基于 K-means 的不平衡数据过采样研究方法

杨云熙 [1]
YANG Yunxi

摘 要 随着机器学习在各领域的广泛应用，不平衡数据问题成为分类任务中的一个关键挑战。传统分类算法在处理类别不平衡时，往往表现出对多数类样本的偏向，导致对少数类样本的分类效果较差。为了解决这一问题，文章提出了一种改进的过采样方法——Pkmeans-SMOTE（Pkmeans synthetic minority over-sampling technique），通过结合 Tomek Links、K-means 聚类和自适应权重分配策略，有效优化了少数类样本的生成过程。实验结果表明，Pkmeans-SMOTE 在多个不平衡数据集上表现出色，特别是在 AUC、F_1-score 和 G-mean 等评估指标上，相较于传统的 SMOTE、Borderline-SMOTE2 和 kmeans-SMOTE 方法，具有更强的性能优势。

关键词 SMOTE；K-means；不平衡数据；聚类方法；过采样

doi：10.3969/j.issn.1672-9528.2025.03.034

0 引言

随着机器学习技术的迅速发展，分类算法在数据分析和知识发现中发挥着重要作用。然而，数据不平衡问题常影响分类模型的表现，特别是少数类样本的识别能力。数据不平衡指某一类别样本数量远超其他类别，导致传统分类算法难以有效区分少数类和多数类，从而使模型偏向多数类，忽视少数类，影响分类效果 [1]。在医疗诊断、金融欺诈检测等领域，少数类样本的正确分类尤为重要，因此，解决不平衡分类问题成为机器学习领域的研究重点。

为应对这一问题，研究者们提出了多种方法，主要分为预处理方法和基于算法的方法 [2]。过采样技术作为一种重要的预处理方法，通过增加少数类样本来平衡不同类别间的样本数量，从而改善模型对少数类的分类性能。SMOTE（synthetic minority over-sampling technique）[3] 及其改进版本（如 Borderline-SMOTE[4] 和 kmeans-SMOTE[5] 等）通过生成新的少数类样本或优化采样策略，提升分类器的识别能力。尽管过采样技术在处理不平衡数据问题时展现出显著优势，但也存在一定的缺陷。

SMOTE 等方法可能引入噪声，生成不真实的少数类样本，进而导致模型过拟合，影响其泛化能力 [6]。同时，当数据集极度不平衡时，过采样可能生成大量冗余样本，增加计算开销，甚至降低训练效率 [7]。因此，如何在保留过采样技术优势的同时，减少其缺点，提升鲁棒性和泛化能力，成为当前研究的重要方向。改进的过采样方法不断涌现，优化样本生成策略、控制噪声等方式已被广泛应用，进一步提高了过采样技术的性能 [8-11]。

1 理论知识

1.1 Tomek Links

Tomek Links 是一种基于集成思想的欠采样方法，主要用于解决数据不平衡问题 [12]。其核心思想是通过识别和去除多数类样本中与少数类样本相近的“边界”样本，从而减少对多数类样本的过拟合，并使分类器的决策边界更加清晰。具体而言，Tomek Links 方法通过计算样本对之间的距离，选择那些既属于多数类，又与少数类样本距离较近的样本对，这些样本对即为 Tomek Links。通过去除这些样本，减少了多数类样本的数量，从而在不增加少数类样本的情况下有效改善了分类器对少数类的识别能力。此方法不仅有助于提高分类性能，还能降低过拟合风险，特别是在处理类别不平衡数据时尤为有效。

1.2 K-means 聚类方法

K-means 是一种常用的聚类算法，通过将数据划分为 K 个簇，使同一簇内样本相似度最大、不同簇间相似度最小。算法以随机选取的 K 个质心为起点，迭代分配样本到最近的质心所在簇，并更新质心位置，直到质心稳定或达到最大迭代次数 [13]。K-means 计算简单、收敛快，适用于大规模数据，广泛应用于图像分割、特征提取和数据预处理等领域 [14]。其不足在于对初始质心敏感且对噪声不鲁棒，实际应用中常结合其他方法优化结果，如用于不平衡数据时提升合成样本分布合理性和分类性能。

