基于异同矩阵的不平衡数据处理算法

易小云,江丹阳,刘相源,郭志军

(湖南师范大学 信息科学与工程学院,湖南 长沙 410081)

摘 要: 不平衡数据分类是机器学习领域的重要挑战。传统方法独立进行特征选择与过采样,一方面扰动特征重要性评估并引入采样噪声;另一方面仅考虑异类样本区分度,忽略同类样本相似度,难以准确识别判别性特征。针对该问题,提出基于异同矩阵的不平衡数据统一处理算法ISDM,通过引入异同矩阵(SDM),利用其同时捕获异类区分度与同类相似度的特性,从多角度量化样本与特征的判别信息,为特征选择和样本质量评估提供统一度量基础。该算法首先基于SDM剔除冗余特征,构建低维特征空间;其次,利用SDM计算的样本重要度对少数类进行聚类与加权插值,确保特征选择与过采样在一致信息度量下协同工作,避免了信息割裂导致的性能损失。在12个基准数据集上的实验表明,ISDM在CART和NN分类器下的F-measure分别为85.69%和84.43%,AUC分别为91.67%和92.23%,相较对比算法平均提升2.22%~9.88%,验证了其对不平衡数据集的分类性能。

关键词: 不平衡数据;异同矩阵;特征选择;过采样;粗糙集理论

DOI: 10. 11907/rjdk. 261065 扫描二维码阅读全文:

中图分类号: TP181 **文献标识码:** A **文章编号:** 1672-7800(XXXX)0XX-0001-08

An Imbalanced Data Processing Algorithm Based on Similarity-Discernibility Matrix

YI Xiaoyun, JIANG Danyang, LIU Xiangyuan, GUO Zhijun

(*School of Information Science and Engineering*, *Hunan Normal University*, *Changsha* 410081, *China*)

Abstract: Imbalanced data classification remains a significant challenge in machine learning. Conventional methods treat feature selection and oversampling as independent processes. On the one hand, this independence distorts feature importance evaluation and introduces sampling noise; on the other hand, feature assessment typically considers only the discernibility between samples from different classes while neglecting the similarity within the same class, making it difficult to accurately identify discriminative features. To address this issue, this paper proposes a unified processing algorithm for imbalanced data based on the similarity-discernibility matrix (ISDM). The algorithm introduces the Similarity-Discernibility Matrix (SDM), which simultaneously captures the discernibility between different classes and the similarity within the same class, enabling multi-perspective quantification of discriminative information for both samples and features, thereby providing a unified measurement basis for feature selection and sample quality evaluation. The algorithm first employs SDM to remove redundant features and construct a low-dimensional feature space, then performs clustering and weighted interpolation on minority class samples based on SDM-computed sample importance, ensuring that feature selection and oversampling collaborate under a consistent information measure, avoiding performance loss caused by information fragmentation. Experimental results on 12 benchmark datasets demonstrate that ISDM achieves F-measure of 85.69% and 84.43% under CART and NN classifiers respectively, and AUC of 91.67% and 92.23% respectively, with average improvements of 2.22% to 9.88% over comparative algorithms, validating the effectiveness of the algorithm.

Key Words: imbalanced data; similarity-discernibility matrix; feature selection; oversampling; rough set theory

---

收稿日期: 2026-02-02

基金项目: 国家自然科学基金项目(11201490);湖南省自然科学基金项目(2021JJ20037);长沙市杰出创新青年培养计划项目(kq1905031)

作者简介: 易小云(2000-),男,CCF学生会员,湖南师范大学信息科学与工程学院硕士研究生,研究方向为数据挖掘、机器学习、粗糙集理论;江丹阳(2006-),女,湖南师范大学信息科学与工程学院学生,研究方向为粒计算与智能信息处理、人工智能;刘相源(2006-),男,湖南师范大学信息科学与工程学院学生,研究方向为人工智能;郭志军(1982-),男,博士,湖南师范大学信息科学与工程学院副教授,研究方向为粒计算与智能信息处理、粗糙集、人工智能。本文通讯作者:郭志军。

0 引言

随着数据规模的持续扩张，类别不平衡问题已成为实际应用中的普遍共性现象，即多数类样本数量远高于少数类，而少数类样本通常对应着具备更高决策价值的关键事件。不平衡数据分类给机器学习带来严峻挑战，传统分类算法倾向于多数类而忽视少数类，导致对关键少数类事件的识别能力不足[1-3]。同时，实际数据往往包含大量冗余特征，进一步加剧了分类难度。针对不平衡数据问题，过采样技术与特征选择是两种主流的应对策略[4]。然而，该类方法通常将特征选择与样本平衡化独立处理，可能导致特征重要性评估扰动和采样噪声，使生成样本分布偏离原始数据结构。此外，传统特征选择方法仅从单一视角评估特征，如基于区分矩阵的方法仅关注异类样本区分性，基于邻域的方法仅关注同类样本相似性，在不平衡数据中易导致特征被多数类主导或无法保持少数类内部结构[5]。因此，亟需一种能够统一处理特征选择与过采样，并从多视角评估特征质量的方法，以提升不平衡数据的分类性能。

1 相关工作

类别不平衡问题广泛存在于医学诊断、金融欺诈检测、工业故障预测和网络异常检测等高风险应用中，少数类样本往往代表着高风险、高价值或关键任务事件[6-8]。针对类别不平衡问题，现有解决方案主要分为数据层面的方法、算法层面的方法以及混合方法三类[9-11]。其中，数据层面的方法因其通用性和有效性而受到广泛关注，主要包括重采样技术和特征选择方法。

