的技术，如智能合约，欺诈者可能利用这些技术进行隐蔽的欺诈。此外，加密市场波动性大，正常价格波动可能被误判为欺诈，这增加了检测难度。

庞氏骗局在加密货币中很普遍[17]。Hu 等[18] 提出了 SCSGuard，一个深度学习框架，用于分析智能合约代码以检测骗局。洗钱检测模型通常利用账户特征，将交易分为非法和正常[19]。对于网络钓鱼，Yuan 等[20] 构建了交易网络数据集，使用 node2vec 提取特征，并用 SVM 将账户分类。然而，提高模型准确性时可解释性往往下降，因此解释模型决策变得愈发重要[21]。

1.3　研究评述

综合上述，现有加密货币交易欺诈检测研究还存在以下不足：

a. 大多数学者认为加密货币交易欺诈检测是一个分类问题。这是因为在加密货币交易欺诈检测中，通常可以获得已知标签的数据，因此可以使用监督学习的方法进行分类，而无需依赖数据中的异常模式。

b. 早期的人工规则和特征工程方法受到专家知识的限制，无法捕捉复杂多变的模式。传统的机器学习模型不足以处理更大的数据量和更复杂的关系结构。现有的深度学习模型需要更多地考虑用户和交易之间的相似性。

c. 在处理像加密货币交易这样极度不平衡的分类问题时，研究还具有挑战性。

d. 随着深度学习模型在欺诈检测方面的提升，解释模型变得更为困难。因此，迫切需要开发方法来解释深度学习“黑盒”模型，提高模型的透明度。

2　模　型

针对这些问题，本文提出了一种可解释不平衡数据分类方法，用于识别加密货币平台上的欺诈交易。首先，利用对比学习中的 SMOTE 过采样和数据增强技术来实现数据平衡。然后，引入了基于 transformer 的深度学习模型 SAINT[22]，该模型专注于样本间注意力，并通过设计对比损失函数进行预训练，从而通过区分正常样本和欺诈样

本，更好地提取出与欺诈真正相关的高阶特征，应用贝叶斯优化方法对预训练模型进行微调，进一步提高欺诈识别效果。最后，结合注意力得分和SHAP方法对模型进行解释，计算出每个特征对欺诈检测的贡献。

2.1 基于SMOTE与随机噪音的少数类样本平衡方法

2.1.1 数据编码

假设$D=\{\boldsymbol{X}_i,\boldsymbol{Y}_i\}_{i=1}^{k}$是有$k$个样本数的交易数据，式中：$\boldsymbol{X}_i$为$n$维的特征向量，为过采样后欺诈样本和正常样本的集合；$\boldsymbol{Y}_i$为标记为正常或欺诈的标签。同时，将[CLS]令牌作为嵌入加到每个样本的特征向量中，则$\boldsymbol{X}_i$的特征向量可表示为$\boldsymbol{X}_i=\{[\mathrm{CLS}],a_1,a_2,\cdots,a_n\}$。

2.1.2 SMOTE过采样

SMOTE算法通过增加少数类别的样本以平衡数据集中的类别分布，同时保持数据的拓扑结构信息。

对于每个少数类的欺诈样本，计算它的k个最近邻少数类别样本，然后从它的k个最近邻样本中随机选择一个样本。则新生成的欺诈样本可表示为

$$\boldsymbol{X}_{\mathrm{fr(new)}}=\boldsymbol{X}_{\mathrm{fr}}+\mathrm{rand}(0,1)\,d(\boldsymbol{X}_{\mathrm{fr}},\boldsymbol{X}_{\mathrm{fr}}^{'}) \tag{1}$$

式中：$\mathrm{rand}(0,1)$为从[0, 1]的均匀分布中随机选择的一个值；$d(\boldsymbol{X}_{\mathrm{fr}},\boldsymbol{X}_{\mathrm{fr}}^{'})$为所选样本$\boldsymbol{X}_{\mathrm{fr}}$与它的一个近邻样本$\boldsymbol{X}_{\mathrm{fr}}^{'}$之间的差值。将生成的合成样本添加到数据集中，使得数据集中的少数欺诈样本数量增加。

