第47卷 第6期 计 算 机 工 程 2021年6月
Vol.47 No.6 Computer Engineering June 2021

·开发研究与工程应用· 文章编号：1000-3428(2021)06-0312-09 文献标志码：A 中图分类号：TP399

面向非平衡数据集的金融欺诈账户检测研究

吕 芳[1,2]，汤丰赫[1,2]，黄俊恒[1,2]，王佰玲[1,2]

（1.哈尔滨工业大学（威海）计算机科学与技术学院，山东 威海 264209；
2.哈尔滨工业大学（威海）网络空间安全研究院，山东 威海 264209）

摘 要：针对非平衡金融数据集，提出一种银行欺诈账户检测框架iForest-SMOTE。基于账户的动态交易特点，从统计、时序、监督信息维度抽取账户交易行为特征。针对过采样技术ADASYN在金融账户数据集中存在的跨区域样本合成问题，提出一种基于iForest算法的数据集均衡预处理策略，通过iForest算法对数据进行混合采样，在去除多数类噪声数据的同时降低分类器对少数类的学习难度。在此基础上，设计随机森林分类器实现金融欺诈账户检测。在真实金融账户交易数据集上进行实验，结果表明，与ADASYN、SMOTE等采样技术相比，iForest-SMOTE在召回率和准确率方面具有明显优势，F-value值至少能够提升2.13个百分点。

关键词：隔离森林；非平衡分类；欺诈账户检测；随机森林；特征挖掘

开放科学（资源服务）标志码（OSID）：

中文引用格式：吕芳，汤丰赫，黄俊恒，等.面向非平衡数据集的金融欺诈账户检测研究[J].计算机工程，2021，47(6)：312-320.

英文引用格式：LÜ Fang，TANG Fenghe，HUANG Junheng，et al.Study on financial fraud account detection based on imbalanced datasets[J].Computer Engineering，2021，47(6)：312-320.

Study on Financial Fraud Account Detection Based on Imbalanced Datasets

LÜ Fang[1,2]，TANG Fenghe[1,2]，HUANG Junheng[1,2]，WANG Bailing[1,2]

（1.School of Computer Science and Technology，Harbin Institute of Technology（Weihai），Weihai，Shandong 264209，China；
2.Research Institute of Cyberspace Security，Harbin Institute of Technology（Weihai），Weihai，Shandong 264209，China）

【Abstract】For the detection of bank accounts involved in fraud，this paper proposes a framework，iForest-SMOTE，which is applicable to the imbalanced financial datasets.Based on the dynamic transaction features of the accounts，the transaction behavior features are extracted from the dimensions of statistical information，sequential order information and supervision information.Then a datasets equalization strategy for data pre-processing is proposed to address the problem of cross-region sample synthesis，which is faced by the oversampling technology，ADASYN，on the financial account datasets.The strategy uses the iForest algorithm for mixed sampling of the data to remove the majority of noisy data and reduce the difficulty of the classifier learning from the minor classes.On this basis，a random forest classifier is designed to implement the detection of the accounts involved in financial fraud.The experimental results on the datasets of actual financial account transactions show that iForest-SMOTE has a clear advantage in the recall rate and accuracy over ADASYN，SMOTE and other sampling techniques.Its F-value is at least 2.13 percentage points higher than that of the other algorithms.

【Key words】isolation forest；imbalanced classification；fraud account detection；random forest；feature mining

DOI：10. 19678/j. issn. 1000-3428. 0058006

0 概述

欺诈可以定义为导致金钱或个人利益损失的不正当或刑事欺骗行为。近年来，欺诈活动的形式和规模随着跨银行交易而变得越来越复杂和庞大，普华永道（PwC）[1]2018年的全球经济犯罪调查结果显示，有49%的公司在过去两年经历过金融欺诈行为，2016年的这一数据仅为36%。面对海量、多样的欺诈手段，基于专家知识、侦查经验的传统欺诈账户识别方法已经难以满足当前金融安全保障的需求。如何从海量金融数据中自动识别少数欺诈账户逐渐成为侦查部门及大数据研究人员关注的问题。

金融欺诈账户检测是一项难度较高的任务，许多学者使用不同方法从多个角度研究检测模型。文献[2]

基金项目：国家重点研发计划“网络空间安全”重点专项（2017YFB0801804）。

作者简介：吕 芳（1990—），女，博士研究生，主研方向为金融安全、数据挖掘、图挖掘；汤丰赫，本科生；黄俊恒，副教授；王佰玲（通信作者），教授、博士。

收稿日期：2020-04-08 **修回日期：**2020-06-19 **E-mail：**wbl@hit.edu.cn

采用广义的定性相应模型（EGB2）来预测企业管理层进行的欺诈活动，文献[3]提出一种成本敏感的决策树欺诈检测方法，文献[4]对比了利用支持向量机（SVM）、逻辑回归和随机森林构建模型对欺诈检测的性能，文献[5]通过比较金融欺诈检测中机器学习算法的性能，得出随机森林算法是最佳的金融欺诈检测技术。在真实的交易数据中，欺诈账户的数据量相对整个数据集来说比例极少，且其具有欺诈倾向的行为活动被淹没在海量、常规的金融交易活动中。若直接采用上述分类模型，由于常规交易（多数类样本）数量多，欺诈交易（少数类样本）数量少，会导致欺诈检测模型在学习分类边界时无法充分捕捉少数类样本的类别特征，从而影响对欺诈账户的检测性能。因此，解决数据集在类间的非平衡问题对提升账户分类模型的检测性能具有重要意义。文献[6]发现不平衡性通常会导致少数类内部形成小杂项（间断和分离），导致其在决策时易被错误地学习，从而降低欺诈检测性能，造成该现象的主要原因是一些典型的少数类样本在少数类中分布稀疏，数量较少。可见，解决小杂项引起的类内不平衡问题也同样值得关注。