1.1 重采样技术

重采样技术通过调整类别分布来实现平衡，其中过采样方法通过为少数类生成合成样本而被广泛应用。例如，Chawla 等[12]提出的合成少数类过采样技术（Synthetic Minority Over-sampling Technique，SMOTE）通过在少数类样本及其 k 近邻构成的特征空间内进行线性插值来生成新样本，解决了简单重复采样导致的样本冗余问题，是该领域的经典方法。近年来，基于 SMOTE 的改进方法不断涌现。例如，He 等[13]提出的 ADASYN 方法根据样本密度自适应生成样本，提高了边界样本的生成比例。Han 等[14]提出的 Borderline-SMOTE 方法关注边界区域样本，仅对边界附近的少数类样本进行过采样。Yi 等[15]提出的 MC-SMOTE 方法通过聚类和簇间插值解决样本分布不均问题，避免了类内样本密集区域的过度采样。Zhang 等[16]提出的 IW-SMOTE 方法通过基于混淆的加权机制来调整种子样本选择概率，增强了对难分样本的关注程度。以上方法在分类任务中表现良好，但主要关注采样策略改进，较少考虑特征空间质量对生成样本的影响。

1.2 粗糙集理论与特征选择方法

粗糙集理论作为处理不确定性和不完整信息的有效数学工具，为特征选择提供了坚实的理论基础[17-18]。定义知识表示系统 $S=(U,C\cup d,V,f)$，其中 U 为论域，$A=C\cup d$ 为特征集合，V 为特征值域，f: $U\times A\rightarrow V$ 为信息映射，粗糙集通过等价关系将论域划分为多个等价类，并利用上下近似集合刻画目标集合的边界特性。基于粗糙集的特征选择方法主要分为两类：第一类方法基于同类样本相似性，利用邻域依赖性、信息熵等度量进行特征评估，能够识别保持类内一致性的特征，但在处理不平衡数据时容易受多数类主导[19-20]；第二类方法基于异类样本可区分性度量特征重要性，主要包括区分矩阵和样本对方法[21-22]。例如 Skowron 等[23]提出的区分矩阵通过构建矩阵 $DM=(\psi_{ij})_{n\times n}$（其中 $\psi_{ij}=c\in C\left|f(x_i,c)\neq f(x_j,c)\right.$）来评估特征对异类样本的区分能力。后续研究在理论和应用层面不断丰富了该方法，在知识推理和数据挖掘等领域发挥着重要作用[24-25]。然而，该类方法仅关注异类样本区分性，忽略了同类样本相似性信息，可能导致选出的特征无法充分保持类内结构。为克服单一视角的局限性，李元江等[5]提出异同矩阵（Similarity-Discernibility Matrix，SDM），其综合考虑样本对的相似性和差异性，为多角度特征评估提供了理论基础。具体来说，异同矩阵可考察异类样本的可区分性，考察同类样本的相似性，当阈值 $\sigma_1=0$ 且 $\sigma_2=0$ 时，异同矩阵退化为经典区分矩阵，表明其是对区分矩阵的扩展。然而，现有研究尚未针对不平衡数据场景对异同矩阵进行优化与应用。

针对不平衡数据的特征选择问题，粗糙集理论也得到了应用与发展。例如，Chen 等[26]提出基于邻域粗糙集与区分矩阵的方法，通过邻域关系处理数值型数据并结合区分矩阵进行特征选择。Sun 等[27-28]提出结合混合采样与自适应邻域的框架，通过动态调整邻域半径来适应不平衡数据的特点。Liu 等[29]将加权熵与经典粗糙集相结合，提出新的约简规则，在特征评估中赋予少数类样本更高权重。

尽管不平衡数据的处理方法取得了较大进展，但仍存在以下局限：①特征选择和过采样通常独立执行，可能导致特征重要性评估扰动和采样噪声，使生成样本分布偏离原始数据结构；②基于区分矩阵的方法仅关注异类样本区分性，基于邻域的方法仅关注同类样本相似性，单一视角评估在不平衡数据中易导致特征被多数类主导或无法保持少数类的内部结构；③特征选择与过采样的独立处理缺乏统一的质量评估标准，难以协同优化。为解决上述问题，本研究提出基于异同矩阵的不平衡数据统一处理算法（Imbalanced Data Processing Algorithm Based on the Similarity-Discernibility Matrix，ISDM），利用 SDM 同时捕获异类样本区分度和同类样本相似度特性，为特征选择与样本质量评估提供一致的度量基础。该算法首先基于 SDM 进行特

征约简，识别对少数类具有判别力的特征；其次在约简后的特征空间中利用SDM计算的样本重要度对少数类进行聚类与加权过采样，以确保特征选择与过采样在一致信息度量下协同工作，避免信息割裂导致的性能损失，实现不平衡数据的协同优化。

2　理论基础

设知识表达系统为 $S=(U,A,V,f)$，其中 U 为论域，$A=C\cup d$ 为特征集合，V 为特征值域，$|U|=n$，$A=C\cup d$，$|C|=m$，$f(x_i,c_t)$ 为对象 x_i 在特征 c_t 上的特征值。由此构建的异同矩阵表示为：

$$SDM=(\varphi_{ij})_{n\times n} \tag{1}$$

$$\varphi_{ij}=\begin{cases}\left[D_{c_t}(x_i,x_j)\right], t=1,2,\cdots,m \quad d(x_i)\neq d(x_j)\\ \left[S_{c_t}(x_i,x_j)\right], t=1,2,\cdots,m \quad d(x_i)=d(x_j)\end{cases} \tag{2}$$

$$D_{c_t}(x_i,x_j)=\begin{cases}1, \left|f(x_i,c_t)-f(x_j,c_t)\right|>\sigma_1\\ 0, \left|f(x_i,c_t)-f(x_j,c_t)\right|\leqslant\sigma_1\end{cases} \tag{3}$$

$$S_{c_t}(x_i,x_j)=\begin{cases}1, \left|f(x_i,c_t)-f(x_j,c_t)\right|<\sigma_2\\ 0, \left|f(x_i,c_t)-f(x_j,c_t)\right|\geqslant\sigma_2\end{cases} \tag{4}$$

式中：$D_{c_t}(x_i,x_j)$ 用于刻画异类对象的特征差异；$S_{c_t}(x_i,x_j)$ 用于刻画同类对象的特征相似性；σ_1 与 σ_2 分别为差异判定阈值与相似判定阈值。