2.1.3 添加噪音

本文考虑到使用SMOTE进行过采样后的数据其结构较为单一，因此为增加样本多样性，采用添加正态分布的随机噪音的方法来模拟现实生活中真实的欺诈样本数据。噪音添加的具体方法如下：

$$\boldsymbol{X}_{2i}=\boldsymbol{X}_{\mathrm{fr(new)}}+\boldsymbol{\sigma} \tag{2}$$

式中：$\boldsymbol{\sigma}$为正态分布的随机噪音；$i=1,2,\cdots,k/2$。

则最后的仿真数据集可表示为

$$D=\{\boldsymbol{X}_i,\boldsymbol{Y}_i\}_{i=1}^{k} \tag{3}$$

2.2 基于对比学习的预训练方法

2.2.1 数据增强

采用对比学习中的数据增强策略做进一步的数据平衡。首先，使用CutMix对原始输入数据进行特征的随机混合；然后，使用mixup对经过嵌入层的向量表示进行随机混合。两种方法的结合能够更有效地提升样本多样性，从而提高整体对比学习的效果。

图1中E表示嵌入层，S表示SAINT，M_1和M_2表示两个作为投影头的全连接层。将CutMix的增强概率记为p，mixup的参数记为α，对于原始输入$\boldsymbol{X}_i$，其经过嵌入层后可记为$\boldsymbol{e}_i$=E($\boldsymbol{X}_i$)，则数据增强后表示为

$$\boldsymbol{X}^{'}=\boldsymbol{X}_i\boldsymbol{V}+\boldsymbol{X}_{\mathrm{a}}(1-\boldsymbol{V}) \tag{4}$$

$$\boldsymbol{e}_i^{'}=\alpha\mathrm{E}(\boldsymbol{X}_i^{'})+(1-\alpha)(\mathrm{E}(\boldsymbol{X}_{\mathrm{b}}^{'})) \tag{5}$$

式中：$\boldsymbol{X}_{\mathrm{a}}$、$\boldsymbol{X}_{\mathrm{b}}$为当前批次的随机样本；$\boldsymbol{X}_{\mathrm{b}}^{'}$为$\boldsymbol{X}_{\mathrm{b}}$经

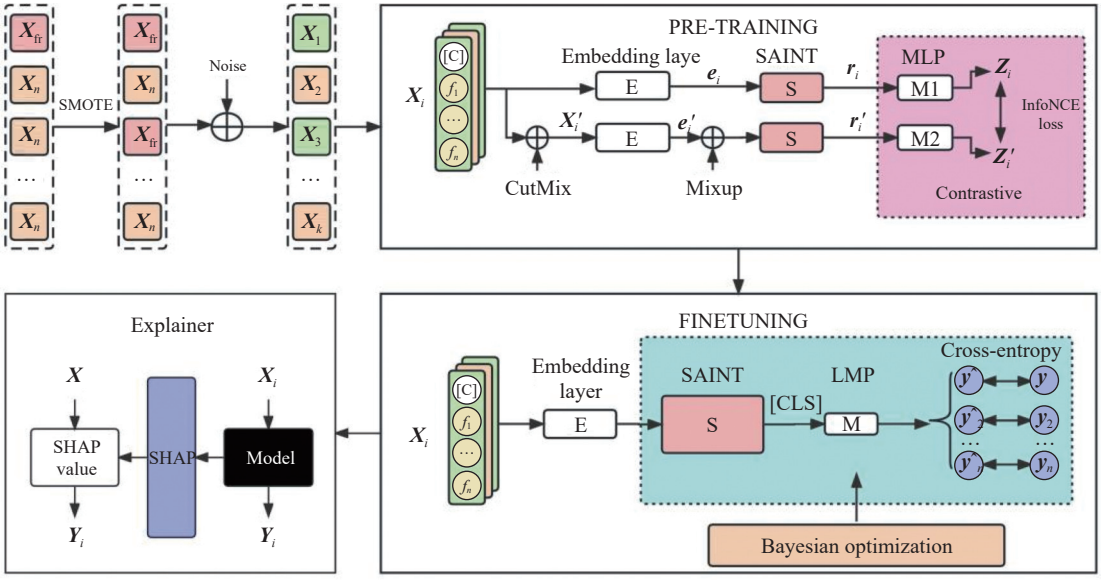

图1 整体流程架构图

Fig.1 Diagram of the overall process architecture

过 CutMix 的版本；$\boldsymbol{V}$为从概率为p的伯努利分布中采样的二值掩码向量，α为 mixup 的参数。首先，通过随机选择要混合的原始样本来获得同一批次中每个数据点的 CutMix 版本。然后，在通过嵌入层后的样本中选择新的混合样本进行 mixup。

