研究与开发

一种基于证据融合的类不平衡分类方法及其在网络流量识别中的应用

和红顺[1]，胡国良[2]，张志鹏[1]，柴鑫刚[1]，高静[1]
（1. 中国移动通信有限公司研究院，北京 100053；
2. 西北农林科技大学信息工程学院，陕西 杨凌 712199）

摘　要：类不平衡分类问题是机器学习中常见的挑战之一，且广泛存在于网络流量识别等实际场景应用中。针对类不平衡分类问题，设计了一种基于证据理论的融合类不平衡分类算法，通过使用不同的欠采样和过采样分类算法进行建模，利用多属性决策方法将多组不同的评价输出转换成证据函数，使用证据组合规则融合得到最终的识别结果。基于人工合成数据集和UCI基准数据集，采用神经网络与随机森林分类器进行交叉验证，并应用于网络流量识别任务。实验结果表明，所提出的算法能更好地应对类不平衡分类问题，在召回率、*F*1-score和*G*-mean等评价指标上均取得显著提升。

关键词：类不平衡分类；欠采样；过采样；证据理论；网络流量识别
中图分类号：TN925；TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.DXKX250278

A combined imbalanced classification approach based on D-S Evidence Theory and its application in network traffic recognition

He Hongshun[1], Hu Guoliang[2], Zhang Zhipeng[1], Chai Xingang[1], Gao Jing[1]
1. China Mobile Research Institute, Beijing 100053, China
2. College of Information Engineering, Northwest A&F University, Yangling 712199, China

Abstract: The imbalanced classification problem is one of the common challenges in machine learning and widely exists in practical applications such as network traffic recognition. To address this issue, a combined imbalanced classification approach based on D-S Evidence Theory was proposed. Different undersampling and oversampling classification algorithms were used for modeling, and multiple attribute decision making methods were used to convert differ-

收稿日期：2025−04−30；**修回日期：**2025−10−31
通信作者：胡国良，huguoliang1991@nwafu.edu.cn
基金项目：国家自然科学基金资助项目（No.32472002）；陕西省自然科学基础研究计划项目（No.2023-JC-QN-0645，No.2023-JC-QN-0684）；西安市科技计划项目（No.24NYGG0040）
Foundation Items: The National Natural Science Foundation of China (No.32472002), The Natural Science Fundamental Research Program of Shaanxi Province (No.2023-JC-QN-0645, No.2023-JC-QN-0684), The Xi'an Science and Technology Planning Project (No.24NYGG0040)

ent evaluation outputs into mass functions. Finally, evidence combination rules were used to combine the mass functions and obtain the final recognition results. Validation experiments were conducted on synthetic datasets and UCI benchmark datasets using neural network classifiers and random forest classifiers. The validated framework was then applied to real-world network traffic identification tasks. The experimental results demonstrate that the proposed approach significantly improves performance in addressing class-imbalanced classification problems, achieving notable enhancements in key evaluation metrics such as recall rate, *F*1-score, and *G*-mean value.

Key words: imbalance classification, undersampling, oversampling, D-S Evidence Theory, network traffic recognition

0　引言

在信息技术快速演进的时代背景下，数据规模呈爆炸式增长，如何科学挖掘和利用海量数据中的有效信息，成为重要的研究任务。然而，在很多实际应用中，类不平衡问题已渗透至多个关键领域，如医疗诊断中疾病样本稀缺（如罕见病数据）[1]、网络流量识别领域存在类不平衡问题[2-4]、模型识别与预测泛化性弱[5-6]等，均对分类模型的泛化能力提出了挑战。传统的监督学习方法，往往基于平衡的训练数据集这一假设，但针对类不平衡数据集，传统监督算法往往表现不佳。因此，如何更好地处理类不平衡数据分类问题、更好地发挥数据价值，成为十分有意义和重要的课题。

为了更好地处理类不平衡分类问题，学者们设计并提出了多种专门针对类不平衡问题的方法，主要可划分为3类：数据处理层面、特征层面和算法层面。

数据处理层面主要利用重采样等手段，尽可能将不同类别数据的数目平衡化，主要可分为欠采样、过采样和混合采样。欠采样通过减少数据中多数类别样本数目，从而使多数类和少数类中的样本数目趋于均衡。最简单的欠采样方法是从多类数据样本中随机抽取与少数类样本同等数目的数据样本。另一类欠采样方法是利用近邻规则等启发式规则[7-9]，去除噪声、边界以及冗余数据样本，达到降低数据不平衡的目的。此外，文献[10]利用聚类方法从多数类数据中筛选出部分数据，在降低数据不平衡度的同时，保留多数类原始数据分布的特征信息。过采样是指通过复制或人工生成等手段，增加少数类样本中的数据数目。合成少数类过采样技术（synthetic minority oversampling technique，SMOTE）[11]是一种常用的类不平衡过采样处理方案，其核心思想是通过随机选取少数类样本的k个同类近邻，利用该样本与近邻样本间线性差值来人工生成新的少数类样本。然而，该方法易引入噪声，且生成样本可能过于集中在某个区域。基于SMOTE算法，学者们提出一系列改进方案，如Borderline-SMOTE算法[12]通过筛选关键少数样本进行过采样，提升分类器边界区域附近的分类能力，可一定限度地降低少数类中噪声点的干扰。Safe-level-SMOTE算法[13]通过量化每个少数类样本的“安全等级”，优先在“安全区域”生成新样本，避免在边界模糊或噪声区域插值，以提升过采样的可靠性。此外，文献[14]利用生成对抗网络（generative adversarial network，GAN）生成少数类的伪样本来达到数据平衡。混合采样则利用组合欠采样和过采样的方案，融合两类算法的优势。例如，SSOMaj-SMOTE-SSOMin算法[15]通过3阶段协同优化实现数据分布平衡与分类性能提升，3阶段分别为：利用自组织映射（self-organizing map，SOM）对多数类样本进行欠采样；采用SMOTE增加少数类样本数量；对合成样本进行SOM优化。

特征层面是另一类解决类不平衡分类问题的重要手段，主要可划分为特征选择和特征提取两类

不同方法。特征选择类的类不平衡分类方法[16-18]，在利用邻域粗糙集等不同数据特征选择方法获取低维表示的同时，同步考虑分类模型更倾向于关注少数类，从而提高对少数类样本的分类性能。此外，特征提取的降维方法也被用于解决类不平衡分类问题。例如，文献[19]利用不同自编码器的特征学习方法，提取可使少数类和多数类分类性能更好的一组特征，以更好地处理类不平衡分类问题。

算法层面的核心思路是通过调整优化分类识别算法，从而提升其对少数类的识别能力，主要可划分为代价敏感学习和集成学习。代价敏感学习将不同类别赋予不同的误分代价，通过最小化总体误分代价目标函数来设计分类器，如代价敏感决策树[20]、代价敏感支持向量机[21]和代价敏感神经网络[22]等。集成学习[23]通过构建并组合多个基分类器，形成一个具有更强鲁棒性和泛化能力的强分类器，常见的集成方法包括Bagging方法、AdaBoost算法和梯度提升技术。其中，Bagging方法[24]通过有放回地采样生成多个不同的训练数据集，基于每个数据集训练一个基分类器，并采用投票机制综合各基分类器的结果，从而提升整体模型性能；AdaBoost算法[25]作为一种迭代式集成学习方法，通过在每一轮训练中调整样本权重，逐步强化对错误样本的学习，最终将多个弱分类器组合为一个强分类器；梯度提升技术[26]则利用负梯度方向近似基分类器的预测误差，并通过迭代方式不断修正残差，实现模型的逐步优化。多模态融合[27]旨在整合来自文本、图像、音频等不同模态的异构信息，突破单一模态所带来的语义局限性，提升系统对复杂任务的理解与表达能力。多模态融合根据技术特点主要分为数据级融合、特征级融合和决策级融合。其中，数据级融合在输入层直接合并不同模态的原始数据（如图像像素+文本词向量），形成统一输入送入模型处理；特征级融合是指各模态独立提取特征后，在中间层通过交互机制融合；决策级融合是指各模态独立生成预测结果，在输出层集成决策（如加权投票、证据理论等）。集成学习、多模态融合已在类不平衡识别问题中展现出良好效果。在处理类不平衡问题方面，一些代表性方法包括SMOTEBoost、EasyEnsemble与BalanceCascade。其中，SMOTEBoost[28]方法结合AdaBoost[25]与SMOTE，通过在训练过程中对少数类样本进行合成扩展，并赋予难以区分的少数类样本更高关注度，从而提高分类器对少数类的识别能力；EasyEnsemble采用无监督随机欠采样策略，将多数类样本划分为若干子集，每个子集与全部少数类样本组成训练集，分别训练基分类器，并通过集成提升整体性能；BalanceCascade[29]则采用有监督的级联式欠采样策略，逐步剔除已被正确分类的多数类样本，使后续训练更专注于难以区分的样本，动态优化训练集分布。综上，集成学习方法在类不平衡识别任务中展现出良好的适应性与有效性，为复杂场景下的智能决策系统提供了有力支撑。