该矩阵具有对称性，计算时仅需处理上三角部分。矩阵元素为 m 维向量，各分量反映对应特征的判定结果。针对异类对象对，当特征值差距大于 σ_1 时记为1，表示该特征可区分两对象；针对同类对象对，当特征值差距小于 σ_2 时记为1，表示该特征不会错误分离两对象。

为阐明异同矩阵与经典区分矩阵的联系，以下给出一个重要的退化定理：当 $\sigma_1=0$ 且 $\sigma_2=0$ 时，异同矩阵退化为经典区分矩阵[5]。证明：令 $\sigma_1=0$，$\sigma_2=0$，对于任意 x_i、x_j、c_t，若 $d(x_i)=d(x_j)$，则 $S_{c_t}(x_i,x_j)=0$；若 $d(x_i)\neq d(x_j)$，当 $f(x_i,c_t)=f(x_j,c_t)$ 时，$D_{c_t}(x_i,x_j)=0$，当 $f(x_i,c_t)\neq f(x_j,c_t)$ 时，$D_{c_t}(x_i,x_j)=1$。此时，φ_{ij} 中值为1的位置对应的特征集合等同于经典区分矩阵的 ψ_{ij}，即异同矩阵简化为区分矩阵，证毕。该定理表明异同矩阵是对区分矩阵的扩展，在改进的同时保留了其优良特性。

设 $S=(U,A,V,f)$ 为知识表达系统，其中 $|U|=n$，$A=C\cup d$，$|C|=m$，SDM为相应的异同矩阵，则特征 c_t 的重要度定义为：

$$FI(c_t)=\sum_{i=1}^{n}\sum_{j=1}^{n}\varphi_{ij}(c_t) \tag{5}$$

式中：$\varphi_{ij}(c_t)$ 表示 φ_{ij} 的第 t 个分量。$FI(c_t)$ 量化了属性 c_t 的全局贡献。该指标越高，表明该属性在实现异类有效区分的同时越能保持同类决策一致性。其作为启发式搜索的评估准则，旨在筛选出信息量大且分类性能稳定的核心特征。

与特征重要度相对应，样本重要度用于评估每个样本在决策空间中的代表性，这对于指导高质量的样本生成至关重要。设 $S=(U,A,V,f)$ 为知识表达系统，其中 $|U|=n$，$A=C\cup d$，$|C|=m$，SDM 为相应的异同矩阵，则样本 x_i 的重要度定义为：

$$SI(x_i)=\sum_{j=1}^{n}\sum_{t=1}^{m}\varphi_{ij}(c_t) \tag{6}$$

式中：$SI(x_i)$ 表征样本的决策代表性与分布质量。在过采样中，该指标提供启发式引导：将高 SI 样本作为可靠基准，确保合成样本位于决策正域内；低 SI 样本则用于定位边界脆弱区，引导算法精准插值。该指标通过兼顾同类一致性与异类区分力，在强化类间边界的同时保持了类内分布的紧凑性。

3　ISDM算法构建

在建立异同矩阵理论框架及相关重要度指标后，图1给出ISDM算法的完整流程。该算法首先构建异同矩阵；其次基于SDM进行特征选择以降低维度；最后在约简后的特征空间中基于样本重要度进行过采样。

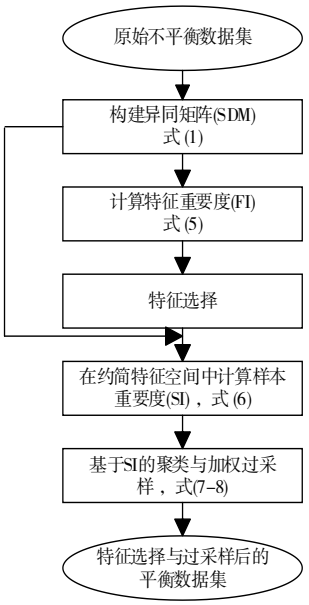

Fig. 1　Process of ISDM algorithm

图1　ISDM算法流程

依据预设阈值遍历训练集中约一半的样本对（共 $\frac{1}{2}n(n-1)$ 对），针对 m 个属性分别比较样本对，并构建异同矩阵，该步骤需执行 $\frac{1}{2}n(n-1)\times m$ 次运算，复杂度为

$O(n^2m)$，空间复杂度为 $O(n^2m)$（需存储每个属性的样本对异同信息）。计算各属性重要度（FI），贪心选取最大特征加入约简集，并将矩阵中满足 $\varphi_{ij}(c^*)=1$ 的元素置为零向量，重复该过程直至矩阵中所有 φ_{ij} 置为零向量。随后，算法将 SDM 缩减至选定的 m' 个特征子集，并进行质量受控的过采样。少数类样本被划分为 p 个簇，其中 p 设定为 $\left\lfloor \sqrt{n^+} \right\rfloor$。这些簇基于少数类样本的重要度 SI 分位数形成，SI 的计算时间复杂度为 $O(n^+m')$。为实现类间平衡，需生成的合成样本总数设为 $G=n^- - n^+$，其中 n^+ 表示少数类样本个数，n^- 表示多数类类样本个数。第 i 个簇 C_i 需生成的样本数 G_i 由其簇规模 $|C_i|$ 和簇内样本平均重要度 $\overline{SI}_i$ 共同决定。计算公式为：

$$G_i=\left\lfloor G\cdot\frac{|C_i|\cdot\overline{SI}_i}{\sum_{j=1}^{p}|C_j|\cdot\overline{SI}_j}\right\rfloor \tag{7}$$

式中：$\lfloor\cdot\rfloor$ 表示向下取整运算；G_i 表示第 i 个簇需生成的合成样本数量。

记 x_i 的 K 个近邻为 $\{x_{i,k}|k=1,2,\cdots,K\}$，从中随机选择一个邻居 $x_{i,r}$，通过线性插值生成新样本。表示为：

$$x_{new}=x_i+\delta\cdot(x_{i,r}-x_i) \tag{8}$$

式中：$\delta\in(0,1)$ 为随机数，用于保证 x_{new} 位于 x_i 与其邻居 $x_{i,r}$ 之间的连线上。该过程的时间复杂度为 $O(Gm')$。算法的总时间复杂度为 $O(n^2m+n^+m'+Gm')$，主要由 $O(n^2m)$ 决定；空间复杂度为 $O(n^2m)$。