在获得未做数据增强的嵌入表示$\boldsymbol{e}_i$和经过数据增强混合后的嵌入表示$\boldsymbol{e}_i'$后，将它们分别输入到 SAINT 模型中，得到$\boldsymbol{r}_i$、$\boldsymbol{r}_i'$，然后输入两个作为投影头的全连接层进行降维，每个全连接层都由一个隐藏层和 ReLU 激活函数所组成。最后，对降维后的两组向量计算对比损失。

2.2.2 Transformer 变体−SAINT 模型

SAINT 是 transformer 编码器的变体，模型由g个相同的模块组成，每个模块包括一个自注意力模块和一个样本间注意力模块。自注意力模块与 transformer 编码器中的相同，而样本间注意力模块与自注意力模块相似。具体来说，样本间注意力就是将单个样本的每个特征的嵌入连接起来，然后计算样本（而不是特征）的注意力，这使得模型能够通过对比其他样本来改进当前样本的表示。

2.2.3 对比损失函数

在预训练 SAINT 模型时，采用度量学习中的 InfoNCE 损失，将同一样本得到的两个不同向量$\boldsymbol{Z}_i$和$\boldsymbol{Z}_i'$进行比较，并缩减它们的距离，对于不同样本$\boldsymbol{Z}_i$和$\boldsymbol{Z}_j$（i不等于j），则使它们相互远离，损失函数为

$$L_{\text{pre\_training}} = -\sum_{i=1}^{k} \log \frac{\exp(\boldsymbol{Z}_i \boldsymbol{Z}_i'/\tau)}{\sum_{m=1}^{k} \exp(\boldsymbol{Z}_i \boldsymbol{Z}_m'/\tau)} \tag{6}$$

式中：$\boldsymbol{Z}_i$=M$_1$($\boldsymbol{r}_i$)，$\boldsymbol{Z}_i'$=M$_2$($\boldsymbol{r}_i'$)；τ为温度参数，用于调节相似性分数的尺度。

2.3 基于贝叶斯优化的微调方法

如图 1 所示，所有未带标签的数据完成在 SAINT 上的预训练后，再使用带有标签的数据对 SAINT 模型进行微调。

通过使用贝叶斯优化来对模型的超参数进行搜索优化，包括学习率、学习步长、迭代次数和批次大小等，以获得预测模型的最高性能。贝叶斯优化分为 2 步，第 1 步是构建一个高斯过程代理模型，该模型用于估计目标函数的潜在分布。计算公式如下：

$$\mu_{\text{new}} = m(\theta_{\text{new}}) + K(\theta_{\text{new}}, \theta)(\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1}(a - m(\theta)) \tag{7}$$

$$\sigma_{\text{new}}^2 = K(\theta_{\text{new}}, \theta_{\text{new}}) - K(\theta_{\text{new}}, \theta)(\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} K(\theta, \theta_{\text{new}}) \tag{8}$$

式中：μ_{new}为在新的采样点上的预测均值，即模型对目标函数值的估计；$m(\cdot)$为先验均值函数；$K(\cdot)$为协方差函数；θ_{new}为最新的采样点；θ为已知的采样点集合；$(\theta_{\text{new}}, \theta)$和$(\theta, \theta_{\text{new}})$均为新采样点和已知采样点采集的样本对；$\boldsymbol{K}$为协方差矩阵；$\sigma^2$为噪声的方差；$\boldsymbol{I}$为单位矩阵；$a$为对应采样点的函数值；$\sigma_{\text{new}}^2$为在新的采样点上的预测方差，用于衡量预测的不确定性；$(\theta_{\text{new}}, \theta_{\text{new}})$为新采样点采集的样本对。