基于欠采样、过采样等数据层面的类不平衡处理算法是效果直观且较为常用的方案，然而，该类方案有各自的缺陷。欠采样的方法通过某种衡量方式从多类样本中去除部分数据样本以达到数据平衡，此操作可能会丢失一些有助于分类的关键信息特征，造成一定程度的信息损失。过采样类方案，如SMOTE算法及其相关的改进方案，通过插值的方式生成人工合成样本，以提高少数类别的分类精度，但人工数据的引入可能会产生噪声及造成过拟合。混合采样方法将欠采样和过采样方案进行组合，既利用了欠采样和过采样的优势，同时也继承了它们的不足，因此如何更好地平衡这两类方案仍存在一定困难。为了充分发挥数据层面方法的优势，本文提出了一种新方法，该方法充分利用不同策略的优势及它们之间的差异互补性，运用决策级融合方法（证据理论）将

其高效组合，以更好地提升分类识别性能。

证据理论[30]，也称信念函数理论，它提供了一种重要的信息融合工具，可以在决策层面很好地表示和组合不确定信息。不同种类的方案均可一定限度地应对类别不平衡问题，且通常会提供一些互补的信息。本文尝试利用证据理论在决策级融合不同的数据层面类不平衡处理方法，以实现更好的类别不平衡分类识别性能。本文提出一种基于证据理论的融合欠采样和过采样类不平衡数据处理的方案，通过不同的欠采样和过采样方法处理数据集并通过分类器进行建模，将多组分类识别结果分别转换成证据函数，利用证据组合规则融合得到最终的识别结果。传统方法（如SMOTE或损失加权）仅通过样本数量或权重调整缓解不平衡，无法从本质上解决类不平衡数据引发的不确定性。证据理论在类不平衡问题中的应用优势主要体现在以下两个方面。

（1）证据理论通过信任函数（belief function）与似然函数（plausibility function）构建置信区间[Bel(A),Pl(A)]，以界定不确定性的范围，形成概率区间估计而非点估计，从而更有效地建模和度量不确定性。

（2）证据理论通过证据组合规则对多个不同模型生成的证据函数进行融合，并采用第六类比例冲突分配（proportional conflict redistribution No. 6，PCR6）规则处理证据间的冲突，从而降低不确定性和噪声的影响，有效解决了传统方法难以应对的不确定性与噪声问题。在多组数据集中的实验以及在KDD Cup网络流量分类中的应用表明，所提出的算法能更好地应对数据类不平衡问题并能有效提升分类性能，从而更好地应用于网络流量分类识别任务。

1　相关工作

本文提出的融合类不平衡处理方案综合了过采样与欠采样算法。为此，本节首先简要介绍方案所使用的3种关键算法：Borderline-SMOTE[12]、Safe-level-SMOTE[13]以及基于K-means的欠采样算法[10]。

1.1　Borderline-SMOTE算法

Borderline-SMOTE算法是基于SMOTE算法的改进工作，可一定限度地降低少数类中噪声点的干扰，通过筛选关键少数样本进行过采样，提升分类器边界区域附近的分类能力。该算法的主要流程如下。

步骤1　少数类样本类别划分。对于每个少数类样本，计算其k近邻，根据其近邻情况划分为以下3类。

（1）噪声点（全为多数类近邻）：不参与采样。

（2）边界点（半数以上为多数类近邻）：优先生成新样本。

（3）安全点（半数以上为少数类近邻）：不参与过采样。

步骤2　边界点过采样。针对边界点随机选取目标样本，从其少数类x_{d}的近邻中随机选取近邻样本x_{neighbor}，通过线性插值生成新样本：

$$x_{\mathrm{new}}=x_{\mathrm{d}}+r\times(x_{\mathrm{neighbor}}-x_{\mathrm{d}}) \tag{1}$$

其中，r为[0,1]区间的随机数。

步骤3　重复直至平衡。根据设定的过采样率，重复步骤2，直至少数类样本数量与多数类平衡。

1.2　Safe-level-SMOTE算法

Safe-level-SMOTE算法是针对传统SMOTE和Borderline-SMOTE的进一步优化，核心目标是解决噪声样本干扰和自适应调整过采样位置。该算法通过量化每个少数类样本的“安全等级”，优先在“安全区域”生成新样本，避免在边界模糊或噪声区域插值，提升过采样的可靠性。该算法的主要流程如下。

步骤1　计算少数类样本的安全等级。对每个少数类样本x_i，计算其k近邻。根据近邻中少数

类和多数类的样本数量，定义安全等级（safe level）：

$$\mathrm{SL}(x_i)=\frac{\mathrm{min\_cout}}{\mathrm{min\_cout}+\mathrm{maj\_cout}} \tag{2}$$

其中，min_cout表示k近邻中属于少数类的样本数，maj_cout表示k近邻中属于多数类的样本数。

步骤2 筛选安全样本对。对每个x_i，从少数类近邻中选择满足$\mathrm{SL}(x_j)\geqslant\mathrm{SL}(x_i)$的样本$x_j$，形成安全样本对集合。若$x_i$无安全近邻，则跳过该样本。

步骤3 自适应生成新样本。对每个安全样本对(x_i,x_j)，根据两者的安全等级差异生成新样本。若$\mathrm{SL}(x_j)=\mathrm{SL}(x_i)$，则在$x_i$和$x_j$的连线上随机插值生成样本：

$$x_{\mathrm{new}}=x_i+r\times(x_j-x_i),r\in[0,1] \tag{3}$$

若$\mathrm{SL}(x_j)>\mathrm{SL}(x_i)$，则新样本靠近安全等级更高的$x_i$，插值计算式调整为：

$$x_{\mathrm{new}}=x_i+r\times(x_j-x_i),r\in[0,0.5] \tag{4}$$

步骤4 重复生成直至平衡。根据过采样倍率，重复步骤2和步骤3，直至少数类样本数量达到目标。

1.3 *K*-means欠采样算法

K-means聚类欠采样通过将多数类样本划分为与少数类样本数量相等的簇，从每个簇中抽取代表性样本（如簇中心或近邻样本），实现数据分布平衡。其核心优势在于保留全局分布特征，同时避免随机欠采样的盲目性。该算法的主要流程如下。

步骤1 *K*-means聚类。利用*K*-means算法对类不平衡数据集做聚类，取少数类的样本个数为聚类簇数K，可得到K个样本簇。

步骤2 样本抽取。使用聚类的各簇质心作为代表样本，得到$X_{\mathrm{maj_reduced}}$。

步骤3 合并数据集。将欠采样后的多数类样本$X_{\mathrm{maj_reduced}}$与原始少数类X_{min}合并，形成平衡数据集：

$$X_{\mathrm{balanced}}=X_{\mathrm{maj\_reduced}}\cup X_{\mathrm{min}} \tag{5}$$

2 基于证据融合的类不平衡分类算法

不同种类的方案均可一定限度地应对类不平衡问题，且不同方案往往蕴含着一定限度的互补性信息。为此，本文提出一种基于证据理论的融合欠采样和过采样类不平衡处理方案，通过融合具备一定差异性的类不平衡处理方法，以提升类不平衡分类性能。本节首先简要介绍所采用的证据理论的基本框架，随后阐述基于证据融合的类不平衡分类算法的具体实现及算法流程。