综上所述，ISDM 算法的伪代码为：

输入：知识表达系统 $S=(U,A,V,f)$，$|U|=n$，$A=C\cup d$，$|C|=m$，阈值 σ_1 和 σ_2，$p=\left\lfloor\sqrt{n^+}\right\rfloor$

输出：基于特征选择与过采样处理后的数据集

1.Initialize $SDM=(\varphi_{ij})_{n\times n}$，$\varphi_{ij}=[0,0,\cdots,0]$

2.FOR $i=1$ *to* n，$j=i+1$ *to* n DO

3. Compute φ_{ij} using Eq.（1）

4.END FOR

5.$Red=\varnothing$，$SDM'=SDM$

6.WHILE $\exists\varphi_{ij}\in SDM'$suchthat$\varphi_{ij}\neq[0,0,\cdots,0]$DO

7. $c^*=arg\,max_{c\in C-Red}FI(c)$

8. $Red=Red\cup\{c^*\}$

9. $\forall\varphi_{ij}$ with $\varphi_{ij}(c^*)=1$：$\varphi_{ij}\leftarrow[0,0,\cdots,0]$

10.END WHILE

11.Compute $SI(x)$ for all $x\in U_{min}$；and partition the samples into p clusters via SI quantiles.

12.Compute G_i for each cluster C_i using Eq.（7）

13.FOR each cluster $i=1,2,\cdots,p$ DO

14. WHILE generated samples < G_i DO

15. Generate x_{new} by interpolating two random samples from cluster i using Eq.（8）

16. END WHILE

17.END FOR

18.RETURN Feature-selected and oversampled dataset

4 实验方法与结果分析

4.1 实验环境与参数设置

实验环境为 Matlab R2024a、Intel Core Ultra 5 225H @ 1.70 GHz 处理器和 32 GB 内存。采用五折交叉验证策略，每个数据集被分为 5 个类别比例一致的部分，其中一部分作为测试集，其余作为训练集。特征选择和过采样仅在训练集上执行，最终结果均取五折结果的平均值。评估指标包括 F-measure 和 AUC，用于评估分类器在不平衡数据上的性能。分类器选用分类回归树（Classification and Regression Tree，CART）和 KNN（K-Nearest Neighbor，$k=3$），均采用默认参数。

为全面验证 ISDM 算法的有效性，选取以下 3 类比较方法：①仅采用特征选择的方法，包括基于 KNN 粗糙集理论在不平衡数据上的特征选择方法 KNMRS（Multi-label feature selection for imbalanced data via KNN-based multi-label rough set theory）[30]；②仅采用过采样的方法，包括 MC-SMOTE（Minority Clustering SMOTE）[15] 和 IW-SMOTE（Instance Weighted SMOTE）[16]；③结合特征选择与过采样的混合方法，包括 RMDPS（Rough set based on Maximal Discernibility Pairs）[22] 和 CDA（Covering-based Discernibility matrix Attribute reduction）[31] 分别与 MC-SMOTE 和 IW-SMOTE 结合，形成 MC-DP、IW-DP、MC-CDA 和 IW-CDA 4 种混合方法。所有混合方法均遵循先特征选择后过采样的流程。

各方法参数设置如下：KNMRS 的所有参数均采用原文推荐设置；MC-SMOTE 的聚类参数 k 在 0.05~0.95 范围内以 0.05 为步长遍历；IW-SMOTE 采用原文推荐设置；RMDPS 使用默认设置；CDA 的参数 ε 在 0.05~0.6 范围内以 0.025 为步长遍历。ISDM 算法参数设置如下：阈值 σ_1 在 0~0.5 范围内以 0.02 为步长遍历，$\sigma_2=0.2\sigma_1$。阈值范围的选择基于以下考虑：①特征值经归一化处理到 $[0,1]$ 区间，σ_1 超过 0.5 意味着特征值差异需超过半个值域才被认为可区分，这样过于严格；②初步实验表明，σ_1>0.5 时特征选择结果趋于稳定，进一步增大阈值不再显著改变特征子集；步长 0.02 在计算效率与参数精度间取得平衡。σ_2 设为 $0.2\sigma_1$ 是由于同类样本的相似性判定应比异类样本的区分性判定更严格，系数 0.2 通过交叉验证确定。对于每组阈值，在训练集上进行五折交叉验证，选择平均 F-measure 最高的参数组合应用于测试集。聚类簇数 p 设为 $\lfloor\sqrt{m'}\rfloor$（其中 n^+ 为少数类样本数），该设置兼顾了计算效率。为公平比较，所有方法的可调参数（如 MC-SMOTE 的聚类参数 k、CDA 的参数 ε）均采用相同的网格搜索策略，在训练集上选取最

优参数。

4.2　数据集

选取来自 UCI（https://archive.ics.uci.edu/datasets/）、KEEL（https://sci2s.ugr.es/keel/datasets.php）和 LIBSVM（https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/）数据库的12个不平衡数据集，详细信息如表1所示。这些数据集在维度和样本规模上差异显著，特征数量范围为6~22 283，样本数量范围为62~10 000，不平衡比例范围为1.76~73.43，涵盖高维和高度不平衡的多种情况。

Table 1　Dataset information

表1　数据集信息

数据集	样本数	特征数	不平衡率
Ecoli1	336	7	3.36
Vehicle1	846	18	2.89
Vehicle2	846	18	2.88
Pima	768	8	1.86
Yeast1	1 484	8	2.45
Kddcup-guest	1 642	41	29.98
Kddcup-buffer	2 233	41	73.43
Kr-vs-k-three	2 935	6	35.23
Colon	62	2 000	1.81
Leukemia	72	7 070	1.88
Gli_85	85	22 283	2.26
Electrical	10 000	14	1.76