第 2 步使用采集函数对新样本进行采样，计算公式如下：

$$C(\theta_i) = \mu(\theta_i) + \beta \sqrt{\frac{\log(t)}{n(\theta_i)}} \tag{9}$$

式中：$C(\theta_i)$为采集函数值，它衡量在点θ_i处的“采样价值”，贝叶斯优化会选择使采集函数值最大化的位置作为下一个采样点；$\mu(\theta_i)$为高斯过程的预测均值；β为一个控制探索与利用权衡的参数；$\log(t)$为当前迭代步骤数的自然对数；$n(\theta_i)$为采样点θ_i被选择的次数。重复上述步骤，直到达到预定的迭代次数，最后得到在这些迭代次数中最优的超参数集合θ^*。

对于原始输入数据，通过嵌入层和 SAINT 得到包含全局信息的向量表示，然后将向量中的 [CLS] 令牌作为嵌入通过一个全连接层做最后的预测，并将预测值与真实值进行对比，然后使用交叉熵损失来对欺诈样本和正常样本的分类效果进行评估。

2.4 基于 SHAP 的解释器

采用 SHAP方法对欺诈检测模型进行解释，过 Shapley 值量化特征在欺诈检测任务中的边际贡献，其计算公式如下：

$$\phi_l(f) = \sum_{S \in N \setminus \{l\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{l\}) - f(S)] \tag{10}$$

式中：$\phi_l(f)$为特征l的 SHAP 值；f为模型的预测函数；N为特征的总数；S为特征索引的一个子集，表示不包括特征l的特征组合；$f(S \cup \{l\})$为包括特征l的特征组合的模型预测；$f(S)$为不包括特征l的特征组合的模型预测。对于特征l的 SHAP 值，对所有不包括特征l的特征组合进行考虑，计算加入特征l后和不加入特征l时模型预测的差异，

并考虑所有可能的组合权重。

3 实验设置与评价

3.1 实验数据

使用Kaggle提供的以太坊交易欺诈检测数据（Ethereum欺诈检测数据集(kaggle.com)），由于数据是有标签的，因此，同样将加密货币交易欺诈检测任务定义为二元分类问题。数据集中的特征能充分反映加密货币交易中的欺诈行为。比如，交易行为特征包括交易金额、时间间隔、交易频率等，数据集中的特征能够广泛涵盖潜在的欺诈活动模式，该数据集对本文中的欺诈检测任务具有较强的支持性和代表性。数据结构如表1所示。

表1 以太坊交易数据结构

Tab.1 Data structures of ethernet transaction

	总样本数	正常样本数	欺诈样本数	比例
过采样前	12147	11232	915	12∶1
过采样后	22464	11232	11232	1∶1

还使用了Elliptic++数据集[23]中的比特币交易数据进行欺诈检测，该数据由总共203370笔交易组成。该数据集除交易行为特征外，还包括交易关系特征，对本文中的欺诈检测任务更具有支持性和代表性。部分特征的数据样本如表2所示。

表2 比特币交易数据样本

Tab.2 Sample of Bitcoin trading data

ID号	时间步	局部特征_1-93	聚合特征_1-72	比特币总额	费用/比特币	大小	类别
3205536	1	…	…	3.44990833	0.0002145	1257	1
16742787	1	…	…	0.0371971	0.001	225	1
16753577	1	…	…	0.0856	0.001	372	1
16754007	1	…	…	0.61417099	0.001	372	1
17387772	1	…	…	1.38101431	0.001	521	1

本文按8∶1∶1的比例划分数据为训练集、验证集和测试集。首先，训练集进行预训练以学习高阶特征；然后，通过训练集和验证集进行微调，将验证集准确率作为贝叶斯优化的目标函数，以提升模型在验证集上的性能；最后，使用测试集评估模型的检测性能。