1. 贵州财经大学信息学院 贵州贵阳 550025

2 Pkmeans-SMOTE 模型

Pkmeans-SMOTE 是一种改进的过采样方法，通过结合 Tomek Links、K-means 聚类和权重分配策略生成高质量的合成样本，应对不平衡数据问题。该方法利用 Tomek Links 技术划分样本为边界样本和安全样本，并通过自适应密度阈值识别边界区域，有效减少噪声干扰。随后，在每个区域内部应用 K-means 聚类，对样本分布进行细分，提取簇的局部特征和质心，精确刻画样本结构。结合样本分布特性设置权重，其中边界样本赋予更高权重，以强化复杂边界区域的生成优先级，同时引入指数衰减策略优化权重分布。最终，通过插值方法在各簇中生成新样本，确保样本分布的合理性，从而显著提升分类器对不平衡数据的性能。边界样本和安全样本的权重分别计算为：

$$w_i = \frac{n_i}{N_{\text{boundary}}} + \exp(-\theta \bullet 2) \tag{1}$$

$$w_i = \frac{n_i}{N_{\text{safe}}} + \exp(-\theta \bullet 3) \tag{2}$$

式中：n_i 表示第 i 个簇的样本数量；N_{boundary} 和 N_{safe} 分别表示边界簇和安全簇中样本总数；θ 表示一个调节指数衰减的参数。权重经过归一化处理后，用于控制不同簇中生成样本的比例。最终通过插值方法在各簇中生成新样本，确保样本分布的合理性，从而显著提升分类器对不平衡数据的性能。

3 实验分析

3.1 数据集

为了评估所提出模型的性能，本文选取 7 个来自不同领域的典型不平衡数据集进行实验。这些数据集涵盖多个应用场景，具有不同的样本数量、特征维度和类别分布特性，能够有效地测试模型在各类实际应用中的表现。表 1 展示了每个数据集的基本信息，包括样本总数、特征数、少数类样本数量，以及不平衡度（IR）。其中，IR 是衡量数据集类别不平衡程度的一个重要指标，定义为多数类样本数量与少数类样本数量的比值。较高的 IR 值意味着数据集的不平衡性越严重。公式为：

$$\text{IR} = \frac{\text{多数类样本}}{\text{少数类样本}} \tag{3}$$

表 1 不平衡数据集

数据集	数量	特征	少数类	IR
haberman	306	3	81	2.78
moon	1 100	2	100	10.00
segment2	2 145	16	165	12.00
seismic	2 584	18	170	14.20
wilt	4 819	5	257	17.75

3.2 实验设计

为验证 Pkmeans-SMOTE 方法的有效性，本文将其与 SMOTE、Borderline-SMOTE2 和 kmeans-SMOTE 进行对比，分别在多个不平衡数据集上合成少数类样本，并使用决策树（DE）和随机森林（RF）进行分类。采用 AUC、F_1-score 和 G-mean 作为性能评估指标 [15]，并使用 5 折分层交叉验证来获得更稳定的评估结果。

3.3 实验结果与分析

表 2 在 5 个不平衡数据集上，比较了 4 种过采样方法（SMOTE、BSMOTE2、KSMOTE 和 Pkmeans-SMOTE） 在决策树（DE）和随机森林（RF）分类器下的表现。