目前，解决数据集不平衡问题的方法主要分为两类。一类从数据层面入手，通过改变数据样本的分布来降低数据的非平衡性，常用方法有欠采样和过采样技术，它们分别对应少数类样本的增加和多数类样本的减少。另一类从算法层面入手，通过调整算法来适应分类不平衡问题，如代价敏感学习、集成学习等。在过采样技术的研究中，文献[7]提出用于不平衡学习的自适应合成采样方法（ADASYN），该方法使用密度分布作为准则为少数类样本分配权重，从而自适应地生成少数类的合成数据样本，以减少由不平衡数据分布引起的偏差。对于处于多数类高密度分布区域内的少数类样本，ADASYN会将该样本作为“较难学习”的样本，赋予其高权重并为其生成更多的合成样本。虽然使用ADASYN会面临跨决策区域合成样本的风险，但作为一种新的学习方法，其基于密度分布自适应地给予样本权重并进行样本合成的思想，可以用于处理不同情况下的不平衡学习问题。除了采用分类模型进行少数类检测，有研究人员将“异常”定义为“离群点”，进而提出众多“异常”检测方法，如基于密度、测量和iForest方法。其中，iForest是由文献[8]提出的基于孤立概念的无监督异常检测方法，其将“异常”定义为“容易被孤立的离群点”。在特征空间中，分布在稀疏区域的点表示某事件在稀疏区域发生的概率很低，iForest认为落在这些区域中的点是“异常”的，因此，通过iForest可以快速高效地检测数据集中分布稀疏且离密度高群体较远的异常点。

欺诈账户交易行为的隐蔽性导致正常账户和欺诈账户的类别边界模糊，严重影响了分类器的检测性能。因此，有必要针对金融账户模糊的类别边界进行分析。模糊边界中的节点集合主要分为少数类的异常点和多数类的异常点。其中，多数类的异常点作为存在于少数类内部或决策边界的冗余样本，是导致决策边界混乱的重要原因；少数类的异常点作为少数类内部的稀疏样本会导致小杂项的产生，是引发类内不平衡问题的重要原因。

本文借鉴iForest检测异常点的算法思想以及ADASYN决策边界样本合成方法，设计一种样本均衡策略。提出一种基于iForest解决分类不平衡问题的金融欺诈账户检测框架（iForest-SMOTE），框架主要包括特征抽取、数据集均衡、欺诈账户检测三个部分。样本的分类特征提取是影响分类器性能的一个关键因素，金融数据同时具有网络、流式数据的特点。因此，为了全面描述账户的交易行为，本文分别从静态交易信息、交易关系和交易周期性三个维度进行特征抽取。具体地，本文分别从交易资金、交易网络和交易周期三个维度设计银行账户的交易行为特征抽取方法。为了解决类别样本不均衡问题，提出一种基于iForest解决非平衡数据集的方法。该方法通过iForest对数据集进行检测以获取预处理样本子集，根据类别不同对其采用不同的调整策略，从而提升欺诈检测的性能，具体地，负采样多数类样本，减轻决策边界的混乱程度，重采样少数类样本，减少内部小杂项的产生，结合ADASYN将决策边界向具有决策影响力的少数类异常点附近移动。在分类器的选择上，结合金融数据分类特征复杂、类间不均衡的特点，本文采用随机森林分类器模型[9]检测金融欺诈账户。

1　相关工作

1.1　iForest异常检测技术

iForest是文献[8]基于样本集中异常样本是稀疏且异于正常样本的两个假设而提出的一种基于孤立点的无监督异常检测方法，该方法使用二值树结构（iTree）将每个实体转化为树结构中的孤立节点。基于异常点对孤立划分更敏感的理论，通过子采样使得异常点相对正常点距离iTree的root节点路径更近。iForest有效解决了异常检测中的淹没效应（异常点和正常点的距离很小）和掩蔽效应（异常点增多，导致其密度增大），因此，iForest可以快速高效地检测离群点。随后，为将iForest扩展到分类、在线异常检测和高维数据中，研究人员进行了一系列探索。文献[10]将iForest扩展到类别数据集上，对用户日志中体现出的用户行为模式进行异常检测。文献[11]改进iForest中的约束条件，实现对多类别正常数据中局部聚集异常数据集合的检测，文献[12]根据iForest中异常分数的热图提出扩展隔离森林（EIF），ELF可以稳定高效地对高维数据进行异常检测。此外，文献[13]基于iForest提出一种自适应方法，实现对网络管理系统的快速异常检测，文献[14]通过iForest对软件进行缺陷预测。

针对金融账户数据，由于正常和欺诈账户在金融交易模式上具有一定的相似性，在特征空间中表现为分布在决策区域附近的样本密度集中且分布混乱，导致iForest在样本密集区域中检测少数类样本的效率较低，不能直接用于金融欺诈账户检测任务。但是，由于iForest检测出的异常点具有孤立的特性，使得该点在不同类别的决策中具有重要作用，因此iForest的异常点可用于样本均衡。

1.2 类别均衡方法

改善数据集类别不均衡问题的方法分为数据级别和算法级别两类。其中,数据级算法主要包括对数据集进行欠采样和过采样。在欠采样方面,文献[15]将聚类与实例选择相结合对不均衡数据集进行欠采样。上述方法加速了分类过程,但对数据进行过度欠抽样时将导致提升分类器性能的样本信息被消除。文献[16]通过欠采样技术去除决策边界的嘈杂和冗余多数类实例,以减少分类器对分类不平衡的敏感度。在银行账户数据集中,一部分多数类样本会成为嘈声存在于少数类内部或决策边界,因此,选择有效的欠采样技术有助于排除降低决策的多数类样本。过采样通过增加少数类样本以达到数据集平衡,若随机复制样本有可能降低样本的泛化能力、加剧少数类中噪音数据对模型的影响。为此,研究人员通过插值生成人工样本,扩大少数类的泛化空间。文献[17]提出SMOTE技术,插入彼此接近的少数类样本以合成新的少数类样本,保证新增少数类样本的质量。然而,SMOTE为所有实例赋予相同的权重,忽略了决策区附近实例对分类的重要性。据此,文献[18]提出了borderline-SMOTE1和borderline-SMOTE2两种改进方法,然而这两种方法均只为决策边界附近的少数类样本分配高采样权重。文献[19]提出一种混合采样的方法,该方法将过采样技术SMOTE与从多数类中消除歧义样本的欠采样技术相结合,通过进行样本均衡来解决数据集的不平衡问题。另外,文献[6]提出用于不平衡学习的基于密度分布的自适应合成采样方法ADASYN,其将分布在高密度多数类中的少数类样本定义为较难学习的样本,设计参数调节较难学习的样本的采样权重,从而自定义地合成更多样本。ADASYN在改善数据集非平衡问题的同时还可以将分类的决策边界自适应地转移到教难学习的样本上。但是,当有大量较难学习样本存在于多数类内部时,ADASYN会在合成少数类样本时跨越决策区域,加剧决策区域的混乱程度。总体而言,ADASYN算法具有较强的泛化能力,通过修改和扩展,可用于解决不同场景下的类别不平衡问题。