2.1 证据理论简介

设集合$\Theta=\{\theta_1,\theta_2,\cdots,\theta_n\}$为辨识框架（frame of discernment，FOD），2^Θ是Θ所有子集构成的集合。若$m:2^\Theta\to[0,1]$满足：

$$\sum_{F\subseteq\Theta}m(F)=1,m(\varnothing)=0 \tag{6}$$

则称m为辨识框架上的基本信度赋值（basic belief assignment，BBA），也称mass函数。若满足$F\subseteq\Theta,m(F)>0$，则F被称为焦元。

对于辨识框架Θ中的某个命题F，定义其信任函数和似然函数分别为$\mathrm{Bel}(F)$和$\mathrm{Pl}(F)$：

$$\mathrm{Bel}(F)=\sum_{B\subseteq F}m(B) \tag{7}$$

$$\mathrm{Pl}(F)=\sum_{B\cap F\neq\varnothing}m(B) \tag{8}$$

区间$[\mathrm{Bel}(F),\mathrm{Pl}(F)]$可用于表示对$F$支持的不确定程度。

Dempster组合规则（简称DS组合规则）是对两个独立mass函数进行融合的方法。设辨识框架Θ上两个独立的mass函数分别为m_1和m_2，命题F满足$\forall F\subseteq\Theta,F\neq\varnothing$，则$m_1$和$m_2$的DS组合规则为：

$$m_1\oplus m_2(F)=\frac{1}{1-K}\sum_{B\cap C=F}m_1(B)m_2(C) \tag{9}$$

其中，$K=\sum_{B\cap C=\varnothing} m_1(B)m_2(C)$称为冲突项。多个证据组合时，DS组合规则满足交换律和结合律。

在获得融合证据后，由证据至概率的转换是实现融合决策的重要步骤。文献[31]定义的Pignistic概率（一种赌博概率）转换是将mass函数转换为概率（即子集F发生的概率值）的一种经典方法，定义如下。

$$\mathrm{BetP}(F)=\sum_{B\subseteq\Theta}\frac{|F\cap B|}{|B|}\frac{m(B)}{1-m(\varnothing)} \tag{10}$$

针对证据中存在的高冲突问题，可采用PCR6组合规则[32]进行融合。PCR6的核心思想是当多个证据源对同一命题的信任分配存在冲突时，将冲突质量按各证据源对该命题的原始信任比例重新分配，而非简单忽略或全局归一化。当存在冲突$K=\sum_{B\cap C=\varnothing} m_1(B)m_2(C)$，将冲突$K$各证据源对冲突焦元的原始信任比例拆分：

$$m_{\mathrm{PCR6}}(F)=\sum_{B\cap C\cdots=F} m_1(B)m_2(C)\cdots+\sum_{i=1}^{k}\frac{m_i(A)^2}{\sum_{j=1}^{k}m_j(A)}K \tag{11}$$

式（11）等号右边第2项为冲突质量的再分配部分。PCR6不强制归一化结果，保留冲突质量的再分配过程，避免DS组合规则高冲突导致“归一化失真”问题。

2.2　基于证据融合的类不平衡分类算法

针对现有类不平衡识别方法的缺陷，本文提出一种基于证据融合的类不平衡分类算法（evidence fusion-based class-imbalanced classification algorithm，ECIC），该算法流程如图1所示。针对类不平衡数据训练样本集，本文首先分别采用不同的过采样方法、欠采样方法进行数据处理，并使用分类器进行训练，然后利用分类器输出的多组不同结果构建评价矩阵，利用证据推理的谨慎有序加权平均证据推理（cautious ordered weighted averaging and evidential reasoning，COWA-ER）[33]生成多组证据函数，最终融合这些证据函数完成识别。COWA-ER是一种面向多属性决策的不确定性推理方法，其核心思想是在多属性决策框架下，针对每种属性，分别以最悲观和最乐观的态度构造两个隶属度函数。在此基础上，进一步根据构造好的隶属度函数生成对应两种态度（最乐观和最悲观）的mass函数。最后，基于DS组合规则以及概率转换方法，做出最终的决策。

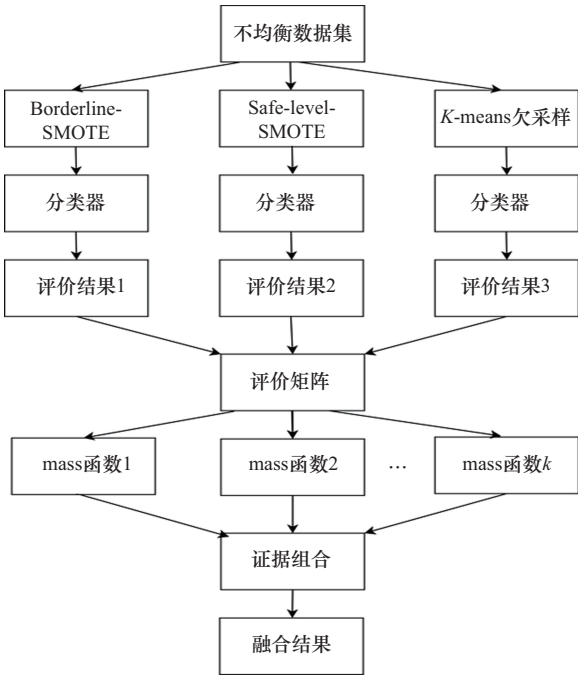

图1　基于证据融合的类不平衡分类算法流程

根据前文所述，基于证据融合的类不平衡分类方法的步骤如下。

步骤1　针对类不平衡的训练数据集D（类别标签为$\theta_j(j=1,2,\cdots,n)$），利用Borderline-SMOTE、Safe-level-SMOTE以及基于K-means的欠采样方法进行类不平衡处理，得到3组平衡后的数据集D_{b1}、D_{b2}、D_{b3}，并分别训练分类器模型A_1、A_2、A_3，基于分类器对待测样本的3组不

同评价值可构造评价矩阵：

$$\boldsymbol{C}=\begin{array}{c} \\ A_1 \\ \vdots \\ A_i \\ \vdots \\ A_k \end{array}\begin{array}{c} \begin{array}{cccccc} \theta_1 & \cdots & \theta_j & \cdots & \theta_n \end{array} \\ \begin{bmatrix} v_{11} & \cdots & v_{1j} & \cdots & v_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ v_{i1} & \cdots & v_{ij} & \cdots & v_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ v_{k1} & \cdots & v_{kj} & \cdots & v_{kn} \end{bmatrix} \end{array} \tag{12}$$

其中k取3（即训练3个分类器模型）。

分别以悲观的方式和乐观的方式选出待测样本对各类别的加权评价值，即选出样本对各类别评价的最小值和最大值，可以得到如下矩阵：

$$\boldsymbol{E}=\begin{bmatrix} \min(v_{i1}) & \min(v_{i2}) & \cdots & \min(v_{in}) \\ \max(v_{i1}) & \max(v_{i2}) & \cdots & \max(v_{in}) \end{bmatrix} \tag{13}$$

步骤2 将矩阵$\boldsymbol{E}$中的每个元素对$\boldsymbol{E}$的最大值归一化，得到矩阵$\boldsymbol{E}_N$，即：

$$\boldsymbol{E}_N=\begin{bmatrix} a_1 & \cdots & a_j & \cdots & a_n \\ b_1 & \cdots & b_j & \cdots & b_n \end{bmatrix} \tag{14}$$

其中，$a_j=\dfrac{\min_i(v_{ij})}{\max(\boldsymbol{E})}$，$b_j=\dfrac{\max_i(v_{ij})}{\max(\boldsymbol{E})}$。