表 2 四种过采样方法分类性能对比

数据集	指标	DE			
		SMOTE	BSM2	KSMOTE	Pkmeans-SMOTE
haberman	AUC	0.552 5	0.566 4	0.535 2	**0.592 1**
	F_1-score	0.729 5	0.743 9	0.734 8	**0.761 3**
	G-mean	0.521 5	0.535 8	0.495 7	**0.578 3**
moon	AUC	0.927 5	0.892 5	0.916 0	**0.957 0**
	F_1-score	0.794 0	0.782 6	0.833 6	**0.856 3**
	G-mean	0.925 8	0.886 3	0.910 9	**0.956 6**
segment2	AUC	0.903 5	0.895 5	0.872 2	**0.923 0**
	F_1-score	0.787 6	0.763 1	0.789 8	**0.845 5**
	G-mean	0.900 1	0.891 6	0.863 8	**0.920 1**
seismic	AUC	0.559 6	0.555 3	0.540 4	**0.565 1**
	F_1-score	0.169 8	0.161 9	0.138 0	**0.175 6**
	G-mean	0.423 0	0.404 6	0.367 2	**0.440 3**
wilt	AUC	0.912 5	0.895 4	0.884 9	**0.928 3**
	F_1-score	0.769 4	0.755 6	0.780 9	**0.792 2**
	G-mean	0.909 5	0.890 5	0.878 4	**0.926 4**
数据集	指标	RF			
		SMOTE	BSM2	KSMOTE	Pkmeans-SMOTE
haberman	AUC	0.679 8	0.669 7	0.682 3	**0.688 4**
	F_1-score	0.756 8	0.766 3	**0.793 5**	0.752 8
	G-mean	0.551 1	0.506 8	0.496 1	**0.566 5**
moon	AUC	0.988 0	0.983 2	0.979 3	**0.988 1**
	F_1-score	0.838 0	0.840 5	0.852 2	**0.859 3**
	G-mean	0.920 6	0.917 2	0.903 8	**0.943 3**
segment2	AUC	0.995 8	0.994 2	0.995 7	**0.995 9**
	F_1-score	0.890 0	0.849 7	0.844 0	**0.893 0**
	G-mean	0.947 7	0.932 7	0.864 6	**0.948 0**
seismic	AUC	**0.738 6**	0.738 5	0.727 8	0.724 2
	F_1-score	0.187 9	0.191 2	0.124 7	**0.194 1**
	G-mean	0.382 1	0.360 5	0.259 5	**0.392 9**
wilt	AUC	0.989 6	0.988 9	0.986 2	**0.990 1**
	F_1-score	0.846 1	0.813 6	0.826 4	**0.857 6**
	G-mean	0.932 0	0.902 4	0.860 4	**0.936 8**

通过 AUC、F_1-score 和 G-mean 三个指标评估，结果显示 Pkmeans-SMOTE 在大多数情况下优于其他方法，尤其在提升 AUC、F_1-score 和 G-mean 方面表现突出。在 AUC 指标上，Pkmeans-SMOTE 在决策树分类器上在多个数据集的表现优于其他方法，特别是在 Haberman 和 Segment2 数据集上，其 AUC 分别为 0.592 1 和 0.923 0，显著高于 SMOTE（分别为 0.552 5 和 0.903 5）。然而，SMOTE 和 KSMOTE 在一些数据集（如 Moon 和 Wilt）上也表现良好，特别是在 Wilt 数据集上，SMOTE 的 AUC 为 0.989 6，接近 Pkmeans-SMOTE 的 0.990 1。在 F_1-score 方面，Pkmeans-SMOTE 在 Haberman 和 Segment2 数据集上的表现最好，分别为 0.761 3 和 0.845 5，明显优于 SMOTE 和 KSMOTE。但 BSMOTE2 在 Wilt 数据集上的 F_1-score 表现较好，达到 0.857 6。在 G-mean 指标上，Pkmeans-SMOTE 在 Haberman 和 Segment2 数据集上表现最佳，分别为 0.578 3 和 0.920 1，显著高于其他方法。尽管在一些数据集如 Wilt，SMOTE 和 KSMOTE 也表现不错，但 Pkmeans-SMOTE 在整体表现上更具优势。