4.3　实验结果分析

4.3.1　分类性能比较

原始数据（RAW）和8个算法在CART和KNN（k = 3）分类器上的F-measure值和AUC分别如表2、表3、表4、表5所示，每组结果的最大值加粗表示。可以看出，经过ISDM算法处理后的数据在F-measure值和AUC两项分类指标上均优于原始数据，表明该算法在减小数据规模和平衡样本的同时能保持良好的数据质量。在CART分类器下，ISDM算法在7个数据集上的F-measure指标达到最优值（含并列第一），总体平均值为85.69%，较KNMRS、MC-ST、IW-ST、MC-DP、IW-DP、MC-CDA和IW-CDA 7种对比算法分别提升9.85%、4.90%、11.88%、5.80%、9.26%、4.94%和2.77%；ISDM算法在7个数据集上的AUC指标达到最优值（含并列第一），总体平均值为91.67%，较7种对比算法分别提升5.22%、3.51%、8.56%、3.34%、7.12%、1.71%和2.34%。

Table 2　F-measure comparison of different algorithms on CART classifier

表2　不同算法在CART分类器上的F-measure比较　　(%)

数据集	RAW	ISDM	KNMRS	MC-ST	IW-ST	MC-DP	IW-DP	MC-CDA	IW-CDA
Ecoli1	78.79	85.58	79.36	79.61	79.94	78.50	81.00	**85.90**	84.05
Vehicle1	48.13	61.09	49.56	54.27	55.25	58.84	58.46	**62.91**	62.25
Vehicle2	90.88	**95.03**	91.09	93.34	89.14	92.78	91.04	89.67	92.93
Pima	58.99	**67.58**	59.82	61.94	59.09	62.01	58.96	67.51	65.95
Yeast1	44.95	57.31	50.00	53.84	52.07	56.25	50.73	**59.48**	57.73
Kddcup-guest	97.23	**100.00**	97.13	98.26	97.23	99.05	98.18	97.23	**100.00**
Kddcup-buffer	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Kr-vs-k-three	96.15	99.53	**100.00**	99.43	96.31	98.28	96.97	97.65	98.29
Colon	64.33	**76.47**	65.28	73.58	43.87	62.23	59.29	76.32	65.58
Leukemia	84.67	94.55	90.51	**96.36**	84.14	91.92	94.14	91.96	90.55
Gli_85	73.52	**91.21**	74.05	75.70	73.29	72.07	78.78	80.89	83.24
Electrical	**99.97**	**99.97**	**99.97**	99.96	99.71	**99.97**	99.66	80.79	**99.97**
平均值	78.13	**85.69**	79.73	82.19	77.50	80.99	80.60	82.53	83.38

注：MC-ST与IW-ST分别为MC-SMOTE和IW-SMOTE的缩写，下同。

Table 3　AUC comparison of different algorithms on CART classifier

表3　不同算法在CART分类器上的AUC比较　　(%)

数据集	RAW	ISDM	KNMRS	MC-ST	IW-ST	MC-DP	IW-DP	MC-CDA	IW-CDA
Ecoli1	92.70	94.83	90.49	87.51	86.34	90.09	86.08	95.99	92.54
Vehicle1	74.19	**79.27**	75.06	75.73	74.17	75.08	74.40	78.89	78.29
Vehicle2	95.14	**97.68**	95.30	95.89	94.78	96.91	95.32	96.06	97.66
Pima	74.35	77.35	71.02	73.54	77.13	73.76	75.96	**78.11**	76.06
Yeast1	69.37	74.57	70.66	70.89	67.30	73.14	73.89	**75.10**	72.26
Kddcup-guest	99.03	**100.00**	98.06	99.94	99.03	99.97	99.94	99.03	**100.00**
Kddcup-buffer	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Kr-vs-k-three	98.24	99.98	**100.00**	**100.00**	99.85	**100.00**	98.91	99.96	99.95
Colon	72.50	84.13	81.75	83.75	58.63	77.88	70.19	91.25	75.19
Leukemia	88.78	**97.98**	92.89	97.89	88.78	97.78	95.78	91.89	93.78
Gli_85	79.64	**95.42**	84.20	79.61	79.50	81.89	81.65	86.77	88.97
Electrical	99.97	**99.98**	**99.98**	99.96	99.76	**99.98**	99.80	87.41	**99.98**
平均值	86.99	**91.77**	88.28	88.73	85.44	88.87	87.66	90.04	89.56

在KNN（k = 3）分类器下，ISDM算法同样表现优异，F-measure指标在7个数据集上达到最优值，总体平均值为84.43%，较7种对比算法分别提升7.58%、8.94%、10.16%、6.46%、9.21%、4.16%和2.22%；AUC指标在6个数据集上达到最优值，总体平均值为92.23%，较7种对比算法分别提升1.69%、1.84%、2.50%、2.36%、3.13%、1.65%和1.56%。

Table 4 F-measure comparison of different algorithms on KNN（k = 3）classifier

表4 不同算法在KNN（k = 3）分类器上的F-measure比较 **(%)**

数据集	RAW	ISDM	KNMRS	MC-ST	IW-ST	MC-DP	IW-DP	MC-CDA	IW-CDA
Ecoli1	77.11	83.88	75.94	81.43	77.24	82.43	78.28	**89.54**	82.33
Vehicle1	52.33	**62.15**	53.03	59.83	57.34	60.24	58.36	62.04	59.94
Vehicle2	89.75	90.46	**94.09**	88.45	87.33	89.77	87.02	83.52	89.33
Pima	58.53	**69.07**	60.13	65.91	65.71	63.16	62.16	67.48	66.51
Yeast1	48.78	59.85	49.09	57.34	56.61	58.26	57.05	**61.61**	59.95
Kddcup-guest	99.05	**100.00**	95.31	**100.00**	**100.00**	98.18	99.05	97.14	**100.00**
Kddcup-buffer	98.18	**100.00**	**100.00**	**100.00**	98.18	**100.00**	98.18	**100.00**	**100.00**
Kr-vs-k-three	90.79	**97.65**	97.53	86.70	91.95	84.68	91.21	89.52	97.11
Colon	59.86	**78.32**	71.23	68.25	54.22	63.05	61.04	78.06	73.66
Leukemia	81.10	94.55	85.57	64.83	75.85	**94.85**	71.11	90.18	92.36
Gli_85	71.67	**89.29**	65.15	70.38	69.91	70.12	78.78	73.94	78.72
Electrical	87.38	87.89	**97.24**	86.89	85.39	86.97	85.43	79.63	91.25
平均值	76.21	**84.43**	78.48	77.50	76.64	79.31	77.31	81.06	82.60

