文章编号:1671-6833(2025)06-0023-09

基于特征转换和少数类聚类的微生物数据扩增算法

温柳英, 郑天浩

(西南石油大学 计算机与软件学院,四川 成都 610500)

摘 要:微生物数据的高维、高零值率特性及少数类样本稀缺导致的类别不平衡,显著降低了分类器的少数类识别能力,而现有扩增算法对高不平衡比(*IR*)敏感且难以有效合成样本。针对此问题,提出了一种基于特征转换和少数类聚类的微生物数据扩增算法(FTMC)。首先,该算法在特征转换阶段采用主成分分析算法对高维数据进行降维,以缓解数据强稀疏性问题;其次,在少数类聚类阶段,使用 *K*-means 算法捕捉少数类的局部特征,获得多个聚类;再次,在聚类筛选阶段,基于每个聚类的密度和难度,结合 *IR* 和权重比来计算其权重值,并以此筛选出核心聚类子集,用于后续样本生成;最后,在样本扩增过滤阶段,利用线性插值算法,对筛选后的每个核心聚类进行样本扩增,并使用局部异常因子算法过滤异常点,确保扩增样本的质量。在 12 个微生物数据集上进行实验,并在 3 个分类器下对比 8 个同类型采样算法的性能,结果表明:FTMC 生成的样本更具多样性,在 *Recall* 指标上平均提高了 26.42%,证明该算法能正确识别更多的阳性样本。

关键词:微生物数据; 高维; 稀疏; 类别不平衡; 聚类; 数据扩增

中图分类号:TP391;Q939.9;TP311.13 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2025.06.006

随着生物技术的迅猛发展,微生物在医学鉴定、疾病诊断和治疗等领域的应用越来越广泛。微生物种类繁多且其在人体内的作用复杂,准确识别和分类微生物对于疾病的诊断和治疗至关重要。然而,微生物数据常存在样本不平衡的问题,影响机器学习模型的训练效果和医学鉴定的准确性[1]。此外,微生物群落数据面临高维数据、类别不平衡和过度稀疏等挑战,尤其在样本数量较少的情况下,OTUs 的数量远超样本数量[2]。尽管分类器能实现较高准确性,但误诊可能对患者造成致命影响,因此必须提高阳性样本的召回率,以识别更多患病样本[3]。

目前,微生物数据扩增的研究主要集中在以下方面:①传统数据扩增方法,典型的如合成少数类过采样法[4](synthetic minority over-sampling technique, SMOTE)、自适应合成采样法[5](adaptive synthetic sampling approach, ADASYN)和欠采样[6]等。这些方法虽然简单易行,但在处理较高 *IR* 的数据集时,往往生成的合成样本质量较低。在此基础上,王曦等[7]基于代价敏感和空间划分提出的 MFCS(cost-sensitive microbial data augmentation through matrix factorization)和温柳英等[8]基于矩阵分解和改进空间划分提出的 MFSP(fusing matrix factorization and space partitioning microbial data augmentation algorithm),虽然在一定程度上改进了传统方法的不足,但在高 *IR* 数据集上的表现仍待提升。②基于生成模型的扩增方法,生成对抗网络(generative adversarial networks, GAN)[9]和变分自编码器(variational auto-encoders, VAE)[10]等生成模型对数据扩增的效果显著,如 Wen 等[11]提出的 KGA(integrating kpca and gan for microbial data augmentation)算法等。这些方法通过学习数据的潜在分布生成合成样本,但在高不平衡比 *IR* 情况下容易受到模式崩溃的影响,导致生成样本的多样性不足。③集成学习与模型融合,通过多种扩增方法结合的方式以提升微生物数据的分类性能,如 Feng 等[12]基于集成边界和欠采样技术提出的集成分类算法。该方法更注重构建平衡数据集,但在高 *IR* 的微生物数据中,阳性样本过于稀少,欠采样难以达到理想平衡。

鉴于现有的扩增算法在高 *IR* 数据集表现不佳和样本多样性不足的局限性,本文提出了一种改进

收稿日期:2025-05-04;**修订日期**:2025-06-11

基金项目:中央引导地方科技发展专项项目(2021ZYD0003)

作者简介:温柳英(1983—),女,四川成都人,西南石油大学副教授,博士,主要从事机器学习、不平衡学习和微生物信息学研究,E-mail:wenliuying1983@163.com。

引用本文:温柳英,郑天浩. 基于特征转换和少数类聚类的微生物数据扩增算法[J]. 郑州大学学报(工学版),2025,46(6):23-31.(WEN L Y, ZHENG T H. Microbial data augmentation algorithm based on feature transformation and minority clustering[J]. Journal of Zhengzhou University (Engineering Science),2025,46(6):23-31.)

的微生物数据扩增算法。该算法运用主成分分析(principal component analysis,PCA)算法进行特征转换,降低特征维度并保留关键信息,以缓解强稀疏性问题;对少数类聚类后,依据聚类密度、难度及 *IR* 和权重比计算权重值,选取合适样本扩增,降低过拟合与欠拟合概率;最后利用局部异常因子(local outlier factor,LOF)算法过滤异常样本,保障合成样本质量与分布合理,提升样本多样性和分类性能。

1 FTMC 算法

针对微生物数据的高维、稀疏以及类别不平衡等问题,本文提出了一种基于特征转换和少数类聚类的数据扩增算法(data augmentation algorithmbased on feature transformation and minority clustering,FTMC)。该算法由特征转换、少数类聚类、筛选核心聚类和样本扩增过滤 4 个阶段组成,如图 1 所示。

(1)特征转换阶段。基于 PCA 算法对归一化数据特征进行变换降维。该过程可以从原始数据中去除冗余信息,实现数据简化和压缩,增强数据的线性可分性,以缓解其强稀疏性问题。