图 1 展示了不同过采样方法在不平衡数据分布上的处理效果，通过可视化合成样本的分布，直观对比五种方法的性能表现：（a）图为 NoSMOTE，表示未进行过采样处理的数据原始分布；（b）图为 SMOTE，展示经典 SMOTE 方法生成的合成样本分布；（c）图为 Borderline-SMOTE2，表示基于边界样本的过采样分布；（d）图为 kmeans-SMOTE，展示引入 K-means 聚类的过采样方法的效果；（e）图为 Pkmeans-SMOTE，代表基于改进聚类的过采样方法的表现。蓝色点为原始多数类样本，红色点为原始少数类样本，红色叉号表示合成的少数类样本。

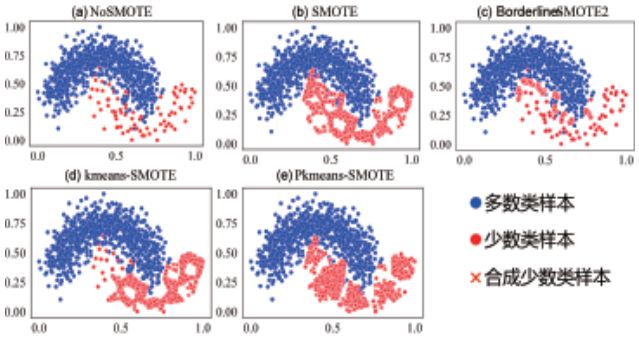

图 1 四种过采样方法在 moon 数据集上

不同过采样方法通过生成合成样本来缓解数据不平衡问题。SMOTE 通过插值生成样本，分布均匀，但未考虑边界信息，可能与原始少数类样本较远。Borderline-SMOTE2 关注边界样本，改善了决策边界的学习，但边界区域样本集中。kmeans-SMOTE 结合 K-means 聚类，在聚类中心生成样本，改善了分布，但仍有局限性。Pkmeans-SMOTE 通过层次化样本划分与局部聚类优化生成的样本，分布更均匀，边界区域更合理，避免了样本过于集中的问题，提升了分类器在边界区域的学习能力。此外，该方法增加了少数类在决策边界附近的代表性，丰富了训练数据的多样性，从而提升模型的泛化能力。

4 结论

本文提出的 Pkmeans-SMOTE 方法结合 Tomek Links 和 K-means 聚类技术，改进了传统过采样方法，显著提升了少数类样本的质量和分布合理性。实验结果表明，Pkmeans-SMOTE 在多个不平衡数据集上，尤其在 AUC、F_1-score 和 G-mean 等指标上优于其他过采样方法。该方法有效避免了样本冗余和噪声问题，并提高了边界区域的学习能力，展现了良好的实际应用潜力。未来的研究可进一步探索如何结合其他技术优化 Pkmeans-SMOTE 的性能，提升其在更复杂和动态数据环境中的适应性。

参考文献：

[1] GUO H X, LI Y J, SHANG J, et al. Learning from class-imbalanced data: review of methods and applications[J]. Expert systems with applications, 2017, 73(5): 220-239.

[2] KAUR H, PANNU H S, MALHI A K. A systematic review on imbalanced data challenges in machine learning: applications and solutions[J]. ACM computing curveys (CSUR), 2019, 52(4): 1-36.

[3] 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(1): 321-357.

[4] HAN H, WANG W Y, MAO B H. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning[J]. Advances in intelligent computing, 2005,3644: 878-887.

[5] DOUZAS G, BACAO F, LAST F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE[J]. Information sciences, 2018, 465(10): 1-20.

[6]KOZIARSKI M, KRAWCZYK B, WOŹNIAK M. Radial-based oversampling for noisy imbalanced data classification[J]. Neurocomputing, 2019, 343:19-33.

[7] FERNÁNDEZ A, GARCÍA S, HERRERA F, et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary[J]. Journal of artificial intelligence research, 2018, 61(1): 863-905.