步骤3 根据$\boldsymbol{E}_N$生成mass函数。$\boldsymbol{E}_N$中以悲观方式得到的a_j用以表征对θ_j的信任度，即$\mathrm{Bel}(\theta_j)=a_j$；以乐观方式得到的$b_j$用以表征对$\theta_j$的似真度，即$\mathrm{Pl}(\theta_j)=b_j$。$[\mathrm{Bel}(\theta_j),\mathrm{Pl}(\theta_j)]$的长度表示不确定程度，即$b_j-a_j$，由此可以生成mass函数：

$$\begin{aligned} &m_j(\theta_j)=a_j \\ &m_j(\overline{\theta_j})=1-b_j \\ &m_j(\theta_j\cup\overline{\theta_j})=m_j(\Theta)=b_j-a_j \end{aligned} \tag{15}$$

步骤4 不同的证据函数$m_j(\cdot)$（$j=1,\cdots,n$）依据证据组合规则（式（9）的DS组合规则或式（11）的PCR6组合规则）进行组合，得到融合证据函数，然后应用式（10）的Pignistic概率转换将证据函数转换为概率并进行决策。

3 实验与分析

为了验证提出的基于证据理论的融合类不平衡分类算法的合理和有效性，本文分别利用人工生成的数据集和UCI数据集进行实验对比及分析，并将其应用于网络流量识别任务。

3.1 实验评估指标

实验评估采用类不平衡分类中常用的召回率（Recall）、$F1$-score以及G-mean这3个指标评估提出的证据融合类不平衡分类算法的性能。这3个指标是基于混淆矩阵得到的。混淆矩阵见表1。

表1 混淆矩阵

预测标签	真实类别	
	正类（positive）	负类（negative）
正类（true）	真正例（TP）	假正例（FP）
负类（false）	假负例（FN）	真负例（TN）

表1中，TP是指实际为正类，预测也为正类；FP是指实际为负类，但预测为正类（误报）；TN是指实际为负类，预测也为负类；FN是指实际为正类，但预测为负类（漏报）。

Recall指实际是正例，分类器预测为正例的比例，计算式为：

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{16}$$

精确率（Precision）指分类器预测为正例的样本中，实际为正例的比例，计算式为：

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{17}$$

$F1$-score为Precision和Recall的调和平均数，计算式为：

$$F1\text{-score}=\frac{2\times\mathrm{Presion}\times\mathrm{Recall}}{\mathrm{Presion}+\mathrm{Recall}} \tag{18}$$

G-mean是指所有类别的准确率的几何平均数，计算式为：

$$G\text{-mean}=\sqrt{\frac{\mathrm{TP}}{\mathrm{TP+FN}}\times\frac{\mathrm{TN}}{\mathrm{TN+FP}}} \tag{19}$$

3.2 基于人工数据集的仿真实验评估

本文利用人工数据集对所提出的算法进行验证，依据式（20）构造人工数据集：

$$\begin{cases} l(1)=\alpha\cos(\phi)+\beta \\ l(2)=\alpha\sin(\phi) \end{cases} \tag{20}$$

其中，α为[0,1]区间的随机数，β取0时构建类别A的数据样本，β取0.6时构建类别B的数据样本，其中A类数据包含200个样本，B类数据包含50个样本。人工数据集示意图如图2所示。

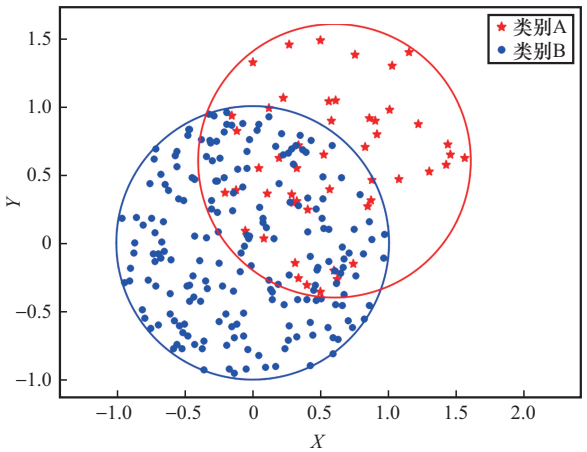

图2　人工数据集示意图

使用人工数据集进行实验时，将数据随机划分为5份，选取其中的3份作为训练集，其余2份作为测试集，且训练集和测试集的数据不平衡比例均与原始数据保持一致。实验随机重复30轮取平均。实验采用全连接的神经网络（fully connected neural network，FCNN）、随机森林（random forest，RF）作为分类器。FCNN层数为4层（输入层+2个隐藏层+输出层），隐藏层神经元数分别为128、64，激活函数为ReLU。RF树的数量为200，最大深度为10。对比算法包括SMOTE、ROS、*K*-means_US、Safe-level-SMOTE、BorSmote和RUSBoost共6种类不平衡处理算法，具体如下。

（1）SMOTE[11]：随机选择少数类中的样本数据，通过线性插值的方式生成人工样本，使得少数类样本与多数类样本趋于平衡。其优势为样本多样性高，但对噪声敏感，且对于中等维度的连续数据处理效果较好。

（2）ROS：即随机过采样方法，随机复制少数类中的样本数据，直到少数类中的样本数据与多数类样本数据达到平衡。该算法实现简单，对小样本数据十分有效，但易产生过拟合，适用于较为稀疏的小样本数据。

（3）*K*-means_US欠采样[10]：*K*-means聚类欠采样通过将多数类样本划分为与少数类样本数量相等的簇，从每个簇中抽取代表性样本（如簇中心或近邻样本），实现数据分布平衡。其优势为保留全局分布、鲁棒性好，但计算复杂度高，处理多数类结构清晰的数据效果更好。

（4）Safe-level-SMOTE[13]：通过量化每个少数类样本的“安全等级”，优先在“安全区域”利用线性插值生成新样本。其优势为抗噪性强、生成样本质量高，对噪声显著的数据处理效果好。

（5）BorSmote[12]：筛选在边界中的少数类样本进行线性插值，生成人工的少数类样本。其优势在于算法强化了分类边界，提升了边界处难样本识别，其局限性是对边界噪声敏感，主要适用于边界模糊的数据场景如欺诈检测等。

（6）RUSBoost（随机欠采样RUS+AdaBoosting）：其核心思想是在Boosting的每轮迭代中引入随机欠采样，动态调整数据分布以平衡类别权重。其主要优势是具备强泛化能力，欠采样可降低计算复杂度，适合大规模数据。

本文采用Recall、*F*1-score以及*G*-mean这3个指标来评估提出的证据融合的类不平衡分类算法的性能。FCNN和RF分类器识别结果分别如图3、图4所示，相较于已有的类不平衡分类算法SMOTE、ROS、*K*-means_US、Safe-level-SMOTE、BorSmote和RUSBoost，本文提出的基于证据融合的类不平衡分类算法（ECIC_DS（使用DS组合规则）、ECIC_PCR6（使用PCR6组合规则））获得了更好的识别效果，尤其是在*G*-mean指标上，所得到的识别性能远远高于对比算法。