(2)少数类聚类。对特征转换后矩阵中的少数类样本进行 *K*-means 聚类,以捕捉数据的局部特征,同时避免在扩增样本时过度集中在某些特征区域。

(3)筛选核心聚类。该阶段是 FTMC 算法中的关键步骤,主要目的是计算每个聚类的密度和难度,并结合 *IR* 和调整权重比 w 来计算权重值,然后筛选权重在均值±标准差内的核心聚类进行后续的样本生成及过滤。分配权重和筛选聚类有助于生成更具代表性的合成样本,避免过度集中在某些特征区域,以缓解过拟合和欠拟合,提高数据多样性。

(4)样本扩增过滤。对每个核心聚类通过线性插值生成合成样本,并使用 LOF 算法检测数据中的异常点进行数据过滤,以确保数据质量,提高模型的鲁棒性。

1.1 特征转换

在此阶段,先将含大量零值的原始数据进行标准化处理,如式(1)所示;再基于 PCA 对标准化数据特征变换[13],如式(2)所示。通过选取主成分映射数据,有效降低特征矩阵维数。

$$\boldsymbol{Z}_{ij}=\frac{\boldsymbol{D}_{ij}-\mu_i}{\sigma_i}; \tag{1}$$

$$\boldsymbol{D}'=\mathrm{PCA}(\boldsymbol{Z},g)。 \tag{2}$$

式中:$\boldsymbol{Z}_{ij}$ 为标准化数据矩阵;$\boldsymbol{D}_{ij}$ 为原始数据矩阵;μ_i 为第 i 个特征的均值;σ_i 为第 i 个特征的标准差;$\boldsymbol{D}'$为特征转换后数据矩阵;$\boldsymbol{Z}$ 为标准化数据矩阵;g 为降维后的维度。

1.2 少数类聚类

该阶段主要针对高维数据中的少数类样本,通过聚类捕捉其分布在不同局部区域的特征,挖掘隐藏特性,识别多种模式,避免局部过度集中,使合成样本更具代表性。

首先,对特征转换后的 $\boldsymbol{D}'$中的少数类样本 $\boldsymbol{M}_+$ 进行 *K*-means 聚类[14],聚类的集合表示为

$$C=\{C_i\mid i=1,2,\cdots,c\}=K\text{-means}(\boldsymbol{M}_+,c)。 \tag{3}$$

其次,计算每个聚类的密度和难度,密度 D_{C_i} 是指在聚类 C_i 中的样本数量。密度越高,表示该聚类中的样本越多,如式(4)所示:

$$D_{C_i}=\sum_{j=1}^{N_{C_i}}C_i。 \tag{4}$$

式中:N_{C_i} 表示聚类 C_i 中样本数量。

难度 F_{C_i} 表示样本到聚类 C_i 中心的平均距离。距离越大,表示该聚类的样本分布越分散,难度越高,如式(5)所示:

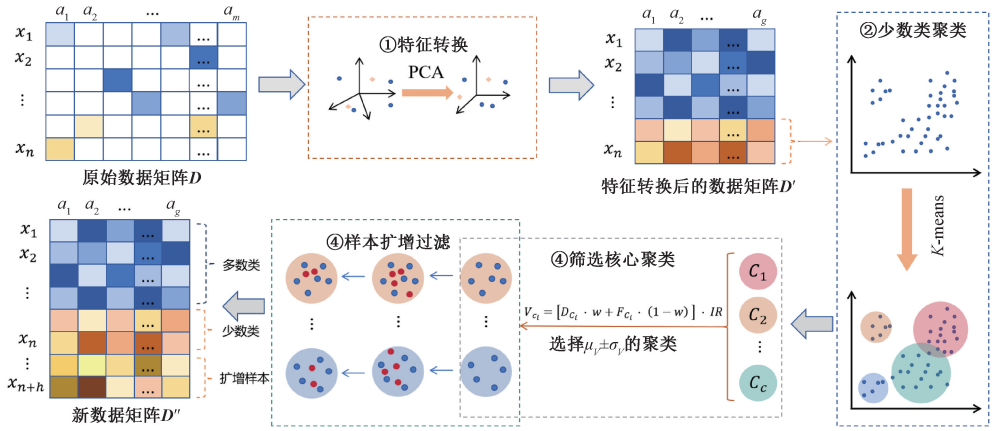


图 1 FTMC 算法框架

Figure 1 FTMC algorithm framework

$$F_{C_i}=\frac{1}{N_{C_i}}\sum_{j=1}^{N_{C_i}}\| C_{i,j}-\text{center}(C_i)\|。\tag{5}$$

式中:N_{C_i} 为聚类 C_i 中的样本数量;$C_{i,j}$ 为聚类 C_i 中的样本;center(C_i) 为聚类 C_i 的中心;$\|\cdot\|$表示欧几里得距离。

1.3　筛选核心聚类

筛选核心聚类是根据聚类的难度和密度信息计算每一个聚类的权重值,以此为基础筛选出核心聚类,用于后续的样本扩增。

聚类 C_i 的权重值 V_{C_i} 计算公式为

$$V_{C_i}=(D_{C_i}\cdot w+F_{C_i}(1-w))\cdot IR。\tag{6}$$

式中:权重比 w 介于 0 和 1 之间,用于平衡难度和密度的影响;IR 为原始矩阵 $\boldsymbol{D}$ 的不平衡比,即多数类样本数量与少数类样本数量的比值。

该过程引入 IR 计算权重值,以放大聚类差异。IR 高时,少数类样本少,各聚类样本的特征区分度不显著,权重值差异小,难以直接有效比较和筛选;IR 低时,少数类样本相对较多,密度 D_{C_i} 和难度 F_{C_i} 能够在一定程度上反映样本的分布及特征,对筛选结果的影响较小。