Table 5 AUC comparison of different algorithms on KNN（k=3）classifier

表5 不同算法在KNN（k=3）分类器上的AUC比较 **(%)**

数据集	RAW	ISDM	KNMRS	MC-ST	IW-ST	MC-DP	IW-DP	MC-CDA	IW-CDA
Ecoli1	94.27	95.23	93.89	93.98	92.35	93.97	92.18	**95.42**	92.58
Vehicle1	80.77	80.28	79.14	**81.89**	80.81	81.90	80.25	81.59	80.20
Vehicle2	98.88	**99.11**	98.99	98.81	98.00	98.99	97.99	96.14	98.35
Pima	76.72	**80.40**	78.29	78.20	77.65	77.78	76.05	79.19	78.46
Yeast1	75.08	76.07	75.29	76.73	74.13	77.07	75.62	**77.50**	76.13
Kddcup-guest	**100.00**	**100.00**	98.02	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Kddcup-buffer	**100.00**	**100.00**	**100.00**	**100.00**	98.33	**100.00**	**100.00**	**100.00**	**100.00**
Kr-vs-k-three	99.31	99.98	**100.00**	99.90	99.32	99.27	99.30	99.91	99.97
Colon	76.44	85.94	**88.44**	81.69	84.69	71.19	79.94	87.13	77.75
Leukemia	93.38	**97.78**	96.00	90.11	90.09	98.67	86.71	94.93	95.47
Gli_85	92.05	**93.97**	78.71	89.45	89.82	86.23	90.39	85.42	92.98
Electrical	96.67	97.62	**99.71**	95.94	94.62	96.14	94.77	91.50	97.88
平均值	90.30	**92.23**	90.58	90.56	89.98	90.10	89.43	90.73	90.81

为进一步评估各算法的整体性能，采用Friedman排序方法计算各算法在不同评价指标下的平均排名[32]。图2为所有算法在CART分类器和KNN（k = 3）分类器上的F-measure和AUC评价指标平均排名情况。可以看出，ISDM算法在4种评估场景中的平均排名介于1.708~2.958之间，在所有场景中均获得最优排名。相比之下，其他方法的排名波动较大，KNMRS的平均排名在4.375~6.083之间，IW-SMOTE在5.167~6.292之间，MC-SMOTE在3.625~4.542之间，基于DP的混合方法（MC-DP、IW-DP）在3.375~5.625之间，而基于CDA的混合方法（MC-CDA、IW-CDA）在2.792~3.750之间。

4.3.2 计算效率分析

表6为不同算法在数据预处理阶段的运行时间比较，该阶段包括特征选择和过采样两个环节。可以看出，ISDM算法与结合特征选择和过采样的混合方法相比展现出更优的计算效率，而与仅采用过采样的算法（如MC-SMOTE和IW-SMOTE）则保持相当的运行时间。在特征选择阶段，ISDM算法在12个数据集上的平均运行时间为18.9 s，分别比混合方法MC-DP、IW-DP和MC-CDA快39.3倍、41.1倍和1.6倍。在过采样阶段，ISDM在12个数据集上的平均运行时间为3.3 s，分别比KNMRS、IW-SMOTE、IW-DP、MC-CDA和IW-CDA快180.9倍、15.6倍、10.9倍、5.6倍和193.9倍。在总预处理时间方面，ISDM在12个数据集上的平均运行时间为22.1 s，相较KNMRS、IW-ST、MC-DP、IW-DP和IW-CDA 5种基准方法分别实现了2.3倍、33.8倍、36.9倍、2.1倍和29.2倍的加速。

4.3.3 特征选择效果分析

表7为不同算法在12个数据集上的特征子集个数（5折交叉验证平均值）比较。可以看出，ISDM算法在特征选择方面表现出色，在大多数数据集上选择的特征数量最少或接近最少。例如，在Kddcup-guess和Kddcup-buffer数据集上，ISDM分别仅选择1.8和1.0个特征，显著少于41个原始特征；在Colon、Leukemia和Gli_85等高维数据集上，ISDM也实现了大幅降维，分别选择7.0、2.0和7.4个特征。

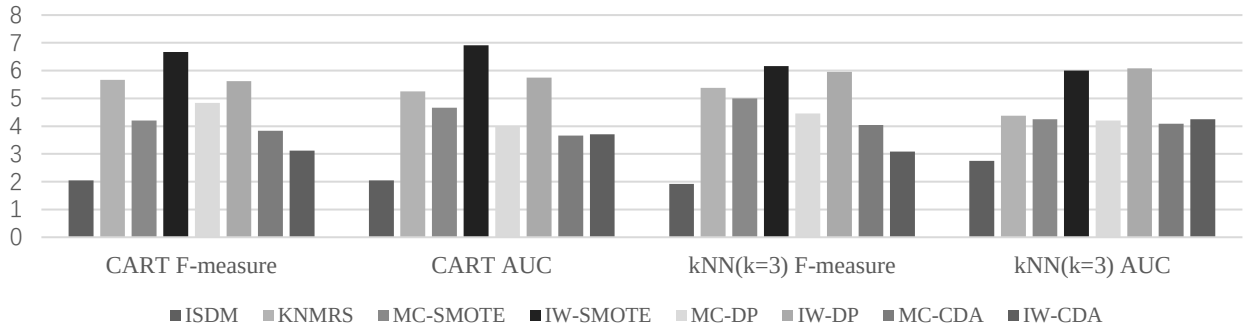

Fig. 2　Average ranking of F-measure and AUC evaluation metrics for various algorithms