由于ADASYN根据多数类的密度分布准则对少数类进行权重分配,当少数类样本分布在多数类内部时,合成样本会面临跨决策边界合成的风险。金融数据的复杂性导致其类别边界模糊,直接使用ADASYN会加剧决策边界的混乱程度。金融数据中不同类别的异常点具有不同的特性,难以确定其能否对决策产生正面影响。为了提高欺诈检测性能,本文对不同类别的异常点实施不同的策略:一方面,将属于多数类的异常点(多数类异常样本)作为嘈杂样本,对该样本和其附近的多数类样本进行筛除,以降低决策边界和少数类内部的混乱程度;另一方面,对于属于少数类的异常点(少数类异常样本),借鉴ADASYN的思想进行样本合成,以在样本均衡的同时减少出现小杂项的风险,并将少数类的决策边界调整到具有典型性的少数类样本附近。

1.3 随机森林分类模型

随机森林[8]是一种由多棵决策树组成的集成学习模型,随机森林在多种分类任务中相对其他机器学习算法具有明显优势,因此受到数据分析、知识管理、模式识别等众多领域研究人员的广泛关注[20]。在异常检测方面,文献[21]使用两种不同的随机森林算法分别训练正常和欺诈交易的行为特征,检测信用卡欺诈行为;文献[22]提出一种采用交易时间序列中固有模式对文件进行汇总的欺诈检测方法,从而评估支持向量机、随机森林等多种分类模型,验证了随机森林具有高效的检测性能。

随机森林在金融数据分类任务中具有明显优势,但非平衡数据集引发的数据稀缺、噪声等问题会大幅降低分类准确性。因此,本文提出iForest-SMOTE框架,对金融数据集进行样本均衡后使用随机森林分类器模型实现欺诈账户检测。

2 iForest-SMOTE框架

iForest-SMOTE框架如图1所示。首先,在银行账户交易数据集中抽取分类特征,包括交易资金、交易网络、交易周期、有监督交易行为等特征,从而构建样本特征数据集;其次,为解决样本不均衡问题,利用iForest进行特征数据集均衡预处理,得到异常样本数据集,并针对其中的多数类异常样本、少数类异常样本分别设计去采样、过采样数据均衡策略,实现样本自适应合成以达到类别数据均衡的目的;最后,采用随机森林分类器对类别均衡特征数据集进行欺诈检测。

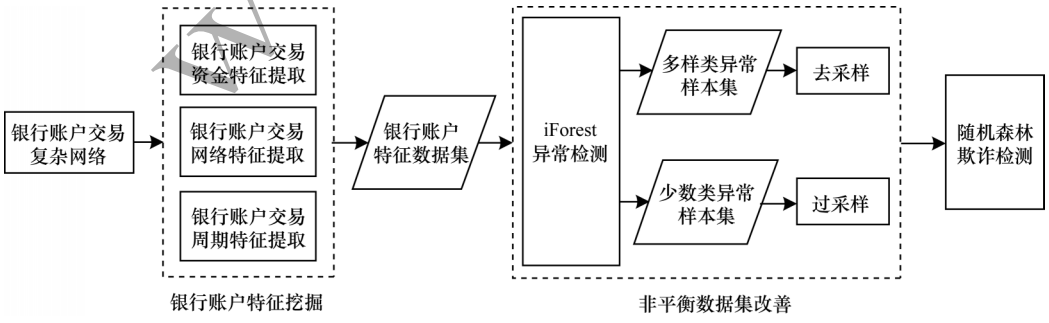

图1 iForest-SMOTE框架

Fig.1 The framework of iForest-SMOTE

2.1　基本定义

在详细描述iForest-SMOTE欺诈账户检测框架之前，本文先给出一些基本的问题说明和定义。

定义1（银行账户数据集）　一个银行账户数据集表示为$D \subseteq C \times B$，其中，$C=\{c_1, c_2, \cdots, c_n\}$为银行账户数据集信息，$c_i$为账户$i$的数据，集合$B=\{T, F\}$作为欺诈账户检测的标记集，$T$和$F$分别代表欺诈标记和正常标记，$B_{c_i}$代表账户$i$的标记。在数据集$D$中，少数类记为$P=\{p_1, p_2, \cdots, p_{\text{pnum}}\}$，$P \subseteq D$，且$B_{p_i}=T$，多数类记为$N=\{n_1, n_2, \cdots, n_{\text{nnum}}\}$，$N \subseteq D$，且$B_{n_i}=F$。

定义2（分类特征集）　设集合$C=\{c_1, c_2, \cdots, c_n\}$是符合定义1的银行账户数据集，$c_i$的$m$维分类特征依次定义为交易行为特征值向量$\boldsymbol{x}_i^a (a=1, 2, \cdots, l_\mu)$、交易网络特征值向量$\boldsymbol{x}_i^b (b=l_\mu+1, l_\mu+2, \cdots, l_v)$、交易周期特征值向量$\boldsymbol{x}_i^c (c=l_v+1, l_v+2, \cdots, l_\xi)$、有监督交易行为特征值向量$\boldsymbol{x}_i^d (d=l_\xi+1, l_\xi+2, \cdots, m)$，由所有$c_i$的交易统计特征向量构成的集合记为银行账户分类特征集C_{x_a}。

定义3（iForest异常标记）　给定银行账户数据集D，其分类特征集为C_{x_a}，采用iForest对D进行异常检测的模型可表示为：

$$\text{iForest}(C_{x_a}, L, N_w) \rightarrow A^n$$

其中，L为iForest中要选择的iTree数量，N_w为采样大小，$A=\{T_{\text{special}}, F_{\text{special}}\}$为iForest对账户的标记集，$T_{\text{special}}$和$F_{\text{special}}$分别代表异常和正常标记，$A_{c_i}$表示iForest对$c_i$的标记。