为了筛选更具代表性的核心聚类集合,需要计算权重值 V 的均值 μ_V 和标准差 σ_V,以此确定聚类筛选的边界。均值和标准差分别为

$$\mu_V=\frac{1}{N_{C_i}}\sum_{1}^{N_{C_i}}V_{C_i};\tag{7}$$

$$\sigma_V=\sqrt{\frac{1}{N_{C_i}}\sum_{1}^{N_{C_i}}(V_{C_i}-\mu_V)^2}。\tag{8}$$

根据均值和标准差,筛选出在上界 b_{upper} 和下界 b_{lower} 内的聚类,作为核心聚类。

$$b_{\text{upper}},b_{\text{lower}}=\mu_V\pm\sigma_V。\tag{9}$$

上界 b_{upper} 外的聚类样本多且分布聚集,分类器对其识别效果相对理想,扩增可能导致过拟合问题;而下界 b_{lower} 外的聚类样本少,含噪声样本,难以代表整体数据特征,扩增易增加欠拟合风险。这种筛选方式能避免聚焦特定区域,以减轻过拟合与欠拟合,提升扩增样本的多样性和代表性。

1.4　样本扩增过滤

样本扩增过滤是通过扩增技术生成合成样本,并通过噪声过滤来增强少数类样本。对于一个少数类 x_i,使用 k 近邻法来搜寻与其距离最近的 k 个少数类样本。此处,样本间的距离定义为 n 维特征空间里的欧氏距离[3],公式如下:

$$d(x_i,x_j)=\sqrt{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2}。\tag{10}$$

式中:x_i 为某个少数类样本;x_j 为其他少数类样本(其中 $j\neq i$);k 为某个少数类样本的某一特征。

将所得距离进行排序,从距离最小的 k 个近邻中随机选取一个 $\hat{x}_i$,生成新的样本。扩增公式为

$$\boldsymbol{x}'=x_i+(\hat{x}_i-x_i)\delta。\tag{11}$$

式中:$\boldsymbol{x}'$ 为新生成的样本矩阵;δ 为介于 0 和 1 之间的随机值。

扩增样本的数量依据数据集中少数类和多数类样本的数量差异来确定。根据不同核心聚类内样本的权重值来决定生成多少个合成样本。首先,将各权重值归一化为相应比例;然后,根据所需生成的样本总量,按此比例对样本进行分配。这种方法可以确保生成的样本数量总量固定,同时不同核心聚类之间的差异性更明显。归一化权重值 V'_{C_i} 和第 i 个核心聚类扩增样本的数量 C_i_num 分别为

$$V'_{C_i}=\frac{V_{C_i}}{\sum_{j=1}^{m}V_{C_j}};\tag{12}$$

$$C_i\_num=\lfloor h\cdot V'_{C_i}\rfloor。\tag{13}$$

式中:h 为扩增样本的总数量。

噪声过滤应用 LOF 算法识别和过滤噪声样本,通过计算每个点的离群因子来判断是否为离群点[15]。若离群因子远大于 1,即为离群点;若离群因子接近 1,则为正常点。离群因子 $LOF_k(i)$ 为

$$LOF_k(i)=\frac{\sum_{o\in N_k(i)}\frac{\rho_k(o)}{\rho_k(i)}}{N_k(i)}。\tag{14}$$

式中:$\rho_k(i)$ 为对象 i 的 k 最近邻点的局部可达密度;$N_k(i)$ 表示与对象 i 之间距离$\leqslant k$ 的对象集合。

经过过滤后保留下的样本矩阵 $\boldsymbol{x}_{\text{new}}$ 为

$$\boldsymbol{x}_{\text{new}}=LOF(\boldsymbol{x}')。\tag{15}$$

最后,将扩增过滤得到的样本矩阵与少数类样本矩阵 $\boldsymbol{M}_+$ 相组合得到 $\boldsymbol{M}'_+$,再与多数类样本矩阵 $\boldsymbol{M}_-$ 结合,得到新的扩增矩阵 $\boldsymbol{D}''$。

1.5　FTMC 算法

算法 1　FTMC 算法。

输入:原始数据集 $\boldsymbol{D}$,降维维度 g,聚类数量 c,权重比 w;

输出:扩增后的数据集 $\boldsymbol{D}''$。

① 对原始数据集 $\boldsymbol{D}$ 标准化得到数据集 $\boldsymbol{Z}$;

② 对 $\boldsymbol{Z}$ 使用 PCA 算法,得到维度为 g 的数据集 $\boldsymbol{D}'$;

③ 划分 $\boldsymbol{D}'$得到多数类样本矩阵 $\boldsymbol{M}_-$ 和少数类样本矩阵 $\boldsymbol{M}_+$;

④ 对 $\boldsymbol{M}_+$ 使用 K-means 算法,划分为 c 个聚类。记为 $C=\{C_i\mid i=1,2,\cdots,c\}$;

⑤ for each C_i in C do
⑥ 根据式(4)计算密度 D_{C_i};
⑦ 根据式(5)计算难度 F_{C_i};
⑧ 根据式(6)计算权重值 V_{C_i};
⑨ end for
⑩ 初始化扩增样本矩阵 $\boldsymbol{x}'$和扩增过滤后样本矩阵 $\boldsymbol{x}_{\text{new}}$ 为空集合;
⑪ 根据式(9)从 C 中筛选出核心聚类,组成核心聚类集合 $C'=\{C_1,C_2,\cdots,C_m\}$;
⑫ for each C_m in C' do
⑬ for each x_i in C_m do;
⑭ 根据式(11)对样本 x_i 进行扩增,得到新的扩增样本,加入 $\boldsymbol{x}'$;
⑮ 使用 LOF 算法对 $\boldsymbol{x}'$中的样本进行异常检测和过滤,将过滤后的样本加入 $\boldsymbol{x}_{\text{new}}$;
⑯ end for
⑰ end for
⑱ 将 $\boldsymbol{x}_{\text{new}}$ 与原始少数类样本矩阵 $\boldsymbol{M}_+$ 合并,得到扩增后的少数类样本 $\boldsymbol{M}'_+$;
⑲ 将 $\boldsymbol{M}'_+$ 与多数类样本矩阵 $\boldsymbol{M}_-$ 合并,得到新的扩增数据集 $\boldsymbol{D}''$;
⑳ 返回扩增后的数据集 $\boldsymbol{D}''$。