图2　各算法F-measure和AUC评价指标平均排名

Table 6　Comparison of running times of different algorithms

表6　不同算法运行时间比较　　**(s)**

数据集	ISDM		KNMRS	MC-ST	IW-ST	MC-DP		IW-DP		MC-CDA		IW-CDA	
	FS	OS	FS	OS	OS	FS	OS	FS	OS	FS	OS	FS	OS
Ecoli1	0.09	0.89	75.50	0.37	8.84	1.42	***0.23***	2.37	7.88	**0.05**	3.38	0.05	219.35
Vehicle1	1.17	1.33	304.90	***0.26***	7.79	37.59	0.31	35.06	6.40	**0.45**	6.83	0.45	274.42
Vehicle2	1.12	1.24	291.92	***0.23***	9.26	35.91	0.24	32.89	6.48	**0.40**	6.67	0.41	260.25
Pima	0.55	1.72	214.51	***0.39***	7.52	35.11	***0.39***	69.40	4.41	0.41	21.33	**0.28**	151.22
Yeast1	1.58	2.71	735.25	0.58	9.77	39.61	***0.56***	36.25	7.95	**0.70**	13.20	0.73	211.29
Kddcup-guest	2.45	1.91	1 840.60	0.32	75.18	72.22	***0.13***	63.14	62.98	**0.15**	3.25	0.18	1 027.00
Kddcup-buffer	3.94	1.64	3 334.60	0.12	295.09	121.50	***0.11***	108.00	126.68	**0.12**	2.03	0.16	3 175.90
Kr-vs-k-three	6.59	0.70	2 354.40	0.25	100.87	112.30	***0.24***	101.75	76.46	**0.22**	5.20	0.26	1 827.50
Colon	0.41	0.48	97.81	***0.13***	14.24	12.04	1.37	9.02	5.49	1.75	1.51	**0.22**	65.16
Leukemia	0.98	0.41	341.72	0.35	21.81	126.41	***0.14***	119.24	29.76	**0.50**	1.26	0.60	63.04
Gli_85	5.77	0.46	1421.20	1.64	55.15	291.48	***0.06***	592.38	73.75	**2.92**	1.41	3.30	89.31
Electrical	202.12	25.79	36 733.70	***5.96***	8.34	8 019.23	8.81	8 141.20	18.99	341.60	152.62	**82.87**	250.08
平均	18.90	3.27	3 978.80	***0.37***	51.15	742.07	1.05	775.89	35.60	29.11	18.22	**7.46**	634.55

注：①FS与OS分别表示特征选择与过采样阶段耗时；②每行中的FS最小值加粗显示，OS最小值以加粗斜体显示。

KNMRS作为纯特征选择方法同样表现良好，但在部分数据集（如Gli_85）上选择的特征数（18.2）多于ISDM。在混合方法中，MC-DP和IW-DP在多数数据集上选择的特征数量较少，表现优于MC-CDA和IW-CDA。其中，仅基于过采样的MC-ST和IW-ST算法保留了全部原始特征，在高维数据集上可能面临维度灾难问题。综上所述，ISDM通过有效的特征选择策略，在保持分类性能的同时显著降低了特征维度，提高了模型的简洁性和可解释性。

Table 7　Comparison of the number of feature subsets across different algorithms

表7　不同算法特征子集个数比较

数据集	RAW	ISDM	KNMRS	MC-ST	IW-ST	MC-DP	IW-DP	MC-CDA	IW-CDA
Ecoli1	7.0	6.6	5.6	7.0	7.0	6.0	6.8	6.0	5.7
Vehicle1	18.0	18.0	11.8	18.0	18.0	17.2	18.0	17.4	14.2
Vehicle2	18.0	17.8	11.0	18.0	18.0	17.0	18.0	17.0	14.1
Pima	8.0	8.0	5.8	8.0	8.0	8.0	8.0	8.0	7.9
Yeast1	8.0	8.0	7.8	8.0	8.0	8.0	8.0	8.0	8.0
Kddcup-guest	41.0	1.8	8.4	41.0	41.0	1.6	28.6	1.4	2.8
Kddcup-buffer	41.0	1.0	3.4	41.0	41.0	1.0	27.0	1.0	1.0
Kr-vs-k-three	6.0	4.0	4.4	6.0	6.0	6.0	6.0	6.0	4.7
Colon	2 000.0	7.0	6.4	2 000.0	2 000.0	3.2	3.4	3.4	7.3
Leukemia	7 070.0	2.0	7.0	7 070.0	7 070.0	2.0	2.0	2.0	4.1
Gli_85	22 283.0	7.4	18.2	22 283.0	22 283.0	4.6	4.6	4.4	6.9
Electrical	14.0	13.0	3.6	14.0	14.0	1.0	13.0	13.0	13.0

5　结语

本文提出的基于异同矩阵的不平衡数据统一处理算法ISDM通过构建统一处理框架，打破了传统方法特征选择与过采样独立执行的局限。该算法基于双重视角评估机制，同时考虑异类区分度和同类相似度，克服了单一视角易受多数类主导的缺陷，实现了特征空间与样本空间在

一致信息度量下的协同优化。实验结果表明,ISDM算法在12个数据集上的F-measure和AUC指标均显著优于对照方法。然而,ISDM在处理大规模数据时仍存在计算开销较大的问题,且阈值参数依赖人工设定,在增量学习和多标签不平衡场景中的适用性尚需进一步验证。未来工作将重点研究基于采样或并行计算的加速策略以适应大规模数据、阈值的自适应设置方法,以及在增量学习和多标签不平衡问题中拓展该算法的应用。

参考文献:

[1] CARVALHO M, PINHO A J, BRAS S. Resampling approaches to handle class imbalance: a review from a data perspective [J]. Journal of Big Data, 2025, 12(1):71.