3.3　基于UCI数据集的实验

本节使用Bupa、Pima、Breast、Diabetes、Hearts

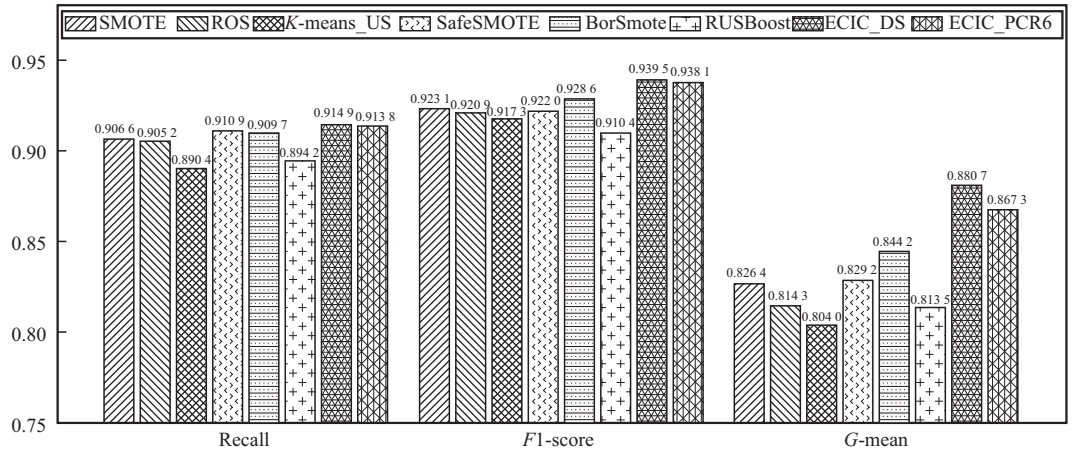

图3 FCNN分类器识别结果

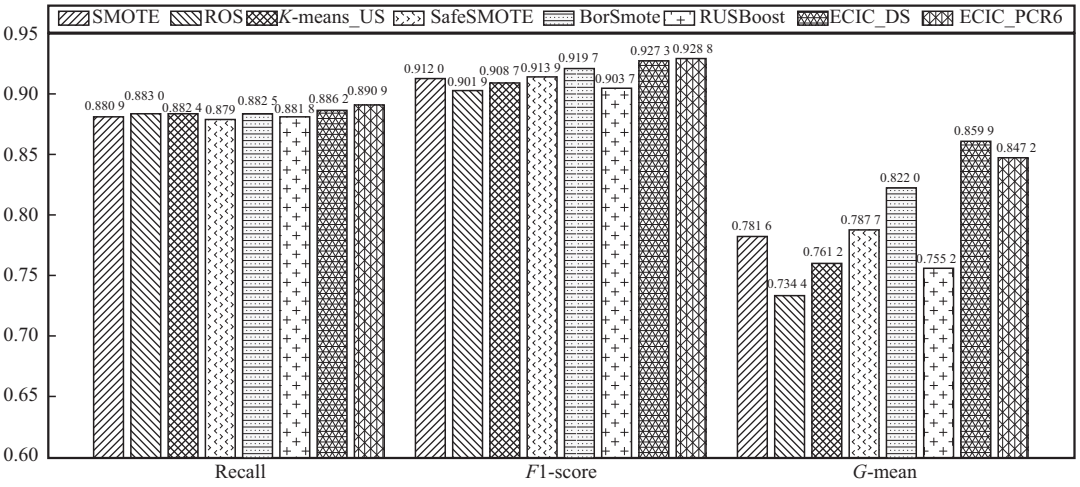

图4 RF分类器识别结果

和Blood共6个UCI公开数据集[34]进行仿真。UCI数据集信息见表2。同样地，对比SMOTE、ROS、*K*-means_US、Safe-level-SMOTE、BorSmote和RUSBoost共6种类不平衡分类算法。

表2 UCI数据集信息

数据集	特征数	样本数	不平衡比
Pima	8	768	1.87
Bupa	7	345	1.38
Diabetes	8	768	1.87
Breast	9	286	2.36
Hearts	13	270	1.25
Blood	4	748	3.61

实验将数据随机划分为5份，选取其中的3份作为训练集，其余2份作为测试集，且训练集和测试集的数据不平衡比例均与原始数据保持一致。实验随机重复30轮取平均。实验采用FCNN、RF两类分类器。FCNN分类器网络层数为4层（输入层+2个隐藏层+输出层），隐藏层神经元数分别为12 864，激活函数为ReLU。RF分类器树的数量为200，最大深度为10。实验同样采用Recall、*F*1-score以及*G*-mean这3个指标来评估本文所提出的证据融合的类不平衡分类算法的性能。采用FCNN分类器的不同指标平均结果对比见表3～表5，采用RF分类器的不同指标平均结果对比见表6～表8，其中，最优的数据结果已加粗。

表 3　UCI 数据集 Recall 平均结果对比（FCNN 分类器）

算法	数据集					
	Pima	Bupa	Diabetes	Breast	Hearts	Blood
SMOTE	0.717 3	0.593 8	0.783 4	0.602 6	0.749 3	0.784 2
ROS	0.727 4	0.598 9	0.769 0	0.596 9	0.693 1	0.781 0
K-means_US	0.726 5	0.620 2	0.762 3	0.596 7	0.762 5	0.811 2
Safe-level-SMOTE	0.728 7	0.616 9	0.778 5	0.605 7	0.722 3	0.783 3
BorSmote	0.728 7	0.592 6	0.765 3	0.580 6	0.683 9	0.764 2
RUSBoost	0.730 5	0.614 6	0.768 5	0.592 8	0.764 0	0.804 2
ECIC_DS	0.733 7	0.622 1	**0.786 9**	**0.607 9**	0.782 8	**0.819 4**
ECIC_PCR6	**0.736 5**	**0.625 0**	0.784 4	0.606 2	**0.784 0**	0.818 2

表 4　UCI 数据集 *F*1-score 平均结果对比（FCNN 分类器）

算法	数据集					
	Pima	Bupa	Diabetes	Breast	Hearts	Blood
SMOTE	0.739 7	0.645 8	0.798 3	0.656 2	0.756 3	0.877 0
ROS	0.743 2	0.642 8	0.798 2	0.652 1	0.722 1	0.875 0
K-means_US	0.758 6	0.640 4	0.789 3	0.651 6	0.782 8	0.871 9
Safe-level-SMOTE	0.761 9	0.644 9	0.786 2	0.650 1	0.743 6	0.874 4
BorSmote	0.768 9	0.650 7	0.781 9	0.664 5	0.716 9	0.862 4
RUSBoost	0.771 3	0.650 2	0.808 7	0.654 4	0.789 2	0.873 2
ECIC_DS	0.770 3	**0.667 1**	**0.821 7**	**0.668 7**	0.793 5	**0.893 6**
ECIC_PCR6	**0.783 4**	0.659 3	0.816 2	0.662 9	**0.797 9**	0.886 4

表 5　UCI 数据集 *G*-mean 平均结果对比（FCNN 分类器）

算法	数据集					
	Pima	Bupa	Diabetes	Breast	Hearts	Blood
SMOTE	0.699 7	0.704 1	0.631 4	0.709 0	0.782 7	0.828 4
ROS	0.718 4	0.697 2	0.609 5	0.704 8	0.740 6	0.818 2
K-means_US	0.728 5	0.714 5	0.600 1	0.704 6	0.802 6	0.797 4
Safe-level-SMOTE	0.726 2	0.708 5	0.614 7	0.707 5	0.766 2	0.780 6
BorSmote	0.728 8	0.718 5	0.589 5	0.705 9	0.736 7	0.804 6
RUSBoost	0.712 8	0.716 9	0.631 3	0.697 0	0.772 6	0.781 4
ECIC_DS	**0.733 6**	0.729 4	0.662 6	0.712 0	0.816 8	**0.872 0**
ECIC_PCR6	0.733 4	**0.734 7**	**0.679 7**	**0.715 7**	**0.817 8**	0.865 5

由表 3～表 5 可知，当使用 FCNN 分类器时，ECIC_DS、ECIC_PCR6 算法在 Bupa、Hearts 和 Blood 这 3 个数据集的 Recall 指标结果明显优于对比算法；ECIC_DS、ECIC_PCR6 算法在 Pima、Breast、Hearts 和 Blood 数据集上 *F*1-score 指标相较于对比算法，表现更为优异；不同数据集的 *G*-mean 结果也高于对比算法。由表 6～表 8 可知，当使用 RF 分类器时，Bupa、Hearts 和 Blood 数据集的 *F*1-score、*G*-mean 指标结果均明显优于对比算法；不同数据集的 Recall 指标结果也优于对比算法。实验结果表明，相较于已有的类不平衡分类算法 SMOTE、ROS、*K*-means_US、Safe-level-