1.6 复杂度分析

FTMC 算法的实现流程涵盖了特征转换、少数类聚类、核心聚类计算、合成样本生成及噪声处理等关键环节。算法 1 中各环节复杂度分析如下。

(1)特征转换:先计算协方差矩阵,时间复杂度为 $O(nm^2)$,其中 n 为原始样本数,m 为原始特征维度;再进行特征值分解,时间复杂度为 $O(m^3)$。该环节的时间复杂度为 $O(nm^2)$。

(2)K-means 聚类:对少数类样本聚类的时间复杂度为 $O(cpt)$,其中,c 为聚类数量;p 为少数类数量;t 为迭代次数。

(3)核心聚类计算:该阶段分为两步,一是计算每个聚类的密度和难度,时间复杂度为 $O(n)$ 和 $O(nt)$;二是计算每个聚类的权重值,时间复杂度为 $O(c)$。该环节的时间复杂度为 $O(nt)$。

(4)合成样本生成:首先,根据少数类样本间的距离确定 k 近邻,时间复杂度为 $O(p^2t)$;其次,进行排序,时间复杂度为 $O(pk\log p)$,其中,$\log p$ 为排序复杂度;最后,通过生成合成样本,时间复杂度为 $O(hgk)$,其中,h 为合成样本数量。该环节的时间复杂度为 $O(p^2t)$。

(5)噪声处理:首先,应用 LOF 计算生成样本间距离,时间复杂度为 $O(h^2t)$;其次,排序确定 k 近邻,时间复杂度为 $O(hk\log h)$;最后,计算可达密度,时间复杂度为 $O(hk)$。该环节的时间复杂度为 $O(h^2t)$。

FTMC 算法针对高维微生物数据集所设计,数据维度规模较大,其中特征转换环节的时间复杂度占据主导地位,其时间复杂度为 $O(nm^2)+O(cpt)+O(nt)+O(p^2t)+O(h^2t)=O(nm^2)$。

2 实验结果及分析

2.1 实验数据

本文所使用 D003015 等 12 个数据集均源自 AutoML[16] 网站(http://39.100.246.211:8050/Dataset),不平衡微生物数据集详细信息见表 1。

表 1 不平衡微生物数据集

Table 1 Unbalanced microbial datasets

数据集	数据量			IR	零值率/%
	特征数	阳性	阴性		
D003015	227	812	1 255	1.55	96.1
D001327	153	462	1 170	2.53	94.1
D015212	162	359	1 201	3.35	94.5
D007410	153	274	1 170	4.27	94.1
D003863	145	228	1 170	5.13	93.8
D012559	145	224	1 170	5.22	86.7
D001714	149	315	1 216	8.16	94.0
D0067877	140	86	1 170	13.60	93.6
D008107	136	62	1 170	18.87	93.4
D007674	136	59	1 170	19.83	93.4
D002446	137	58	1 170	20.17	93.4
D004827	137	31	1 170	37.74	93.4

2.2 对比算法

2.2.1 扩增算法

SMOTE(SMO)通过在少数类样本及其 k 近邻连线随机生成新样本;根据少数类样本的分布密度,ADASYN(ADA)在较为稀疏的区域生成更多样本;K_means_ SMOTE (KSMO)[17] 结合 K-means 与 SMOTE,先聚类再在各簇内生成样本;GAN 由生成器与判别器对抗生成新样本;VAE 是基于变分推断通过学习少数类的潜在分布并采样解码生成新样本;MFCS 利用矩阵分解将原始数据分解为对象和特征子空间后进行扩增;KGA 通过 KPCA 映射到低维空间,再用 GAN 扩增正样本并控制比例;MFSP 包括矩阵分解、空间划分及引入代价因子,最后根据类内和类间距离过滤合成样本。

2.2.2 聚类算法

DBSCAN[18] 是一种基于密度的聚类算法,通过

邻域半径和最小点数识别核心点和噪声点,可发现任意形状的聚类。OPTICS[19]是其改进版,计算可达距离生成有序列表,支持按需截取聚类。AGNES[20]是一种凝聚式层次聚类算法,从各点独立开始,逐步合并最近类形成聚类树,无须预设聚类数。

2.3　评价指标

本文使用表 2 的混淆矩阵,基于 *Recall*、*G-mean* 和 *AUC* 这 3 个指标来评估不同采样算法的分类性能:

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

$$G\text{-}mean = \sqrt{\frac{TP}{TP + FN} \cdot \frac{TN}{FP + TN}}; \tag{17}$$

$$AUC = \frac{\sum_{i \in M_+} rank_i - \frac{(TP + FN)(TP + FN + 1)}{2}}{(TP + FN)(FP + TN)}。 \tag{18}$$

式中:$\sum_{i \in M_+} rank_i$ 为正样本的序号之和。

表 2　混淆矩阵

Table 2　Confusion matrix

分类	预测正类	预测负类
真实正类	*TP*	*FN*
真实负类	*FP*	*TN*

2.4　扩增结果

本节使用 *t*-SNE 算法进行数据可视化。图 2 为 D003863、D007674 和 D004827 经过 SMOTE、MFSP 和 FTMC 这 3 种不同扩增算法的前后对比效果。