定义4（样本预处理）　给定标记集A，$D_{\text{special}} \subseteq C$为$C$中属于异常标记的预处理样本子集，其中，$D_{\text{special}}$满足如下条件：

$$\forall c_i \in D_{\text{special}}, A_{c_i}=T_{\text{special}}$$

定义5（异常样本集）　给定D_{special}，其中，属于少数类的样本组成少数类异常样本集P_{special}，属于多数类的样本组成多数类异常样本集N_{special}，则P_{special}和N_{special}的数学定义如下（P、N详见定义1）：

N_{special}定义为：$\forall c_i \in N_{\text{special}}$，$c_i \in N$且$c_i \in D_{\text{special}}$

P_{special}定义为：$\forall c_j \in P_{\text{special}}$，$c_i \in P$且$c_i \in D_{\text{special}}$

2.2　数据均衡策略

受到iForest检测出的异常样本在不同类别中具有不同特性的启发，本文设计一种样本均衡策略。

多数类异常点指远离多数类的离群点。文献[23]采用去采样多数类（记为$x \in S_{\text{maj}}$）的方法减弱噪声数据对分类器的影响。去采样的核心是确定要筛除的多数类样本。远离多数类的离群点会成为噪声数据，致使分类器依据错误的样本学习。因此，本文将多数类异常点作为噪声源点，并将多数类异常点近邻的多数类样本构成的集合作为噪声簇，将多数类异常点和其对应的噪声簇从多数类中去除。

少数类异常点指在特征空间中分布稀疏、数量较少的离群点。过采样技术通过对少数类（记为S_{min}）进行人工合成数据，以解决小样本数据不均衡问题。过采样算法的核心[7]是确定每个少数类样本$x \in S_{\text{min}}$的合成样本数量k。ADASYN首先计算$\forall x_i \in S_{\text{min}}$在$S_{\text{maj}}$中的密度分布$\hat{r}_i$，并将$\hat{r}_i$作为权重衡量准则来确定$x_i$的过采样次数$k_i$。可见，$\hat{r}_i$值正比于集合$S=S_{i-\text{near}} \cap S_{\text{maj}}$的大小，其中，$S_{i-\text{near}}$为$x_i$的KNN邻近样本集，高$\hat{r}_i$值样本分布在多数类高密度区域，该样本在分类器中难以被学习，因此，ADASYN根据$\hat{r}_i$值赋予该类样本更多的过采样次数，使分类器更加关注难以学习的样本。

从上述分析可以看出，过采样通过对少数类进行样本合成从而使分类器充分地对少数类进行学习，进而提升决策性能，去采样因筛除了噪声数据而提升决策性能，过采样改善了数据集的不平衡性问题。然而，ADASYN在处理S集合过大或决策边界混合严重的问题时，会面临跨决策区域合成数据的风险。欺诈账户的隐蔽性导致金融账户数据集中存在一定数量的少数类样本分布在决策边界和多数类内部，使用多数类的密度分布$\hat{r}_i$计算并合成样本会使多数类内部和决策边界出现大量的少数类合成数据，提高了分类器模型错误地学习样本的几率并加剧了决策边界的混乱程度。

为解决上述问题，本文利用异常点在特征空间的密度改进ADASYN中的权重衡量准则$\hat{r}_i$，以提升分类器的欺诈检测性能。

2.3　特征抽取

在分类框架设计时需要考虑如何表示样本的类别特征以及避免特征集合冗杂等问题。根据定义2，银行账户的交易行为可量化为资金特征、网络特征、周期特征以及有监督的交易特征。

2.3.1　交易资金特征

将账户视为单一个体，其历史交易数据视为静态时序数据，可从统计角度表示其交易资金特征，则定义2中的$\boldsymbol{x}_i^a (a=1, 2, \cdots, l_\mu)$具体表示为账号$i$收入和支出两种交易类型分别对应的资金相关统计项，如交易金额、交易次数等，交易资金特征如表1所示。

表1　交易资金特征汇总

Table 1　Summary of transaction capital characteristics

特征编号	特征描述
1	账户银行卡交易收入总金额
2	账户银行卡交易支出总金额
3	账户银行卡收入总交易次数
4	账户银行卡支出总交易次数
5	账户银行卡收入平均交易次数
6	账户银行卡支出平均交易次数

2.3.2 交易网络特征

账户与其直接交易账户集合之间的资金流动构成了自我中心金融关系网络,据此,将账户的交易行为转化为一个局部中心网络,该网络的属性特征可视为账户的交易特征,则定义2中的$\boldsymbol{x}_i^b(b=l_\mu+1, l_\mu+2, \cdots, l_v)$为账户$i$的一阶关系网络特征,具体特征项如表2所示。

表2 交易网络特征汇总

Table 2 Summary of transaction network characteristics

特征编号	特征描述
7	账户的黑洞节点标记
8	账户的白洞节点标记
9	账户的交流边标记
10	账户的交易次数中转节点标记
11	账户的交易金额中转节点标记
12	账户入度
13	账户出度
14	账户的LeaderRank值

如表2所示,$\boldsymbol{x}_i^b(b=l_\mu+1, l_\mu+2, \cdots, l_v)$包括账户$i$的交易入度$d_{\text{in}}$、出度$d_{\text{out}}$、根据进出交易对比得到的账户$i$的黑洞(账户转账远大于出账)和白洞(账户出账远大于转账)节点标记、根据网络计算出的LeaderRank值[24]和对流边[25]账户之间的频繁交易等特征。

2.3.3 交易行为周期特征

账户的交易行为反映了持卡者的社会经济活动,则社会活动的周期性、规律性也会体现在交易数据上。以一个月为一个活动周期单位,分析账户交易的周期波动,则账户i的交易周期特征$\boldsymbol{x}_i^c(c=l_v+1, l_v+2, \cdots, l_\xi)$如表3所示。

表3 交易行为周期特征汇总

Table 3 Summary of transaction behavior cycle characteristics

特征编号	特征描述	特征类别
15	月进最大日总金额占月进总金额的比值	交易金额
16	月进最多日总金额占月进总金额的比值	交易金额
17	月进出金额差值	交易金额
18	账户月进对手个数中值	交易对手
19	账户月出对手个数中值	交易对手
20	账户月进出对手差值比值	交易对手
21	月进最大日方差和异常系数	日期差异
22	月进最多日方差和异常系数	日期差异
23	月出最大日方差和异常系数	日期差异
24	月出最多日方差和异常系数	日期差异
25	月进出最大金额日期的方差	日期差异
26	月进出最多金额日期的方差	日期差异
27	月进最大金额和出最多金额日期的方差	日期差异
28	月进最多金额和出最大金额日期的方差	日期差异