表6　UCI数据集Recall平均结果对比（RF分类器）

算法	数据集					
	Pima	Bupa	Diabetes	Breast	Hearts	Blood
SMOTE	0.709 6	0.630 2	0.770 5	0.798 9	0.829 5	0.807 3
ROS	0.715 2	0.627 6	0.774 2	0.802 5	0.805 2	0.809 8
K-means_US	0.705 3	0.621 8	0.764 4	0.796 6	0.834 4	0.809 7
Safe-level-SMOTE	0.713 3	0.627 6	0.773 0	0.796 9	0.823 0	0.803 8
BorSmote	0.718 9	0.602 5	0.767 8	0.778 8	0.821 3	0.801 7
RUSBoost	0.712 4	0.625 0	0.772 3	0.801 4	0.811 4	0.798 5
ECIC_DS	0.721 9	**0.633 9**	0.776 1	0.815 4	0.837 3	0.814 2
ECIC_PCR6	**0.726 5**	0.631 1	**0.784 6**	**0.822 4**	**0.838 6**	**0.815 8**

表7　UCI数据集*F*1-score平均结果对比（RF分类器）

算法	数据集					
	Pima	Bupa	Diabetes	Breast	Hearts	Blood
SMOTE	0.758 5	0.665 4	0.836 8	0.836 9	0.798 8	0.844 9
ROS	0.752 4	0.661 8	0.824 4	0.837 6	0.794 3	0.846 3
K-means_US	0.755 5	0.668 7	0.829 5	0.839 0	0.808 0	0.848 1
Safe-level-SMOTE	0.761 1	0.669 4	0.836 1	0.837 4	0.806 9	0.847 7
BorSmote	0.764 7	0.668 4	0.830 2	0.837 3	0.800 7	0.848 9
RUSBoost	0.760 9	0.669 3	0.836 9	0.816 5	0.801 9	0.837 7
ECIC_DS	**0.767 7**	**0.676 8**	0.838 6	0.841 7	0.815 5	0.850 5
ECIC_PCR6	0.766 7	0.675 7	**0.839 7**	**0.844 2**	**0.825 3**	**0.852 1**

表8　UCI数据集*G*-mean平均结果对比（RF分类器）

算法	数据集					
	Pima	Bupa	Diabetes	Breast	Hearts	Blood
SMOTE	0.694 5	0.716 5	0.704 6	0.758 6	0.830 8	0.612 8
ROS	0.688 9	0.724 5	0.668 7	0.759 7	0.818 7	0.625 1
K-means_US	0.685 8	0.722 0	0.679 7	0.762 2	0.837 5	0.623 3
Safe-level-SMOTE	0.696 3	0.726 1	0.693 9	0.760 3	0.831 7	0.616 0
BorSmote	0.700 3	0.714 5	0.680 1	0.761 1	0.827 6	0.615 5
RUSBoost	0.690 4	0.709 7	0.690 2	0.746 5	0.822 8	0.627 9
ECIC_DS	0.708 2	0.728 0	**0.709 9**	**0.766 5**	0.840 3	**0.634 3**
ECIC_PCR6	**0.710 5**	**0.729 4**	0.708 1	0.765 3	**0.847 1**	0.632 7

SMOTE、BorSmote和RUSBoost，本文提出的基于证融合的类不平衡分类算法（ECIC_DS、ECIC_PCR6）在处理不平衡数据时可获得更好的识别效果。

3.4　网络流量分类中的应用

KDD Cup数据集[35]是包含不同攻击类型的网络流量数据，本节选取其中的U2R、R2L、Probe这3种类型数据进行实验测试。KDD Cup网络流量信息见表9。由表9可知，3类网络攻击数据存在类不平衡的问题，特别是U2R数据相较于Probe、R2L两类数据，所占比例仅为0.98%。实验时将区分Probe、R2L和U2R的多分类任务

划分为{Probe，others}、{R2L，others}和{U2R，others}3个二分类任务。

表9　KDD Cup网络流量信息

攻击类型	数据流数目	特征维度	所占比例
Probe	4 107	41	77.71%
R2L	1 126	41	21.31%
U2R	52	41	0.98%

本文利用提出的算法和SMOTE、ROS、*K*-means_US、Safe-level-SMOTE、BorSmote和RUSBoost这6种类不平衡分类算法进行对比实验。实验将数据随机划分为5份，选取其中的3份作为训练集，其余2份作为测试集，且训练集和测试集的数据不平衡比例均与原始数据保持一致。实验随机重复30轮取平均。实验采用FCNN、RF两类分类器。分类器参数设置与UCI数据集实验保持一致。FCNN分类器的U2R、R2L、Probe流量类型识别结果对比如图5～图7所示，RF分类器的U2R、R2L、Probe流量类型识别结果对比如图9～图10所示。

本文采用Recall、*F*1-score以及*G*-mean这3个指标来评估所提出的基于证据融合的类不平衡分类算法的性能。由图5～图7可知，在U2R流量识别任务中，该算法的提升最为明显（U2R流量占比最少，不平衡度最高），ECIC_PCR6算法在Recall指标上达到了0.712 2，显著优于对比算法；在R2L流量识别任务中，所提出的ECIC_DS