3.2 评价方法

采用了P、R和F_1作为模型的评价指标：

$$P=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}} \tag{11}$$

$$R=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}} \tag{12}$$

$$F_1=\frac{2PR}{P+R} \tag{13}$$

式中：T_{P}为被正确预测为正类的样本；F_{P}为被错误预测为正类的样本；F_{N}为实际为正样本而未被预测为正的样本。

4 实验结果分析

4.1 敏感性实验

通过敏感性实验寻找模型的最佳参数，并且通过相同预训练与微调的步骤，实验结果如表3所示。

对于本文使用的SAINT模型，最优的模型参数为：自注意力的层数设置为4，注意力头设置为8，输入维度设置为32。在对模型进行预训练时，优化器选择AdamW，学习率为0.0001，批次大小为64。在对模型进行微调时，经过100次的贝叶斯优化实验后，得到了一组使模型性能表现最好的超参数：学习率为0.00069，迭代次数为20，批次大小为64，学习步长为5，优化器为AdamW。

表3 敏感性实验结果

Tab.3 Results of sensitivity experiments

层数	注意力头	输入维度	P	R	F_1
2	4	16	0.7495	0.8502	0.7966
4	8	32	0.8473	0.927	0.8854
8	16	64	0.7611	0.8653	0.8099

4.2 消融实验

为了考察不同模块对检测结果的影响，分别比较不同模块组合，并通过实验结果来判断其影响，实验结果如图2所示。

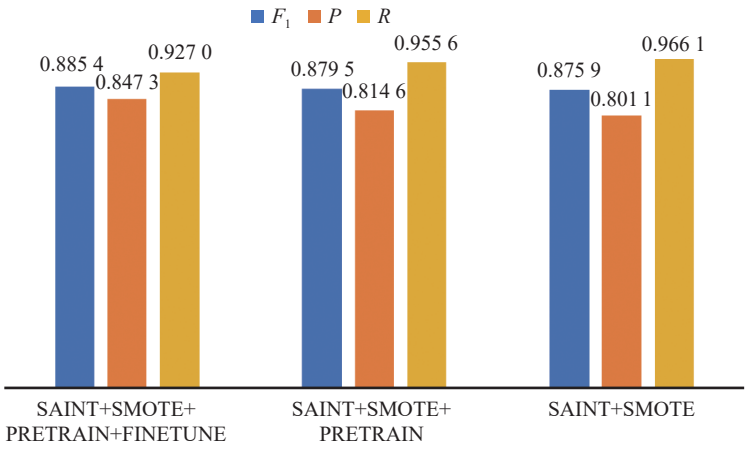

图 2　消融实验结果柱状图

Fig.2　Histogram of ablation experiment results

从图 2 可以看出，完整模型的召回率很高，表明它能有效地捕捉欺诈行为并减少漏报。同时，高 F_1 值表明召回率和精确度之间达到了平衡。

去掉预训练和微调后，模型能够更准确地预测欺诈样本，从而提高召回率。但是，该模型错误地将更多正常样本预测为欺诈样本，从而降低了精确率。此外，F_1 值下降，表明召回率和精确率之间的平衡被打破，整体性能下降。因此，本文提出的预训练策略和微调策略的必要性和有效性得到了证明。

4.3　对比实验

本文比较的基线方法如下：

LightGBM：是一种梯度提升框架，是由微软开发的一种高效的机器学习算法。

RF：是一种集成学习方法，用于解决分类和回归问题。

XGBoost：是一种梯度提升算法，通过组合多个弱学习器来构建一个强大的集成模型解决分类和回归问题。

TabTransformer：是一种用于表格数据的深度学习模型，旨在处理表格数据的结构化信息。

对所有比较方法在加密货币交易欺诈检测任务中的性能进行评估，相应的 P、R 和 F_1 值如表 4 和表 5 所示。可以看出，一些机器学习模型的 P 指标较高，而 R 指标较低，这说明模型在所有预测的欺诈样本中预测出了大部分正确样本，但预测出的欺诈样本数量很少；综上所述，SAINT 的总体性能优于上述 4 种基线模型，验证了以太币和比特币加密货币交易欺诈检测性能良好。

为了使模型取得更好的效果，通过对比学习进行预训练来提高模型性能，结果如表 5 所示。可以看出，模型的整体性能提高了约 0.4%，这说明对比学习在应用于表格数据时也能发挥良好的作用。本文在对模型进行微调时，采用贝叶斯优化来寻找最优超参数，可以看出，经过贝叶斯优化后，模型的整体性能提高了近 0.6%。