2.3.4 有监督的交易特征

在异常检测任务中,若将已知的专家知识量化为分类特征,对优化分类器具有重要作用。这类特征与具体的欺诈类型相关,金融欺诈的实施方式、欺诈团伙的牟利模式、欺诈组织的运营方式等,均直接影响有监督交易特征的定义和量化。本文以传销欺诈组织为例,对此类特征进行说明。传销组织的资金流通方式多呈现金字塔形式,会员费(本文称为申购资金)自底向上流经固定的申购账户汇集到顶层账户;提成(本文称为返利资金)按比例从顶层经由返利账户下发给各会员。针对涉及传销的账户i,其$\boldsymbol{x}_i^d(d=l_\xi+1, l_\xi+2, \cdots, m)$的各特征分量如表4所示。

表4 有监督的交易特征汇总

Table 4 Summary of supervised transaction characteristics

特征编号	特征描述	特征类别
29	申购金额次数	申购返利
30	申购金额总额	申购返利
31	申购对手数量	申购返利
32	返利金额次数	申购返利
33	返利金额总额	申购返利
34	返利对手数量	申购返利
35	对手账户的申购金额次数	申购返利
36	对手账户的申购金额总额	申购返利
37	对手账户的申购金额数量	申购返利
38	对手账户的返利金额次数	申购返利
39	对手账户的返利金额总额	申购返利
40	对手账户的返利金额数量	申购返利

需要指出的是,本文提出的特征为串联关系,因此,若异常检测任务缺乏背景知识则特征值向量可忽略此类特征。

2.4 基于iForest的数据均衡预处理

如上文所述,金融交易数据中正常账户、欺诈账户样本的不均衡问题,严重影响欺诈账户检测模型的性能。为此,本文提出一种基于iForest改善非平衡数据集的策略。采用iForest进行异常子集筛选,以获取银行账户特征数据集中的异常样本集,进而将其划分成多数类异常样本和少数类异常样本,分别对上述两类样本采用欠采样和自适应生成合成样本的方式实现类别均衡。

2.4.1 基于iForest的异常子集筛选

本文首先对所构建的银行账户特征数据集C_{x_α}进行iForest异常检测,为每个账户样本分配一个异常账户检测标记,其次根据样本的异常检测标记对样本进行预处理,最后根据预处理样本子集中样本的欺诈标记对样本进行筛选,以获取少数类异常样本集和多数类异常样本集。具体过程如下:

1)通过iForest对特征数据集C_{x_a}进行检测并得到每个特征样本$C_{x_a}^i$的标记集:

$$\text{iForest}(C_{x_a}^i, L, N_w) \rightarrow A_i$$

2)将标记集A^n中标记为T_{special}的样本加入到D_{special}中,对于$\forall c_i \in C$,如果$A_{c_i}=T_{\text{special}}$,则$D_{\text{special}}=D_{\text{special}} \cup c_i$。

3)对预处理样本子集的样本进行筛选:对于$\forall c_j \in D_{\text{special}}$,如果$\exists c_j \in N$,则$N_{\text{special}}=N_{\text{special}} \cup c_j$,如果$\exists c_j \in P$,则$N_{\text{special}}=N_{\text{special}} \cup c_j$。

在具体实现过程中,C_{x_a}、L、N_w分别表示银行账户特征数据集、iTree的数量、数据采样大小,N、P是符合定义1的多数类和少数类,A_{c_i}是符合定义3中c_i样本的异常标记,D_{special}是符合定义4的预处理样本子集,N_{special}和P_{special}分别为符合定义5的多数类异常样本集和少数类异常样本集。

2.4.2 多数类样本去采样

本节将对2.4.1节筛选的多数类异常样本进行欠采样处理,以减少嘈声样本对决策的影响,具体过程如下:

1)对于每一个多数类异常样本$c_i \in N_{\text{special}}$,计算距离$c_i$最近并且属于多数类的$K_1$个邻近样本$c_{i-\text{near}}$,将$c_{i-\text{near}}$构成$c_i$的噪声簇$M_{i-\text{near}}^{\text{Maj}}$:

$$M_{\text{i-near}}^{\text{Maj}} = \bigcup_{j=1}^{K_1} c_{i-\text{near}\,j}$$

2)将每一个多数类异常样本$c_i \in N_{\text{special}}$和$c_i$对应的噪声簇$M_{\text{i-near}}^{\text{Maj}}$从多数类$N$中去除:

$$N = N - c_i - M_{\text{i-near}}^{\text{Maj}}$$

样本之间距离计算采用欧几里得距离:

$$D(x,y)=\sqrt{\sum_{i=0}^{n}(x_i-y_i)^2}$$

其中,x、y为空间中的任意两个样本,x_i和y_i为对应的i维度的数值。

2.4.3 少数类样本过采样

本节将少数类异常点的密度分布作为权重衡量准则$\hat{r}_i$,提出一种基于iForest的少数类自适应合成样本方法,以解决决策边界混合严重的金融账户数据集不平衡问题。具体地,计算每个少数类样本$c_i \in P$在少数类异常样本集P_{special}中的密度分布$\hat{r}_i$,即c_i的近邻样本中少数类异常点所占比重,从而确定c_i进行合成样本的数量k_i,当c_i越靠近少数类异常点时,$\hat{r}_i$值越大,合成的样本越多。继而在少数类异常点附近合成更多样本,从而在改善少数类内不平衡问题的同时降低分类器学习少数类异常点的难度,具体过程如下:

1)计算需要生成的合成数据数量G:

$$G=(L_{\text{len}}(N)-L_{\text{len}}(P))\times\theta$$

其中,$\theta \in [0,1]$为用户定义参数,用于指定生成合成数据的水平,当$\theta=1$时将得到完全平衡的样本集。

2)计算针对每个少数类样本$p_i \in P$需要合成的数据数量g_i,计算过程如下:

对于$\forall p_i \in P$,首先计算距离p_i最近的K_2个近邻样本构成的近邻样本集$D_{i-\text{near}}$,其次计算$D_{i-\text{near}}$中少数类异常样本$c_j \in P_{\text{special}}$所占的比重$r_i$:

$$r_i=\frac{\Delta_i}{K_2},\ r_i \in [0,1]$$

其中,Δ_i是近邻样本集$D_{i-\text{near}}$中c_j的样本数量,$\Delta_i=L_{\text{len}}(D_{i-\text{near}} \cap P_{\text{special}})$。对$r_i$规范化得到$\hat{r}_i$:

$$\hat{r}_i=\frac{r_i}{\sum_{i=1}^{p} r_i}$$

其中,$\hat{r}_i$是r_i的密度分布,$\hat{r}_i$值越大,在该样本附近合成的数据越多。

最后,计算p_i需合成的数据数量$g_i=\hat{r}_i \times G$,G为前文计划由少数类生成的目标数据总量。

3)对少数类样本进行样本合成。对于每一个少数类样本p_i,进行g_i次样本合成,在合成人工数据时,本文选择近似SMOTE[17]中的数据合成方法,具体过程如下:

对每个少数类样本p_i进行g_i次循环,每次循环步骤为:

步骤1 计算距离p_i最近的K_3个属于少数类的近邻样本并构成近邻样本集$M_{i-\text{near}}^{\text{Min}}$。

步骤2 在$M_{i-\text{near}}^{\text{Min}}$中随机选择一个少数类样本$p_{zi}$。

步骤3 根据p_{zi}和p_i的特征进行人工数据合成,合成公式如下:

$$s_{xi}=p_{xi}+(\boldsymbol{p}_{xzi}-\boldsymbol{p}_{xi})\times\lambda$$

其中,s_{xi}是合成样本的特征,$\boldsymbol{p}_{xi}$和$\boldsymbol{p}_{xzi}$分别是少数类样本p_i和p_{zi}符合定义2对应的特征向量,$(\boldsymbol{p}_{xzi}-\boldsymbol{p}_{xi})$为$n$维空间中特征的差失量,$\lambda$是随机数,$\lambda \in [0,1]$。

步骤4 赋予合成的特征向量少数类标签$B_{s_i}=T$,并将对应的样本s_i加入少数类中,$P=P \cup s_i$。

结束循环。

上述过程修改了$\hat{r}_i$的衡量标准,因此,与ADASYN相比具有相同的时间和空间复杂度。

本文通过赋予少数类异常点和其临近样本更高的权重来调整合成样本的数量,不仅实现了样本均衡还降低了跨区域合成的风险,同时合成的样本会提高少数类异常样本附近的少数类密度,降低内部小杂项出现的概率,通过合成样本能够转移少数类的决策边界。

2.5 欺诈账户检测模型

iForest-SMOTE首先通过对银行账户数据进行特征抽取并生成特征数据集,再通过银行特征数据集实现类别均衡,得到样本均衡数据集D_{balance},随后采用随机森林分类模型检测欺诈样本,分类器的输入为D_{balance}中样本平衡特征数据集C_{x_a},输出为分类模型对每个样本的分类结果。

3 实验与结果分析

3.1 实验环境与数据集

本文实验的硬件环境为Inter®Core™i7-7700HQ，内存(RAM)为16 GB。软件环境为Python语言，Windows 10操作系统。实验数据为由经侦部门提供的脱敏资金交易数据，其中包括正常金融账户和欺诈账户四年内产生的银行交易数据，每条交易数据包括交易双方账户、交易方向、交易时间、交易金额等属性，共涉及账户15 633个，传销账户为1 303个。数据集含有总账户交易数据227 179条，传销账户交易数据64 630条。实验将数据转化为7 859条银行账户数据，其中属于少数类的账户数据共778条，属于多数类的账户数据共7 081条，多数类和少数类节点比为10∶1。随机抽取数据集中70%的数据作为训练集，其余30%的数据作为测试集。

3.2 分类效果衡量指标

随机森林是用于分类和预测的组合分类器，分类效果是评价分类器性能的典型指标。本文使用混淆矩阵作为分类器的性能衡量指标，混淆矩阵详见表5。

表5 混淆矩阵

Table 5 Confusion matrix

预测值	真实值	
	欺诈	正常
欺诈	TP(Ture Positive)	FP(False Positive)
正常	FN(False Negative)	TN(Ture Negative)

其中，TP表示真实值和分类结果均为欺诈，FN表示真实值为欺诈而分类结果为正常，FP表示真实值为正常而分类结果为欺诈，TN表示真实值和分类结果均为正常。

本文采用准确率、召回率、精确率、F-value值评价模型的分类效果。准确率Accuracy为分类模型所有判断正确的样本数占样本总数的比例；召回率Recall为在模型预测为欺诈的样本集合中，真实值也为欺诈的样本数占所有真正为欺诈的样本总数的比例；精确率Precision为在被模型预测为欺诈的所有样本集合中，真正为欺诈的样本比例；F-value值从少数类的角度综合评价随机森林的性能，它是召回率和精确率的组合。

3.3 实验结果

3.3.1 采样均衡策略评估

在非平衡数据欺诈检测问题中，由于欺诈类别属于少数类，因此少数类的分类准确率对于评价分类模型更有意义，本文采用召回率Recall、精确率Precision、F-value值等指标在少数类上的平均得分来评价不同欺诈检测模型的性能。为了验证本文iForest-SMOTE框架对不均衡数据集的优化效果，统一对不同算法处理后的特征数据集采用随机森林进行欺诈检测。特征数据集包括分别经过随机过采样算法(RamdonOverSampler)、ADASYN算法、SMOTE算法、iForest-SMOTE框架处理后的数据集以及只进行特征提取的数据集。随机森林对不同特征数据集的检测效果如表6所示。其中，使用下划线标出每项指标的最佳取值，并加粗显示本文算法(iForest-SMOTE)的各项指标取值。

表6 不同方法的性能比较结果

Table 6 Performance comparison results of different methods %

方法	Accuracy	Precision	Recall	F-value
数据集未处理	93.35	81.14	68.73	74.42
RamdonOverSampler	88.90	58.98	90.30	71.35
ADASYN	89.32	60.46	87.53	71.52
SMOTE	91.88	69.06	85.11	76.25
iForest-SMOTE	**93.42**	**78.93**	**77.83**	**78.38**