[2] ZHANG C M. A classification method for imbalanced data based on fuzzy rough nearest neighbor[J]. Software Guide, 2020, 19(11): 37-41.
章春梅. 基于模糊粗糙最近邻算法的不平衡数据分类[J]. 软件导刊, 2020,19(11):37-41.

[3] ZOU C A, WANG J B, FU G H. An imbalanced classification algorithm combining MetaCost and resampling—RS-MetaCost[J]. Software Guide, 2022, 21(3): 34-41.
邹春安,王嘉宝,付光辉. MetaCost与重采样结合的不平衡分类算法——RS-MetaCost[J]. 软件导刊,2022,21(3):34-41.

[4] CHEN W, YANG K, YU Z, et al. A survey on imbalanced learning: latest research, applications and future directions [J]. Artificial Intelligence Review, 2024, 57:137.

[5] LI Y J, QUAN J S, TAN Y Y, et al. Attribute reduction for high-dimensional data from dual perspectives of similarity and difference [J]. Journal of Computer Applications, 2023, 43(5):1467-1472.
李元江,权金升,谭阳奕,等. 基于相似和差异双视角的高维数据属性约简[J]. 计算机应用,2023,43(5):1467-1472.

[6] CHEN Y, CALABRESE R, MARTIN B, et al. Interpretable machine learning for imbalanced credit scoring datasets [J]. European Journal of Operational Research, 2024, 312(1):357-372.

[7] HUANG H, LIU B, XUE X, et al. Imbalanced credit card fraud detection data: a solution based on hybrid neural network and clustering based under sampling technique [J]. Applied Soft Computing, 2024, 154: 111368.

[8] REN Z, LIN T, FENG K, et al. A systematic review on imbalanced learning methods in intelligent fault diagnosis [J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72:1-35.

[9] ELREEDY D, ATIYA A F, KAMALOV F. A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning [J]. Machine Learning, 2024, 113(7):4903-4923.

[10] ZHAO H, ZHAO C, ZHANG X, et al. An ensemble learning approach with gradient resampling for class imbalance problems [J]. INFORMS Journal on Computing, 2023, 35(4):747-763.

[11] NIZAM O H, ORMAN Z. A heuristic based hybrid sampling method using a combination of SMOTE and ENN for imbalanced health data [J]. Expert Systems, 2024, 41(8):e13596.

[12] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority oversampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16:321-357.

[13] HE H B, BAI Y, GARCIA E A, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning [C]//Proceedings of 2008 IEEE International Joint Conference on Neural Networks, 2008: 1322-1328.

[14] HAN H, WANG W, MAO B. Borderline SMOTE: a new oversampling method in imbalanced data sets learning[C]//Proceedings of International Conference on Intelligent Computing, 2005:878-887.

[15] YI H, JIANG Q, YAN X, et al. Imbalanced classification based on minority clustering synthetic minority oversampling technique with wind turbine fault detection application [J]. IEEE Transactions on Industrial Informatics, 2021, 17(9):5867-5875.

[16] ZHANG A M, YU H L, ZHOU S L, et al. Instance weighted SMOTE by indirectly exploring the data distribution[J]. Knowledge Based Systems, 2022, 249:108919.

[17] PAWLAK Z. Rough sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5):341-356.

[18] WANG G Y, YAO Y Y, YU H. A survey on rough set theory and applications[J]. Chinese Journal of Computers, 2009, 32(7):1229-1246.
王国胤,姚一豫,于洪. 粗糙集理论与应用研究综述[J]. 计算机学报, 2009,32(7):1229-1246.

[19] DENG Z X, LI T R, ZHANG P F, et al. Feature selection for label distribution learning based on neighborhood fuzzy rough sets [J]. Applied Soft Computing, 2025, 169:112542.

[20] SANG B B, YANG L, XU W H, et al. VCOS: multi scale information fusion to feature selection using fuzzy rough combination entropy[J]. Information Fusion, 2025, 117:102901.

[21] KOMOROWSKI J, PAWLAK Z, POLKOWSKI L, et al. Rough sets: a tutorial[M]. Singapore: Springer, 1999:3-98.

[22] DAI J H, HU H, WU W Z, et al. Maximal discernibility pair based approach to attribute reduction in fuzzy rough sets[J]. IEEE Transactions on Fuzzy Systems, 2017, 26(4):2174-2187.

[23] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems[M]. Berlin: Springer, 1992:331-362.

[24] JENSEN R, SHEN Q. New approaches to fuzzy rough feature selection [J]. IEEE Transactions on Fuzzy Systems, 2008, 17(4):824-838.

[25] SOWKUNTLA P, PRASAD P S V S S. Parallel attribute reduction in high dimensional data: an efficient MapReduce strategy with fuzzy discernibility matrix [J]. Applied Soft Computing, 2025, 172:112870.

[26] CHEN H, LI T, FAN X, et al. Feature selection for imbalanced data based on neighborhood rough sets[J]. Information Sciences, 2019, 483: 1-20.

[27] SUN L, LI M, DING W, et al. Adaptive fuzzy multi neighborhood feature selection with hybrid sampling and its application for class imbalance data[J]. Applied Soft Computing, 2023, 149:110968.

[28] SUN L, SI S, DING W, et al. TFSFB: two stage feature selection via fusing fuzzy multi neighborhood rough set with binary whale optimization for imbalanced data [J]. Information Fusion, 2023, 95:91-108.

[29] LIU J, HU Q, YU D. A weighted rough set based method developed for class imbalance learning [J]. Information Sciences, 2008, 178(4): 1235-1256.

[30] XU W H, LI Y Z. Multi-label feature selection for imbalanced data via KNN-based multi-label rough set theory[J]. Information Sciences, 2025, 715:122220.

[31] WANG C Z, HE Q, CHEN D G, et al. A novel method for attribute reduction of covering decision systems [J]. Information Sciences, 2014, 254:181-196.

[32] DEMŠAR J. Statistical comparisons of classifiers over multiple data sets [J]. Journal of Machine Learning Research, 2006, 7:1-30.

(责任编辑:尹晨茹)