图 2 显示了 D003863 原始数据中正负样本随机分布,难以区分的数据经 SMOTE 与 MFSP 算法扩增后,虽数量平衡,但样本分布集中甚至重叠。D007674 和 D004827 因 *IR* 高,原始数据的正样本稀缺且分布无序,经 SMOTE 和 MFSP 算法扩增后虽缓解了数量不平衡,但样本分布仍较为集中。相比之下,FTMC 算法扩增后,3 个数据集的正负样本分布规律明显,边界清晰,可区分度更高,表明其在数据质量与样本多样性方面具有明显优势,更利于分类器建模与评估。

2.5　分类性能

实验将数据集按 8∶2划分为训练集和测试集,保持测试集的不平衡比与原始数据一致。每组实验重复 10 次,并取均值以减少随机性影响。随后在 12 个数据集上对比了 FTMC 与 8 种扩增方法的分类性能。

图 3 为各个扩增方法在 3 种分类器下的分类结果。其中,KNN 表示 *K*-nearest neighbour;SVM 表示 support vector machine;LR 表示 logistic regression。

由图 3 可知,在 12 个数据集以及 KNN、SVM 和 LR 分类器的实验环境下,FTMC 算法在 *Recall*、*G-mean* 和 *AUC* 关键指标上,均展现出超越其他最新采样算法的性能优势,特别在高 *IR* 数据集上表现突

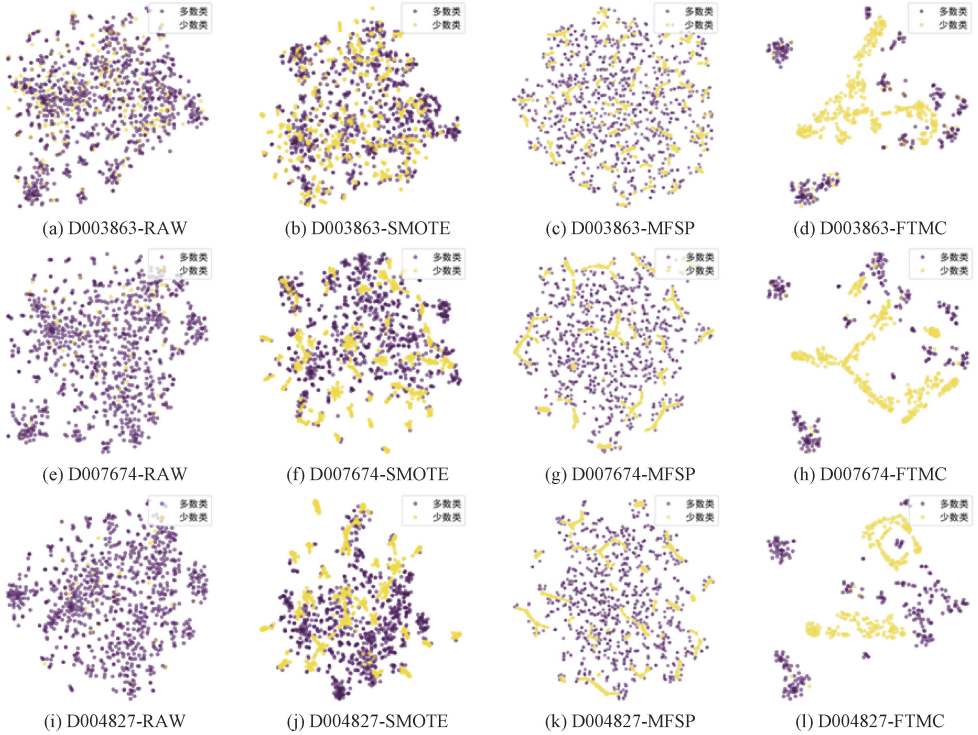
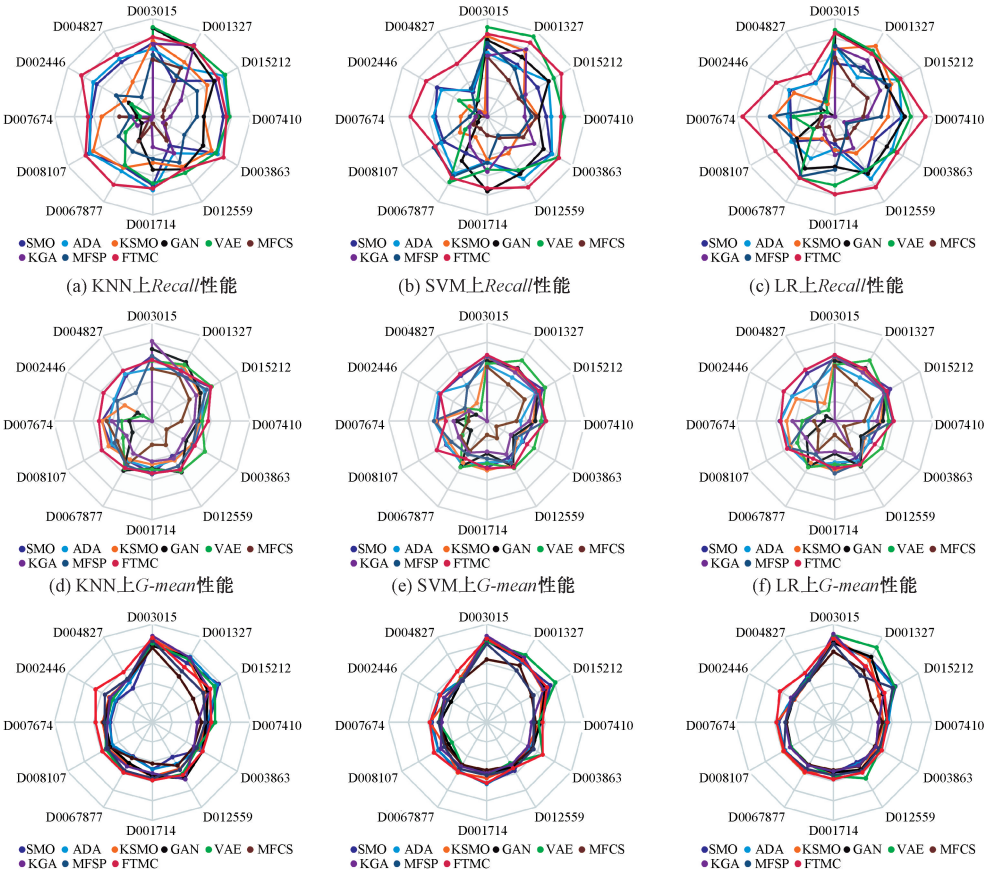