由表6可知，尽管某些算法(如ADASYN)的召回率Recall指标具有较高水平，但其他指标大多处于较低的水平，导致综合指标F-value值偏低。ADASYN的F-value值较低说明其存在跨区域合成样本的风险，不适合用来解决金融数据集的非平衡问题。与其他算法相比，本文iForest-SMOTE模型在召回率和准确率方面都处于较高的水平，F-value相比对比算法至少提升2.13个百分点。综合各项指标得出，iForest-SMOTE框架能够为检测模型提供更好的特征集合筛选功能，可以明显提高分类器的欺诈账户检测能力。

ROC曲线可以描述分类器的性能，是针对不平衡技术的重要判断依据，ROC曲线越靠近左上角表示非平衡技术越能提升分类器的性能。图2所示为金融账户数据集的ROC曲线。

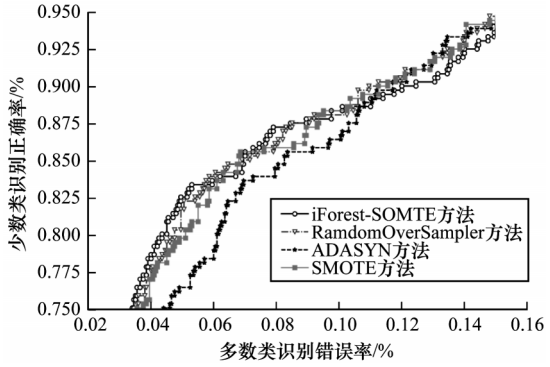

图2 不同分类方法的ROC曲线

Fig.2 ROC curves of different classification methods

从图2可以看出，各个方法的分类性能较为接近，其中，iForest-SOMTE具有相对较高的少数类识别正确率。ROC曲线下的面积可以用来度量非平衡分类模型的功效，通常将该度量值称为AUC，AUC值介于0和1之间，其中，0.5为随机猜测值。在非平衡

数据集中，AUC值更加能够体现两个类别的正确性。不同方法的AUC值如表7所示。

表7　不同方法的AUC值

Table 7　AUC values of different methods　%

方法	AUC值
RamdonOverSampler	96.66
ADASYN	96.21
SMOTE	96.63
iForest-SMOTE	**96.73**

由表7可知，iForest-SMOTE具有较高的AUC值，表明其对金融不平衡数据集具有更好的处理效果。

3.3.2　分类特征重要性评估

通过随机森林对特征重要性的评估，可以了解每种特征在构建决策模型时的重要性，这为后续的特征筛选提供了一定支撑，有利于提高模型的鲁棒性。本节对提取的每维分类特征在决策中的重要性进行评估。

随机森林特征重要性评估的思想为：比较每个特征在随机森林的所有决策树上分类贡献的平均值，然后比较特征之间的贡献值大小。本文采用基尼指数评估重要性，对于特征x_j，计算在随机森林的每一颗决策树中由特征x_j形成的分支节点的基尼指数Gini(p)下降程度之和（基尼不纯度下降程度）。其中，基尼指数Gini(p)为：

$$\sum_{k=1}^{K} p_k(1-p_k) = 1-\sum_{k=1}^{K} p_k^2$$

其中，K代表类别个数，$\sum_{k=1}^{K} p_k = 1$。

特征x_j的重要性评估过程具体如下：

1）计算特征x_j在决策树中节点m处的下降程度$V_{jm}^{(\mathrm{Gini})}$：

$$V_{jm}^{(\mathrm{Gini})} = G_m - G_l - G_r$$

其中，G_l和G_r表示在决策树中节点m分支前后两个新节点的Gini指数。

2）计算特征x_j在决策树i上的特征重要性：

$$V_j^{(\mathrm{Gini})} = \sum_{m \in M} V_{jm}^{(\mathrm{Gini})}$$

其中，m为特征x_j在决策树i中出现的节点，M为节点m的集合。

3）计算特征x_j在随机森林中的分类重要性：

$$V_j^{(\mathrm{Gini})} = \sum_{i=1}^{n} V_{ij}^{(\mathrm{Gini})}$$

其中，n为随机森林中的决策树数量。

4）对所有特征的重要性评分进行归一化处理，特征x_j的重要性评分为：

$$V_j = V_j \Big/ \sum_{i=1}^{c} V_i$$

其中，c为特征的总数量。

根据上述方法，本文提取的金融账户分类特征集合中每维特征的重要性如图3所示，其中，银行账户特征中LeaderRank值（编号14）、入度（编号12）、出度（编号13）等特征的贡献占比较高，由此可知，这三个特征对辨识欺诈账户尤为关键，表示交易网络特征（编号7~编号14）对欺诈账户检测具有重要作用。此外，银行账户交易资金特征（编号1~编号6）的特征贡献度总体相对较低，但体现账户交易敏感资金和交易敏感次数的申购返利特征（编号29~编号40）具有较高的贡献占比，说明在传销账户识别中，账户的申购和返利交易能有效区分欺诈账户和正常账户，即有监督交易特征在提升欺诈账户检测性能中具有重要作用。

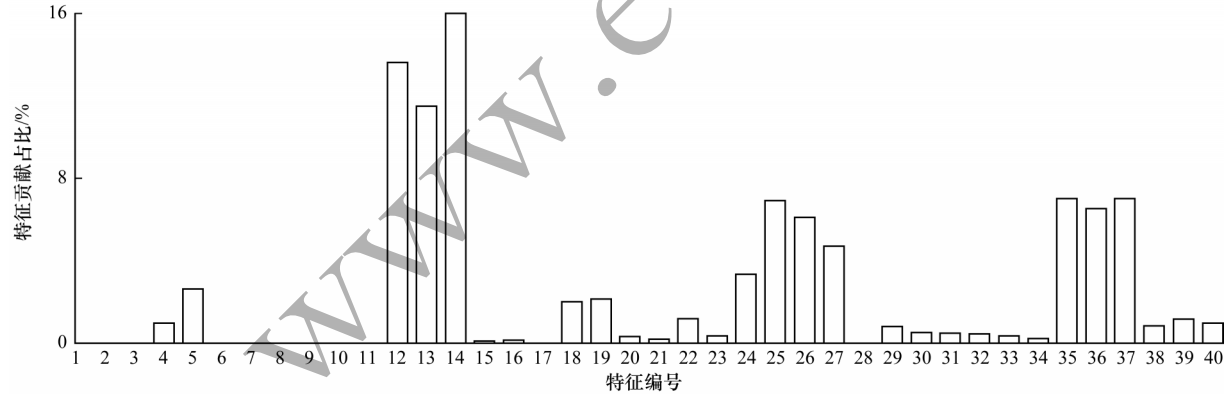

图3　分类特征的重要性程度

Fig.3　Importance degree of classification features

4　结束语

本文设计一种欺诈账户检测框架iForest-SMOTE。针对实际数据中欺诈样本不均衡的问题，结合iForest对异常边界的识别能力与ADASYN对决策边界的样本合成思想，改善分类器的训练数据集。分析样本在交易的时序、关系、周期及有监督异常行为方面体现出的判别特征，进而组合生成分类特征数据集。iForest-SMOTE中的随机森林分类模型用于提高分类准确性并实现对各分类特征的重要性评