表 4　以太坊交易欺诈检测性能比较

Tab.4　Comparison of Ethernet transaction fraud detection performance

基线模型	P	R	F_1
LightGBM	0.672 3	0.506 7	0.577 9
RF	0.877 6	0.637 6	0.738 6
XGBoost	**0.966 4**	0.356 1	0.520 5
TabTransformer	0.733	**0.978 3**	0.838 1
本文	0.847 3	0.927	**0.885 4**

表 5　比特币交易欺诈检测性能比较

Tab.5　Comparison of Bitcoin transaction fraud detection performance

基线模型	P	R	F_1
LightGBM	0.632 5	0.489	0.551 3
Random Forest	**0.975**	0.719	0.828
XGBoost	0.793 1	0.718 2	0.754 1
TabTransformer	0.701 0	**0.924 3**	0.797 1
本文	0.824 1	0.895 4	**0.858 2**

通过实验得到以下结论：

机器学习在面对大量且关系结构复杂的数据时，性能通常没有深度学习的好。同是深度学习模型的 TabTransformer 所表现出的性能没有 SAINT

好，说明了样本间注意力在欺诈检测任务中，通过对比欺诈和正常样本间的相似度，从而提升了模型识别欺诈样本的能力。

通过对比学习的预训练，模型学习到了欺诈样本与正常样本的整体特征表示，从而使模型在下游分类任务中的检测性能提高。并且，搜索超参数来优化模型是一个很重要的步骤，相比手动调参和网格搜索等，贝叶斯优化在效率更高的同时，还能达到更好的效果。

4.4 可解释分析

4.4.1 基于注意力机制的可解释分析

本文基于 Transformer 模型计算的注意力分数，分析了欺诈预测过程中交易特征的影响，并选取前 5 个显著特征进行解释性分析，如表 6 所示。模型发现，新用户注册、买入货币数量异常高、采用多次小额交易规避监测、买入与卖出不匹配以及短暂活跃天数，对于欺诈预测具有显著影响。

4.4.2 基于 SHAP 的可解释分析

在本文研究中，使用 SHAP 解释框架对欺诈预测的交易特征进行了高阶分析。图 3 所示的 SHAP 蜂群图将每个特征和样本的 SHAP 值可视化，以直观地解读模型输出。图 3 显示了影响欺诈预测的 5 个最关键特征：接收方交易之间的时间间隔、账户活动持续时间、总交易次数、账户注册时长和平均货币销售价值。主要观察结果为：

表 6 一个欺诈样本的特征值和注意力权重（前 5 个）

Tab.6 Eigenvalues and attention weights for a sample of frauds (top 5)

特征	值	权重
账户注册时长	776	0.036 341 2
基尼系数	0.852 56	0.035 214 56
每日最大交易额/比特币	12	0.035 204 89
买卖平衡比	−1.546	0.033 307 28
活跃天数	122	0.031 768 18