图 2　不同算法在数据集 D003863、D007674 和 D004827 上的扩增结果

Figure 2　Augmentation results of different algorithms on datasets D003863, D007674 and D004827



(a) KNN上*Recall*性能 (b) SVM上*Recall*性能 (c) LR上*Recall*性能

(d) KNN上*G-mean*性能 (e) SVM上*G-mean*性能 (f) LR上*G-mean*性能

(g) KNN上*AUC*性能 (h) SVM上*AUC*性能 (i) LR上*AUC*性能

图 3 *Recall*、*G-mean* 和 *AUC* 性能对比结果

Figure 3 Comparison results of *Recall*, *G-mean* and *AUC* performance

出,例如在 D002446 数据集上,FTMC 算法的 *Recall* 达到 0.89,*G-mean* 和 *AUC* 分别为 0.59 和 0.74,各项指标不仅稳定且全面超越对比算法,充分体现了 FTMC 算法在分类性能上的优越性,证明其在识别正样本上优于其他算法。

2.6 消融实验

为探究特征转换阶段、少数类聚类阶段与筛选核心聚类阶段以及应用不同聚类算法对生成样本质量和分类性能的影响,本小节分为算法阶段和聚类阶段两个部分进行消融实验。

2.6.1 算法阶段

本小节选取 D003863、D007674 和 D004827 数据集对 FTMC 算法进行消融实验,实验结果如表 3 所示。在表 3 中,FTMC-PCA 表示消融特征转换的算法;FTMC-*K*-means 表示消融少数类聚类的算法;FTMC-Screening 表示消融筛选核心聚类的算法。

由表 3 可知,FTMC 算法中的特征转换阶段、少数类聚类阶段以及筛选核心聚类阶段均对分类性能的提升起到了重要的作用。特别是特征转换阶段,在 *Recall* 和 *G-mean* 指标上展现了显著的提升效果,并且在 *AUC* 指标上也有一定的提高。说明 FTMC 算法能够有效解决微生物数据的高维性、稀疏性和类别不平衡问题,使得分类器识别出更多的阳性样本,以验证 FTMC 算法的可行性和有效性。

2.6.2 聚类阶段

在 12 个数据集上的聚类阶段应用不同聚类算法(*K*-means、DBSCAN、OPTICS 和 AGNES)进行消融实验,如图 4 所示。

表 3 消融实验

Table 3 Ablation experiment

模型	D003863			D007674			D004827		
	Recall	*G-mean*	*AUC*	*Recall*	*G-mean*	*AUC*	*Recall*	*G-mean*	*AUC*
FTMC-PCA	0.05	0.21	0.52	0.16	0.38	0.52	0.13	0.31	0.47
FTMC-*K*-means	0.37	0.17	0.54	0.49	0.21	0.47	0.47	0.14	0.51
FTMC-Screening	0.60	0.44	0.53	0.65	0.45	0.52	0.42	0.45	0.50
FTMC	**0.73**	**0.47**	**0.57**	**0.72**	**0.55**	**0.53**	**0.51**	**0.58**	**0.60**

由图 4 可知，*K*-means、DBSCAN 和 OPTICS 的分类性能良好，且受影响较小，而 AGNES 却大多表现欠佳；*K*-means 表现优秀，在处理高 *IR* 的数据集时，优势明显。综合 *Recall*、*G-mean* 和 *AUC* 指标来看，*K*-means 算法的适应性与优势更为显著。

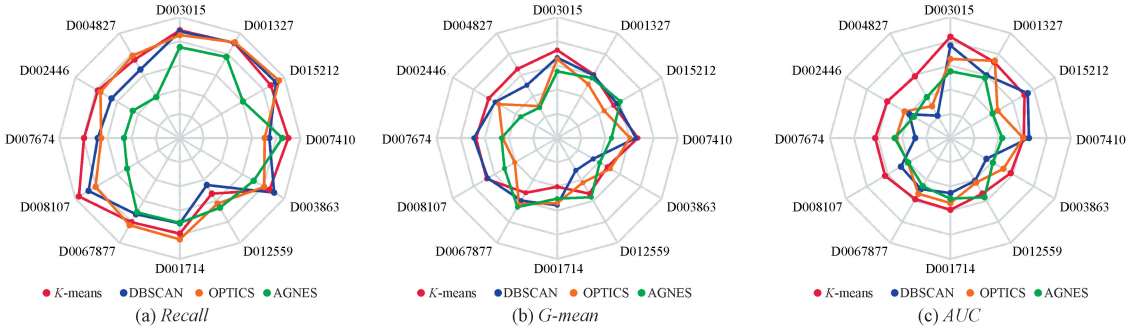
2.7　超参实验

FTMC 模型有 3 个关键的超参数：降维维度 g、聚类数量 c 和权重比 w。本节通过 D007674 数据集在 3 个分类器上测试不同取值的影响，结果如图 5 所示。