[8] SEIFFERT C, KHOSHGOFTAAR T M, VAN HULSE J, et al. An empirical study of the classification performance of learners on imbalanced and noisy software quality data[J]. Information sciences, 2014, 259: 571-595.

基于改进 U-Net 的牙齿图像分割和抑噪研究

张 磊[1] 滕开良[1]
ZHANG Lei TENG Kailiang

摘 要 牙齿 X 光医学图像具有低对比度和高噪声特性，牙齿的形态和排列复杂性也增加了图像识别与分割的难度。基于此，文章提出了一种改进的 U-Net 模型，设计了动态注意力融合模块（dynamic attention fusion module, DAFM）增强特征提取能力。该模块对通道和空间两个维度的注意力特征进行自适应加权，降低冗余信息影响，并增强对关键特征的捕捉，提升模型对边界特征的识别能力。实验结果表明，改进后的模型在分割精度、边界一致性及细节还原方面优于原始 U-Net 模型，同时在乘性噪声、椒盐噪声和高斯噪声环境条件下仍保持稳定的分割性能。

关键词 牙齿分割；U-Net; 注意力；特征融合

doi：10.3969/j.issn.1672-9528.2025.03.035

0 引言

随着数字图像在医学领域的快速扩展，医学图像分割技术不断提升。牙齿医学图像具有低对比度和高噪声的特点，且因牙齿形态和排列多样，传统图像处理方法在分割任务上面临挑战[1]。目前，医学图像去噪方法可分为基于变换域[2]、统计域[3]及机器学习[4]三大类。传统的变换域和统计域方法依赖噪声先验信息，应用范围受到限制。而深度学习去噪方法可保留图像边缘细节，减少传统去噪方法中因平滑操作造成的信息损失[5]。

在牙齿医学图像分割的研究中，U-Net[6]的编码 - 解码结构通过跳跃连接将浅层和深层特征结合，保留更多的细节信息。ResUNet[7]的残差结构增强特征传递的稳定性。Attention U-Net[8]引入注意力机制，提升模型在复杂背景下对目标区域的分割性能。SegNet[9]模型使用池化索引保留了位置信息，使模型能够在解码过程中复原部分空间结构。

随着注意力机制在医学图像分割中的应用逐渐增多，自注意力机制和卷积块注意力在处理复杂结构的牙齿图像时展现出优势。自注意力机制关注每个像素与其他像素之间的关

1. 广西民族大学人工智能学院 广西南宁 530006

[9] CHEN B Y, XIA S Y, CHEN Z Z, et al. RSMOTE: a self-adaptive robust SMOTE for imbalanced problems with label noise[J]. Information sciences, 2021, 553(4): 397-428.

[10] LIU R. A novel synthetic minority oversampling technique based on relative and absolute densities for imbalanced classification[J]. Applied intelligence, 2023, 53(1): 786-803.

[11] SUN Z B, SONG Q B, ZHU X Y, et al. A novel ensemble method for classifying imbalanced data[J]. Pattern recognition, 2015, 48(5): 1623-1637.

[12]ZENG M, ZOU B J, WEI F R, et al. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data[C]//2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS). Piscataway: IEEE, 2016: 225-228.

[13]KANUNGO T, MOUNT D M, NETANYAHU N S, et al. An efficient k-means clustering algorithm: Analysis and implementation[J]. IEEE transactions on pattern analysis and machine intelligence, 2002, 24(7): 881-892.

[14]TIAN J L, ZHU LIN, ZHANG S Q, et al. Improvement and parallelism of k-means clustering algorithm[J].Tsinghua science & technology, 2012, 10(3):277-281.

[15] THABTAH F, HAMMOUD S, KAMALOV F, et al. Data imbalance in classification: experimental evaluation[J]. Information sciences, 2020, 513(3): 429-441.

【作者简介】

杨云熙（2000—），女，贵州贵阳人，硕士研究生，研究方向：计算机应用技术。

（收稿日期：2025-01-03）