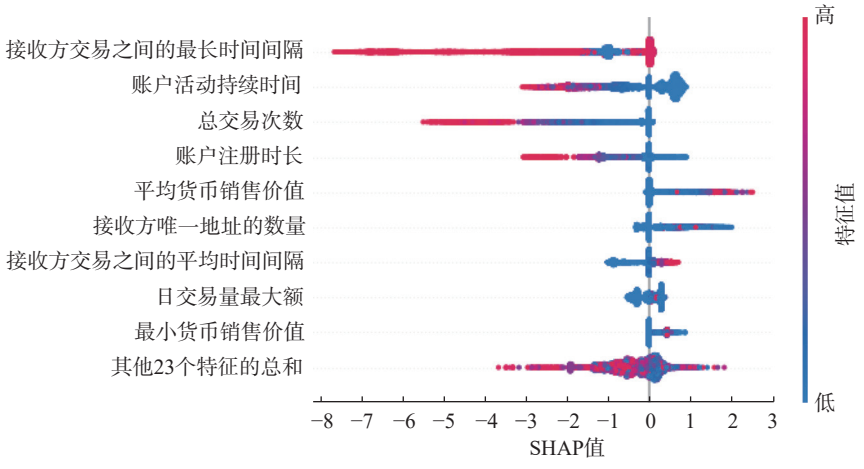

图 3 SHAP 蜂群图

Fig.3 SHAP swarm diagram

a. 接收方交易之间的时间间隔较长通常表示活动正常。

b. 账户活动持续时间、总交易次数和账户注册时长与欺诈呈反向关系；活跃的长期账户不太可能从事欺诈活动。

c. 平均货币销售价值越高，模型中的 SHAP 值就越高，这可能表明存在欺诈行为。正常交易主要涉及货币购买，因此大规模卖出是不正常的。

模型主要从以下几个方面识别加密货币交易中的欺诈行为：

a. 欺诈通常发生在活跃度较低的新用户交易中。这些用户可能还不太了解加密货币的工作原理，因此更容易成为欺诈者的目标。

b. 欺诈者的交易通常具有以下特点：单次金额不高，但交易频率高且交易总额较高。他们故意减小单笔交易金额，以避免引起其他用户或机构的注意。高频交易帮助他们实现更高的总交易额。

c. 大多数用户的交易中，买入货币总量通常多于卖出货币，因为加密货币市值不断上升。然而，在欺诈交易中，卖出货币的总量通常少于买入货币，这表明用户受到欺诈手段，将货币卖给不法分子。

5　结　论

为应对加密货币交易欺诈检测的复杂性，提出了一种用于加密货币交易欺诈检测的可解释不平衡数据分类方法。该方法采用了 SMOTE 过采样和对比学习的数据增强策略，以及对比损失的预训练和基于贝叶斯优化的微调策略，以提取与欺诈相关的高阶和高维特征，从而更好地将欺诈样本与正常样本区分开来。最后，设计了一个基于 SHAP 的解释器，结合注意力分数，对模型的预测进行解释。

通过实验得出，本文提出的模型在召回率方面表现出色，同时，在 F_1 值上取得最佳表现，并且证明了所提出的数据平衡方法和预训练−微调策略的必要性。此外，可解释性分析表明了各种交易特征在欺诈检测中的重要性。本研究不仅丰富了金融欺诈检测的研究体系，还推动了加密货币市场的健康发展。但是，该模型还需要在真实的加密货币交易环境中进行更广泛的验证，以确保模型在实际场景中的可行性和有效性。同时，考虑到加密货币市场的多样性，未来的研究可以探讨如何使得模型具有更好的跨平台适用性，以适应不同交易平台和加密货币的特点。

参考文献：

[1] CHANDRADEVA L S, AMARASINGHE T M, DE SILVA M, et al. Monetary transaction fraud detection system based on machine learning strategies[C]//Proceedings of the 4th International Congress on Information and Communication Technology. London: Springer, 2020: 385–396.

[2] SÜRÜCÜ O, YEPREM U, WILKINSON C, et al. A survey on ethereum smart contract vulnerability detection using machine learning[C]//Proceedings of SPIE 12117, Disruptive Technologies in Information Sciences VI. Orlando: SPIE, 2022: 121170C.

[3] KURSHAN E, SHEN H D, YU H J. Financial crime & fraud detection using graph computing: application considerations & outlook[C]//Proceedings of 2020 Second International Conference on Transdisciplinary AI. Irvine: IEEE, 2020: 542–549.

[4] BENEISH M D. The detection of earnings manipulation[J]. Financial Analysts Journal, 1999, 55(5): 24−36.

[5] DAZELEY R P. To the knowledge frontier and beyond: a hybrid system for incremental contextual-learning and prudence analysis[D]. Hobart: University of Tasmania, 2006.

[6] RYMAN-TUBB N F, KRAUSE P, GARN W. How Artificial Intelligence and machine learning research impacts payment card fraud detection: a survey and industry benchmark[J]. Engineering Applications of Artificial Intelligence, 2018, 76: 130−157.

[7] SEHRAWAT D, SINGH Y. Auto-encoder and LSTM-based credit card fraud detection[J]. SN Computer Science, 2023, 4(5): 557.