估。在真实含有传销欺诈账户的数据集上进行实验,结果表明,iForest-SMOTE在严重不均衡数据集中仍能取得较高的识别准确率。下一步将在无监督的数据集上实现异常边界调整,以改进无标签非平衡数据的异常检测效果。

参考文献

[1] PwC's global economic crime and fraud survey 2018[EB/OL]. [2020-03-05]. https://www.pwc.com/gx/en/services/advisory/forensics/economic-crime-survey.html.

[2] HANSEN J V, MCDONALD J B, MESSIER W F, et al. A generalized qualitative-response model and the analysis of management fraud[J]. Management Science, 1996, 42(7): 1022-1032.

[3] SAHIN Y, BULKAN S, DUMAN E. A cost-sensitive decision tree approach for fraud detection[J]. Expert Systems with Applications, 2013, 40(15): 5916-5923.

[4] BHATTACHARYYA S, JHA S, THARAKUNNEL K, et al. Data mining for credit card fraud: a comparative study[J]. Decision Support Systems, 2011, 50(3): 602-613.

[5] TAE C M, HUNG P D. Comparing ML algorithms on financial fraud detection[C]//Proceedings of the 2nd International Conference on Data Science and Information Technology. Washington D.C., USA: IEEE Press, 2019: 15-26.

[6] JO T, JAPKOWICZ N. Class imbalances versus small disjuncts[J]. ACM SIGKDD Explorations News Letter, 2004, 6(1): 40-41.

[7] HE H, BAI Y, GARCIA E A, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning[C]// Proceedings of IEEE International Joint Conference on Neural Networks. Washington D.C., USA: IEEE Press, 2008: 125-136.

[8] LIU F T, TING K M, ZHOU Z. Isolation forest[C]// Proceedings of 2008 IEEE International Conference on Data Mining. Washington D.C., USA: IEEE Press, 2008: 1225-1229.

[9] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[10] SUN L, VERSTEEG S, BOZTAS S, et al. Detecting anomalous user behavior using an extended isolation forest algorithm[EB/OL]. [2020-03-05]. https://export.arxiv.org/pdf/1609.06676.

[11] ARYAL S, TING K M, WELLS J R, et al. Improving iForest with relative mass[J]. Advances in Knowledge Discovery and Data Mining, 2014, 8444(2): 510-521.

[12] HARIRI S, KIND M C, BRUNNER R J. Extended isolation forest[EB/OL]. [2020-03-05]. https://arxiv.org/pdf/1811.02141.pdf.

[13] XIAO Chunhui, SU Chen, BAO Congxiao, et al. Anomaly detection in network management system based on isolation forest[C]//Proceedings of 2018 Annual International Conference on Network and Information Systems for Computers. Washington D.C., USA: IEEE Press, 2018: 145-168.

[14] DING Z, MO Y, PAN Z. A novel software defect prediction method based on isolation forest[C]//Proceedings of 2019 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering. Washington D.C., USA: IEEE Press, 2019: 122-136.

[15] TSAI C F, LIN W C, HU Y H, et al. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection[J]. Information Sciences, 2018, 447: 47-54.

[16] KUBAT M, MATWIN S. Addressing the curse of imbalanced training sets: one-sided selection[C]// Proceedings of the 40th International Conference on Machine Learning. Washington D.C., USA: IEEE Press, 1997: 100-109.

[17] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2011, 16(1): 321-357.

[18] HUI H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[C]// Proceedings of 2005 International Conference on Advances in Intelligent Computing. Washington D.C., USA: IEEE Press, 2005: 11-23.

[19] HANSKUNATAI A. A new hybrid sampling approach for classification of imbalanced datasets[C]//Proceedings of 2018 International Conference on Computer and Communication Systems. Washington D.C., USA: IEEE Press, 2018: 67-71.

[20] CHEN Mincheng, YUAN Jingling, WANG Xiaoyan, et al. Parallelization of random forest algorithm based on discretization and selection of weak-correlation feature subspaces[J]. Computer Science, 2016, 43(6): 55-58, 90. (in Chinese)
陈旻骋,袁景凌,王啸岩,等. 基于弱相关化特征子空间选择的离散化随机森林并行分类算法[J]. 计算机科学,2016,43(6):55-58,90.

[21] XUAN S, LIU G, LI Z, et al. Random forest for credit card fraud detection[C]//Proceedings of 2018 IEEE International Conference on Networking, Sensing and Control. Washington D.C., USA: IEEE Press, 2018: 102-112.

[22] SEYEDHOSSEIN L, HASHEMI M R. Mining information from credit card time series for timelier fraud detection[C]// Proceedings of International Symposium on Telecommunications. Washington D.C., USA: IEEE Press, 2010: 117-123.

[23] ZHANG Jianjun. Undersampling near decision boundary for imbalance problems[C]//Proceedings of 2019 International Conference on Machine Learning and Cybernetics. Washington D.C., USA: IEEE Press, 2019: 11-25.

[24] ZHU H, YIN X, MA J, et al. Identifying the main paths of information diffusion in online social networks[J]. Physica A: Statistical Mechanics and Its Applications, 2016, 452: 320-328.

[25] LÜ Fang, TANG Fenghe, HUANG Junheng, et al. Frequent path discovery algorithm for financial network[J]. Chinese Journal of Network and Information Security, 2019, 5(5): 48-55. (in Chinese)
吕芳,汤丰赫,黄俊恒,等. 金融网络频繁链路发现算法[J]. 网络与信息安全学报,2019,5(5):48-55.

编辑 吴云芳