结果表明，$10 \leqslant g \leqslant 14$ 时，*Recall* 稳定且较高，*G-mean* 波动较小，*AUC* 相对稳定且小范围浮动；$5 \leqslant c \leqslant 30$ 时，*Recall* 波动大，$20 \leqslant c \leqslant 25$ 时，*G-mean* 稳定，*AUC* 小范围内浮动；$0.2 \leqslant w \leqslant 1.0$ 时，*Recall* 波动大，$w = 0.6$ 附近较优，$0.6 \leqslant w \leqslant 0.8$ 时，*G-mean* 稳定，*AUC* 小范围内浮动，$0.8 \leqslant w \leqslant 1.0$ 时，AUC 略有下降。综合来看，不同参数取值对算法性能的影响各异，在实际应用中需根据具体情况选择合适的参数，以优化算法性能。

2.8　FTMC 局限性

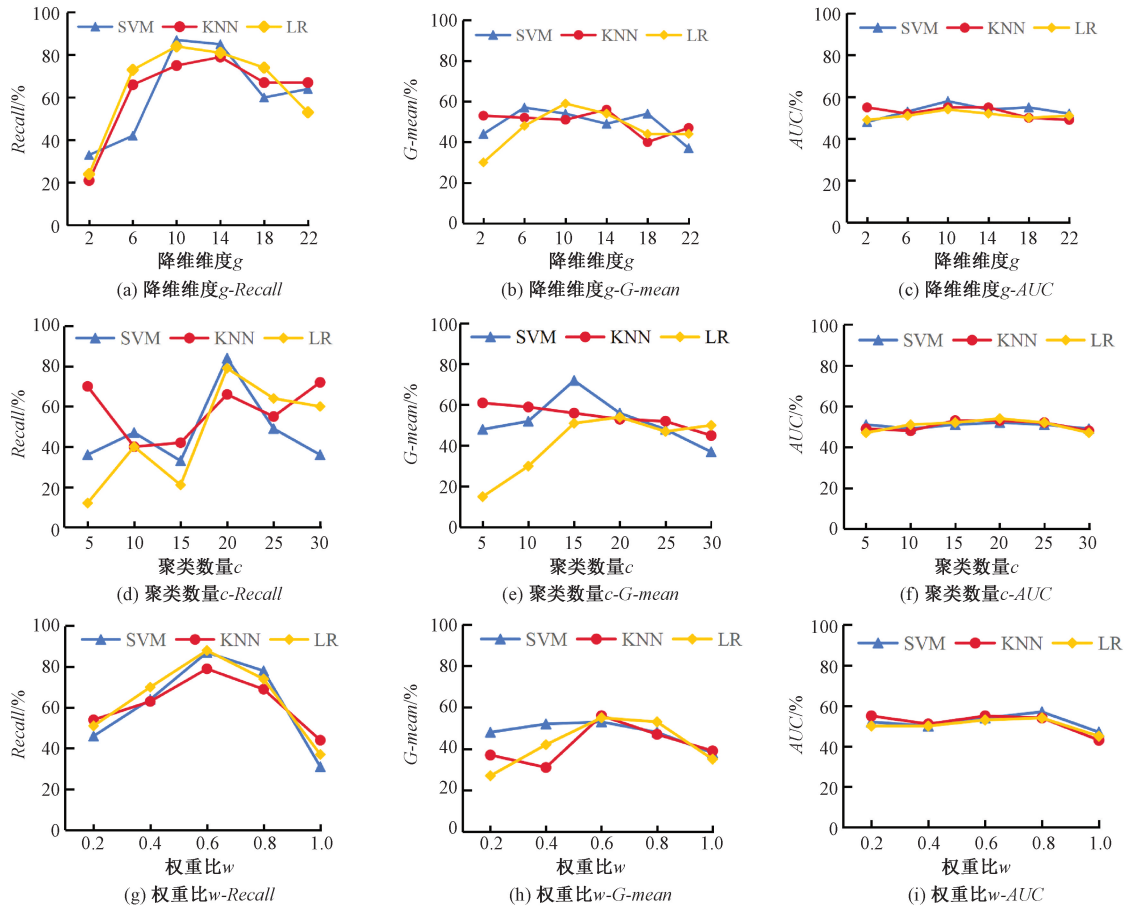
FTMC 算法有效缓解了微生物数据的高维性、



(a) *Recall*　(b) *G-mean*　(c) *AUC*

图 4　FTMC 算法应用不同聚类算法的各指标结果

Figure 4　Results of various indicators for the FTMC algorithm with different clustering algorithms



(a) 降维维度*g*-*Recall*　(b) 降维维度*g*-*G-mean*　(c) 降维维度*g*-*AUC*

(d) 聚类数量*c*-*Recall*　(e) 聚类数量*c*-*G-mean*　(f) 聚类数量*c*-*AUC*

(g) 权重比*w*-*Recall*　(h) 权重比*w*-*G-mean*　(i) 权重比*w*-*AUC*

图 5　FTMC 算法取不同 g、c 和 w 值时在 3 个分类器上的性能

Figure 5　Performance results of the FTMC algorithm on three classifiers with different values of g, c and w

稀疏性和类别不平衡问题,在 *Recall*、*G-mean* 和 *AUC* 上表现出色,均优于现有采样算法。其各阶段对分类性能的提升均起到重要作用,生成的扩增样本兼具多样性和代表性,但该算法仍存在一定的局限性。

在分类性能方面,虽然 FTMC 算法在大多数数据集上表现良好,但是处理 *IR* 过低的数据集时,FTMC 可能不再具有显著的优势。

就参数选择而言,FTMC 模型涉及数据维度、聚类数量和权重比等关键超参数。从图 5 的超参实验的结果来看,不同的参数组合对算法性能影响较大,且在实际应用中很难预先确定最优参数组合,这增加了算法应用的难度与不确定性,参数的调整过程也需要耗费大量计算资源与时间成本。