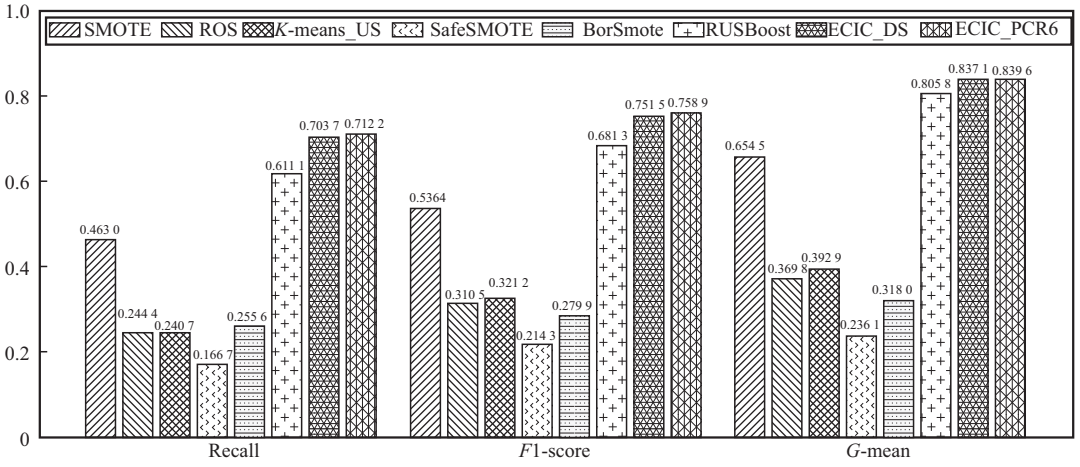

图5　FCNN分类器的U2R流量类型识别结果对比

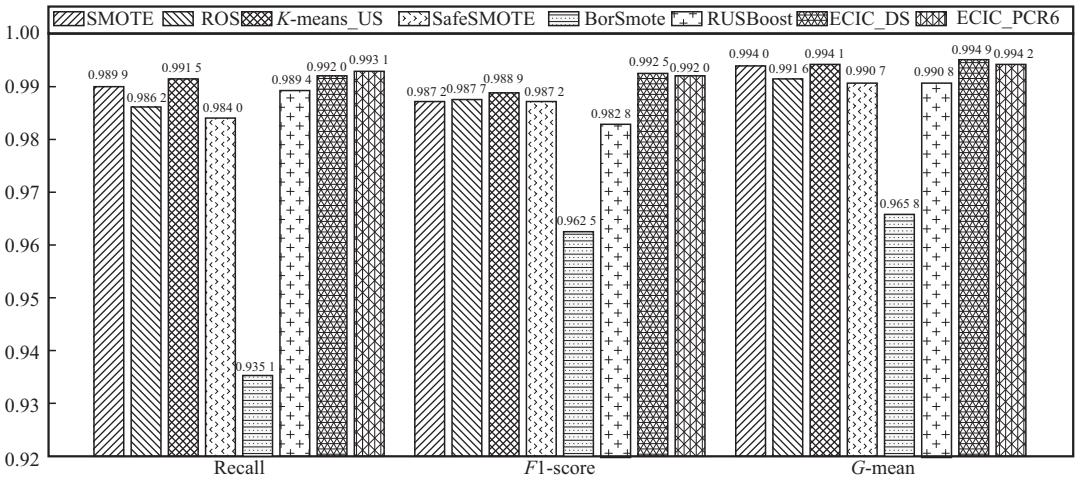

图6　FCNN分类器的R2L流量类型识别结果对比

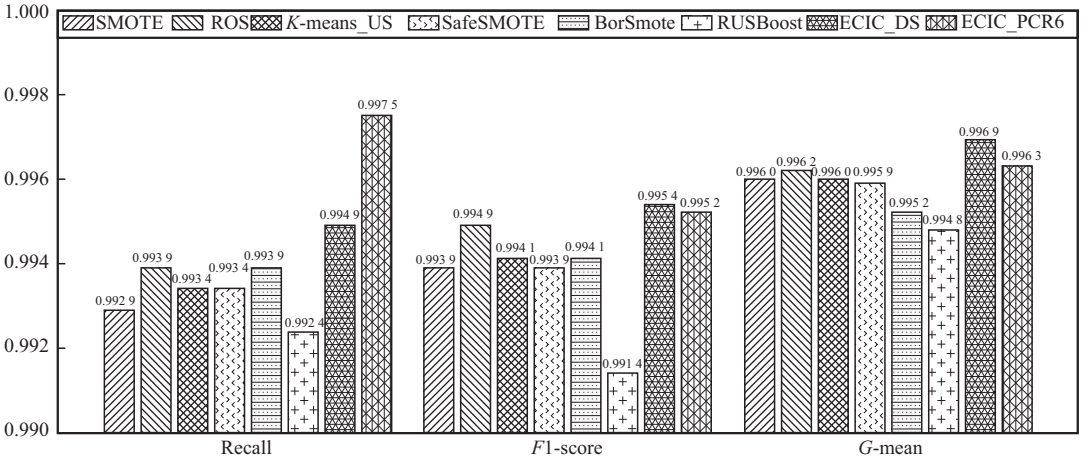

图7　FCNN 分类器的 Probe 流量类型识别结果对比

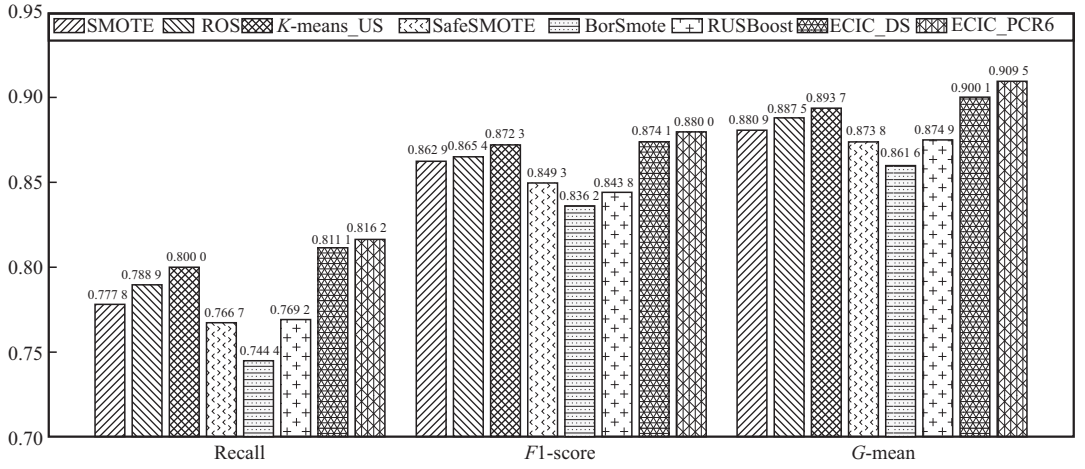

图8　RF 分类器的 U2R 流量类型识别结果对比

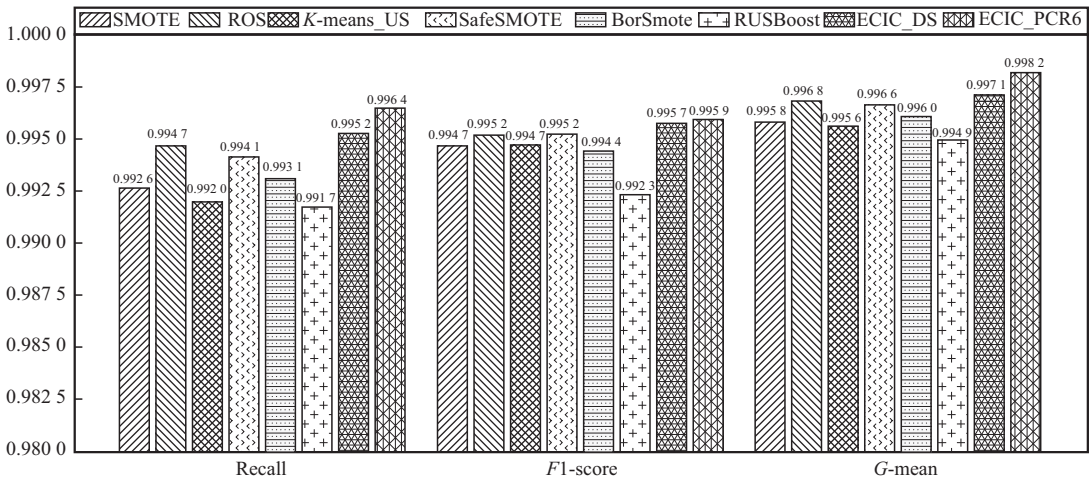

图9　RF 分类器的 R2L 流量类型识别结果对比

和 ECIC_PCR6 算法在 Recall、*F*1-score 以及 *G*-mean 这 3 个指标上的表现均优于对比算法；在 Probe 流量识别任务中，由于其样本数目占比最高，ECIC_DS 和 ECIC_PCR6 算法的识别效果与

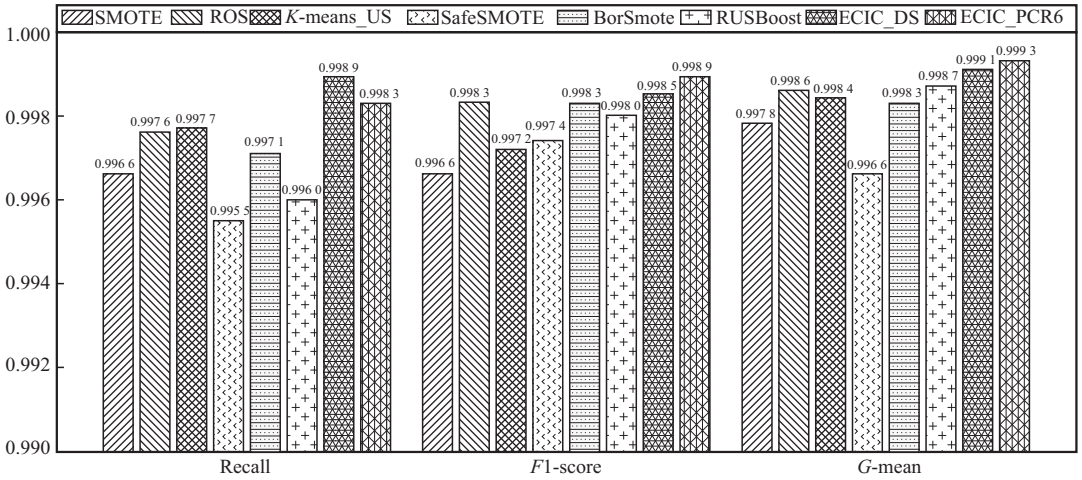

图 10　RF 分类器的 Probe 流量类型识别结果对比

对比算法在 Recall、*F*1-score 以及 *G*-mean 这 3 个指标上持平或略高。由图 8～图 10 可知，对于占比最少的 U2R 流量识别任务，ECIC_DS 和 ECIC_PCR6 算法的识别效果相较于对比算法提升更为明显；在 R2L 和 Probe 流量识别任务中，所提出算法的性能略高于对比算法。综合来看，基于证据理论的融合类不平衡分类算法（ECIC_DS、ECIC_PCR6）在 KDD Cup 类不平衡流量识别数据集上表现出了更优的识别性能。