[8] AJITHA E, SNEHA S, MAKESH S, et al. A comparative analysis of credit card fraud detection with machine learning algorithms and convolutional neural network[C]//Proceedings of 2023 International Conference on Advances in Computing, Communication and Applied Informatics. Chennai: IEEE, 2023: 1–8.

[9] MIZHER M Z, NASSIF A B. Deep CNN approach for unbalanced credit card fraud detection data[C]//Proceedings of 2023 Advances in Science and Engineering Technology International Conferences. Dubai: IEEE, 2023: 1–7.

[10] WANG J Y, WEN R, WU C M, et al. FdGars: fraudster detection via graph convolutional networks in online app review system[C]//Proceedings of 2019 World Wide Web Conference. San Francisco: ACM, 2019: 310–316.

[11] LI A, QIN Z, LIU R S, et al. Spam review detection with graph convolutional networks[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing: ACM, 2019.

[12] LIU Z W, DOU Y T, YU P S, et al. Alleviating the inconsistency problem of applying graph neural network to fraud detection[C]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020.

[13] DOU Y T, LIU Z W, SUN L, et al. Enhancing graph neural network-based fraud detectors against camouflaged fraudsters[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. New York: ACM, 2020.

[14] IANSITI M, LAKHANI K R. The truth about blockchain[J]. Harvard Business Review, 2017, 95(1): 118−127.

[15] THARANI J S, CHARLES E Y A, HÓU Z, et al. Graph based visualisation techniques for analysis of blockchain transactions[C]//Proceedings of 2021 IEEE 46th Conference on Local Computer Networks. Edmonton: IEEE, 2021: 427–430.

[16] YUAN Y, WANG F Y. Blockchain and cryptocurrencies: model, techniques, and applications[J]. IEEE Transactions

on Systems, Man, and Cybernetics: Systems, 2018, 48(9): 1421−1428.

[17] MUKHERJEE S, LARKIN C, CORBET S. Cryptocurrency Ponzi schemes[M]//CORBET S. Understanding Cryptocurrency Fraud. Berlin: De Gruyter, 2021: 111–120.

[18] HU H W, BAI Q L, XU Y D. SCSGuard: deep scam detection for ethereum smart contracts[C]//Proceedings of the IEEE Conference on Computer Communications Workshops. New York: IEEE, 2022: 1–6.

[19] ALARAB I, PRAKOONWIT S, NACER M I. Competence of graph convolutional networks for anti-money laundering in Bitcoin blockchain[C]//Proceedings of the 2020 5th International Conference on Machine Learning Technologies. Beijing: ACM, 2020: 23–27.

[20] YUAN Q, HUANG B Y, ZHANG J, et al. Detecting phishing scams on Ethereum based on transaction records[C]//Proceedings of 2020 IEEE International Symposium on Circuits and Systems. Sevilla: IEEE, 2020: 1–5.

[21] KUTE D V, PRADHAN B, SHUKLA N, et al. Deep learning and explainable artificial intelligence techniques applied for detecting money laundering–a critical review[J]. IEEE Access, 2021, 9: 82300−82317.

[22] SOMEPALLI G, GOLDBLUM M, SCHWARZSCHILD A, et al. SAINT: improved neural networks for tabular data via row attention and contrastive pre-training[DB/OL]. [2021-06-02] https://arxiv.org/abs/2106.01342.

[23] ELMOUGY Y, LIU L. Demystifying fraudulent transactions and illicit nodes in the Bitcoin network for financial forensics[C]//Proceedings of 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Long Beach: ACM, 2023: 3979–3990.

（编辑：董　伟）

（上接第 460 页）

[27] REYNOLDS C W. Flocks, herds and schools: a distributed behavioral model[C]//Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques. New York: Association for Computing Machinery, 1987: 25−34.

[28] LEI L, ESCOBEDO R, SIRE C, et al. Computational and robotic modeling reveal parsimonious combinations of interactions between individuals in schooling fish[J]. PLoS Computational Biology, 2020, 16(3): e1007194.

[29] 刘磊, 葛振业, 林杰, 等. 基于鱼群涌现行为启发的集群机器人硬注意力强化模型 [J]. 计算机应用研究, 2024, 41(9): 2737−2744.

（编辑：丁红艺）