3 结论

针对微生物数据的高维性、稀疏性以及类别不平衡等问题,本文提出一种基于特征转换和少数类聚类的数据扩增算法 FTMC。首先,利用 PCA 算法进行特征转换,借此降低特征维度、保留重要信息,以缓解强稀疏性问题;其次,通过对少数类进行聚类,分析识别其不同模式,防止样本扩增时过度聚焦特定特征,使扩增样本特征覆盖更全面;随后,综合各聚类的密度、难度,结合 *IR* 与权重比调整来计算权重值,选择权重值在均值和标准差范围内核心聚类内的样本进行扩增,降低过拟合与欠拟合概率,有效提升了扩增样本的多样性和代表性;最后,应用 LOF 算法检测并过滤异常样本,确保合成样本集的数据质量与分布合理性,提高样本多样性及分类性能。实验结果表明,通过 FTMC 算法生成的扩增样本,在数据质量与分布合理性方面均表现出色,尤其是在处理 *IR* 较高数据集时,这种优势更为凸显。这些样本更具多样性和代表性,有效解决了微生物数据的高维性、稀疏性以及类别不平衡等关键问题。

参考文献:

[1] MEGAHED F M, CHEN Y J, MEGAHED A, et al. The class imbalance problem[J]. Nature Methods, 2021, 18(11): 1270-1272.

[2] 田鸿朋,张震,张思源,等. 复合可靠性分析下的不平衡数据证据分类[J]. 郑州大学学报(工学版),2023,44(4):22-28.
TIAN H P, ZHANG Z, ZHANG S Y, et al. Imbala-nced data evidential classification with composite reliability [J]. Journal of Zhengzhou University (Engine-ering Science), 2023, 44(4): 22-28.

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

[4] 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.

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

[6] MOREO A, ESULI A, SEBASTIANI F. Distributional random oversampling for imbalanced text classification [C]//Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Informat-ion Retrieval. New York: ACM, 2016: 805-808.

[7] 王曦, 温柳英, 闵帆. 融合矩阵分解和代价敏感的微生物数据扩增算法[J]. 数据采集与处理, 2023, 38(2): 401-412.
WANG X, WEN L Y, MIN F. Fusing matrix factorization and cost-sensitive microbial data augmentation algorithm[J]. Journal of Data Acquisition and Processing, 2023, 38(2): 401-412.

[8] 温柳英, 吴俊, 闵帆. 融合矩阵分解和空间划分的微生物数据扩增方法[J]. 山东大学学报(理学版), 2025, 60(1): 14-28,44.
WEN L Y, WU J, MIN F. Fusing matrix factorization and space partition microbial data augmentation algorithm [J]. Journal of Shandong University (Natural Science), 2025, 60(1): 14-28, 44.

[9] LI Y, HUANG C, DING L Z, et al. Deep learning in bioinformatics: introduction, application, and perspective in the big data era[J]. Methods, 2019, 166: 4-21.

[10] LI Y X, CHAI Y, YIN H P, et al. A novel feature learning framework for high-dimensional data classification [J]. International Journal of Machine Learning and Cybernetics, 2021, 12(2): 555-569.

[11] WEN L Y, ZHANG X M, LI Q F, et al. KGA: integrating KPCA and GAN for microbial data augmentation [J]. International Journal of Machine Learning and Cybernetics, 2023, 14(4): 1427-1444.

[12] FENG W, HUANG W J, REN J C. Class imbalance ensemble learning based on the margin theory[J]. Applied Sciences, 2018, 8(5): 815.

[13] ABDI H, WILLIAMS L J. Principal component analysis [J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(4): 433-459.

[14] AHMED M, SERAJ R, ISLAM S M S. The *k*-means algorithm: a comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8): 1295.

[15] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: identifying density-based local outliers[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2000: 93-104.

[16] HE X, ZHAO K Y, CHU X W. AutoML: a survey of the state-of-the-art[J]. Knowledge-Based Systems, 2021, 212: 106622.

[17] 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: 1-20.

[18] DENG D S. DBSCAN clustering algorithm based on density[C]//2020 7th International Forum on Electrical Engineering and Automation (IFEEA). Piscataway: IEEE, 2020: 949-953.

[19] ANKERST M, BREUNIG M M, KRIEGEL H P, et al. OPTICS: ordering points to identify the clustering structure[C]//Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1999: 49-60.

[20] SCHILT K. The importance of being Agnes[J]. Symbolic Interaction, 2016, 39(2): 287-294.

Microbial Data Augmentation Algorithm Based on Feature Transformation and Minority Clustering

WEN Liuying, ZHENG Tianhao

(School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu 610500, China)

Abstract: The high-dimensional characteristics of microbial data, the high zero-value rate, and the scarcity of minority-class samples, which led to class imbalance, had significantly weakened classifiers′ ability to identify minority class. Existing augmentation algorithms are sensitive to high imbalance ratios (*IR*) and struggle to effectively synthesize samples. In this study a microbial data augmentation algorithm based on feature transformation and minority class clustering (FTMC) was presenteed. Firstly, the feature transformation stage used the principal components analysis algorithm to down thescale high-dimensional data to alleviate the problem of strong data sparsity. Subsequently, in the minority class clustering stage, the *K*-Means algorithm was used to capture the local features of the minority classes and obtain multiple clusters. In the cluster screening stage, based on the density and difficulty of each cluster, combined with the IR and weight ratio, its weight value was calculated and used to screen a subset of core clusters for subsequent sample generation. Finally, in the sample augmentation and filtering stage, a linear interpolation algorithm was used augment to the samples for each core cluster, and a local anomaly factor algorithm was used to filter outliers to ensure the quality of the augmented samples. The experiments were conducted on 12 microbial datasets and the performance was compared with 8 sampling algorithms of the same type with 3 classifiers. Results indicated that samples generated by FTMC were more diverse, with an average improvement of 26.42% in the Recall metric. This demonstrated that the algorithm could correctly identify more positive samples.

Keywords: microbial data; high-dimensional; sparsity; class imbalance; cluster; data augmentation