4　结束语

本文提出一种基于证据融合的类不平衡分类方法，采用不同的欠采样和过采样方法处理数据集并通过分类器进行建模，同时将多组分类识别结果分别转换成证据函数，利用证据组合规则融合得到最终的识别结果。在多组人工数据集、UCI 数据集上进行测试的结果表明，本文提出的算法能更好地应对数据类不平衡问题并有效提升分类识别性能。本文将该算法进一步在网络流量数据集上进行识别应用，为类不平衡流量识别任务提供一定的理论识别依据，具备一定的应用推广价值。未来将进一步研究新型类不平衡分类方法的融合以及融合算法的优化，持续提升算法性能和泛化能力，并将其应用到 6G 业务识别等更多领域。

参考文献：

[1] Zhou D D, Xu Q, Wang J, et al. Alleviating class imbalance problem in automatic sleep stage classification[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 4006612.

[2] Dong S, Xia Y J. Network traffic identification in packet sampling environment[J]. Digital Communications and Networks, 2023, 9(4): 957-970.

[3] Dong S, Xia Y J, Peng T. Traffic identification model based on generative adversarial deep convolutional network[J]. Annals of Telecommunications, 2022, 77(9): 573-587.

[4] Dong S, Xia Y J, Wang T. Network abnormal traffic detection framework based on deep reinforcement learning[J]. IEEE Wireless Communications, 2024, 31(3): 185-193.

[5] 阿克弘, 胡晓东. 基于 GAN 数据重构的电信用户流失预测方法[J]. 电信科学, 2023, 39(3): 135-142.
A K H, Hu X D. GAN data reconstruction based prediction method of telecom subscriber loss[J]. Telecommunications Science, 2023, 39(3): 135-142.

[6] 余立, 李哲, 高飞, 等. 改进自训练模型在业务质差用户识别中的应用[J]. 电信科学, 2021, 37(10): 136-142.
Yu L, Li Z, Gao F, et al. Application of improved self-training model in the identification of users with poor service quality[J]. Telecommunications Science, 2021, 37(10): 136-142.

[7] Hart P. The condensed nearest neighbor rule (Corresp.) [J]. IEEE Transactions on Information Theory, 1968, 14(3): 515-516.

[8] Kubat M, Matwin S. Addressing the curse of imbalanced train-

ing sets: one sided selection[C]//Proceedings of the 14th International Conference on Machine Learning. Nashville: Morgan Kaufmann, 1997: 179-186.

[9] Laurikkala J. Improving identification of difficult small classes by balancing class distribution[C]//Artificial Intelligence in Medicine. Berlin, Heidelberg: Springer, 2001: 63-66.

[10] 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.

[11] 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: 321-357.

[12] Han H, Wang W Y, Mao B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[M]//Advances in Intelligent Computing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005: 878-887.

[13] Bunkhumpornpat C, Sinapiromsaran K, Lursinsap C. Safe-level-SMOTE: safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem[M]//Advances in Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009: 475-482.

[14] Engelmann J, Lessmann S. Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning[J]. Expert Systems with Applications, 2021, 174: 114582.

[15] Susan S, Kumar A. SSO maj-SMOTE-SSO Min: three-step intelligent pruning of majority and minority samples for learning from imbalanced datasets[J]. Applied Soft Computing, 2019, 78: 141-149.

[16] Chen H M, Li T R, Fan X, et al. Feature selection for imbalanced data based on neighborhood rough sets[J]. Information Sciences, 2019, 483: 1-20.

[17] Moreno-Torres J G, Herrera F. A preliminary study on overlapping and data fracture in imbalanced domains by means of Genetic Programming-based feature extraction[C]//Proceedings of the 2010 10th International Conference on Intelligent Systems Design and Applications. Piscataway: IEEE Press, 2010: 501-506.

[18] Salekshahrezaee Z, Leevy J L, Khoshgoftaar T M. Feature extraction for class imbalance using a convolutional autoencoder and data sampling[C]//Proceedings of the 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). Piscataway: IEEE Press, 2021: 217-223.

[19] Ng W W Y, Zeng G J, Zhang J J, et al. Dual autoencoders features for imbalance classification problem[J]. Pattern Recognition, 2016, 60: 875-889.

[20] 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.

[21] Cao P, Zhao D Z, Zaiane O. An optimized cost-sensitive SVM for imbalanced data learning[M]//Advances in Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013: 280-292.

[22] Arar ö F, Ayan K. Software defect prediction using cost-sensitive neural network[J]. Applied Soft Computing, 2015, 33: 263-277.

[23] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
Zhou Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016.

[24] Bauer E, Kohavi R. An empirical comparison of voting classification algorithms: bagging, boosting, and variants[J]. Machine Learning, 1999, 36(1): 105-139.

[25] Hastie T, Rosset S, Zhu J, et al. Multi-class AdaBoost[J]. Statistics and Its Interface, 2009, 2(3): 349-360.

[26] Friedman J H. Stochastic gradient boosting[J]. Computational Statistics & Data Analysis, 2002, 38(4): 367-378.

[27] Datsi T, Aznag K, Benali B A, et al. A short survey on multimodal data fusion in image classification[C]//Proceedings of the 2024 International Conference on Global Aeronautical Engineering and Satellite Technology (GAST). Piscataway: IEEE Press, 2024: 1-4.

[28] Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: improving prediction of the minority class in boosting[M]//Knowledge Discovery in Databases: PKDD 2003. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003: 107-119.

[29] Liu X Y, Wu J X, Zhou Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics Part B, Cybernetics, 2009, 39(2): 539-550.

[30] Shafer G. A mathematical theory of evidence[M]. Princeton: Princeton University Press, 1976.

[31] Smets P. Data fusion in the transferable belief model[C]//Proceedings of the Third International Conference on Information Fusion. Piscataway: IEEE Press, 2000: PS21-PS33.

[32] Smarandache F, Dezert J. On the consistency of PCR6 with the averaging rule and its application to probability estimation[C]//Proceedings of the 16th International Conference on Information Fusion. Piscataway: IEEE Press, 2013: 1119-1126.

[33] Tacnet J M, Dezert J. Cautious OWA and evidential reasoning

for decision making under uncertainty[C]//Proceedings of the 14th International Conference on Information Fusion. Piscataway: IEEE Press, 2011: 1-8.
[34] Fisher R A, Forina M, Asuncion A, et al. UCI machine learning repository[EB]. 1987.
[35] Hettich S, Bay S D. KDD cup 1999[EB]. 2007.

[作者简介]

和红顺（1990– ），男，博士，现就职于中国移动通信集团有限公司研究院，主要研究方向为机器学习、业务识别、计算机视觉、6G创新业务相关技术。

胡国良（1991– ），男，博士，西北农林科技大学信息工程学院讲师，主要研究方向为计算机体系结构、计算机双目视觉、图像处理、业务识别。

张志鹏（1972– ），男，博士，现就职于中国移动通信有限公司研究院，主要研究方向为机器学习、计算机视觉、AI智慧工业关键技术与产品创新。

柴鑫刚（1976– ），男，中国移动通信有限公司研究院高级工程师，主要研究方向为视联网、计算机视觉、6G业务识别及创新业务关键技术与产品创新。

高静（1976– ），女，现就职于中国移动通信有限公司研究院，主要研究方向为大视频AI技术、6G沉浸媒体业务关键技术与产品创新。