ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 2,Mar. 2023,pp. 401－412
DOI:10. 16337/j. 1004-9037. 2023. 02. 015

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

融合矩阵分解和代价敏感的微生物数据扩增算法

王　曦，温柳英，闵　帆

（西南石油大学计算机科学学院，成都 610500）

摘　要：微生物会对人类健康产生直接影响，对相关数据的分析有助于疾病诊断。然而，采集到的数据存在类不平衡与高稀疏性两个问题。现有的过采样方法在一定程度上可缓解数据的类不平衡，但是难以应对微生物数据的高稀疏性。本文提出了一种融合矩阵分解和代价敏感的数据扩增算法，其包含3个技术。首先，将原始矩阵分解为样本子空间和特征子空间；其次，利用样本子空间的正向量及其近邻向量生成合成向量；最后，根据合成向量与所有负向量的距离对其过滤。实验在8个微生物数据集上进行，同时与5种过采样算法对比，结果表明本文所提算法能够增强正样本的多样性，在识别出更多正样本的同时，分类结果的代价更低。

关键词：矩阵分解；代价敏感；微生物数据；高稀疏性；样本子空间；特征子空间

中图分类号：TP181　　**文献标志码：**A

Fusing Matrix Factorization and Cost-Sensitive Microbial Data Augmentation Algorithm

WANG Xi, WEN Liuying, MIN Fan

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

Abstract: Microorganisms have a direct impact on human health, and the analysis of relevant data is helpful for disease diagnosis. However, the collected data suffers from two problems: class imbalance and high sparseness. Existing oversampling methods can alleviate the class imbalance of data to a certain extent, but it is difficult to cope with the high sparsity of microbial data. This paper proposes a data augmentation algorithm that fuses matrix factorization and cost-sensitive, which consists of three techniques. First, the original matrix is decomposed into a sample subspace and a feature subspace. Second, the positive vectors of the sample subspace and their neighbor vectors are used to generate synthetic vectors. Finally, the synthetic vectors are filtered according to their distance from all negative vectors. The proposed algorithm is compared with five oversampling algorithms on 8 microbial datasets. The results show that the proposed algorithm can enhance the diversity of positive samples and identify more positive samples with lower classification cost.

Key words: matrix factorization; cost-sensitive; microbial data; high sparsity; sample subspace; feature subspace

---

基金项目：中央引导地方科技发展专项项目(2021ZYD0003)；西南石油大学启航计划(2018QHR007)。
收稿日期：2022-05-18；**修订日期：**2022-11-22

引 言

人类微生物组研究是通过分析人体内(表)微生物群落的结构变化来了解微生物对人体健康的影响,其已成为疾病检测的诊断工具[1]。这类微生物数据有两个特点,一是正样本少而价值高,二是数据具有高稀疏性。

在微生物不平衡数据中,由于患病样本的数量远少于健康样本的数量,即使将所有患病样本预测为健康样本,仍然可以得到非常高的准确率[2]。但是在现实中,如果将病患判定为健康状态,将会付出巨大的代价[3]。如何识别更多的患病样本,降低不平衡数据分类结果的总代价,是研究微生物数据分类的核心问题之一。

目前改进样本类别不平衡的方法可在分类算法层面、代价敏感层面和数据处理层面上进行。分类算法层面一般通过增加分类器对正样本的重视程度来提高分类器的性能[4]。常用的方法有调整损失函数、改变决策阈值和训练多个基分类器等。代价敏感层面是为正样本的误分类结果分配更高的代价[5-6],但是在某些应用较少的数据集中,它们的代价通常是未知的[7],难以得到验证,为相关应用带来挑战。数据处理层面一般指的是通过重采样技术处理数据,从而实现数据类别的平衡。重采样技术分为过采样[8]和欠采样[9-10],现今使用最广泛的过采样算法是Chawla等[11]提出的SMOTE方法。基于该方法,研究者们提出了许多改进算法,例如Borderline-SMOTE[12]、Safe-Level-SMOTE[13]、V-synth[14]和KMeansSMOTE[15]等方法。这些方法在一定程度上解决了数据不平衡问题,然而,还没有学者研究不平衡数据的高稀疏性问题。

目前流行的过采样方法都是在数据的原始空间中进行扩增,而在微生物数据中,数据集的稀疏性达到90%以上。这导致其他的过采样方法难以捕捉到有效的数据,使得这类方法对稀疏数据的应用效果较差。如何能够抓住稀疏数据的本质,通过所提取的潜在特征进行过采样,是研究不平衡稀疏数据的重点。

本文针对高稀疏性数据,提出了一种融合矩阵分解和代价敏感的数据扩增算法(MFCDA)。该方法由3个阶段组成,分别是矩阵分解、数据扩增和数据过滤。在矩阵分解阶段,将原始数据矩阵分解为样本子空间和特征子空间。矩阵分解技术提取了数据隐藏的k维属性形成了样本子空间,大大减少了数据的稀疏性,使得后续技术在一个稠密的矩阵中进行。在数据扩增阶段,对样本子空间中的正向量进行过采样,找出每个正向量的5个近邻向量,选择其中最近向量、最远向量和原始正向量组成一个封闭区域,在这区域中随机插值,构造合成向量。在数据过滤阶段,根据样本子空间中合成向量与所有负向量的距离进行筛选,将与负向量距离更远的合成向量标记为正。

本文的主要贡献包括3个方面:

(1) 提出了一种融合矩阵分解和代价敏感的数据扩增算法,能够在处理不平衡问题的同时缓解微生物数据的稀疏性;

(2) 在数据扩增过程中引入了代价因子,使得扩增后的数据内部达到代价平衡;

(3) 在不同的微生物数据集中的实验结果表明,提出的MFCDA算法可以正确预测更多的正样本,使得分类结果的总代价最低。

1 相关工作

1.1 矩阵分解

奇异值分解(Singular value decomposition, SVD)[16]是矩阵分解的一个主要分支,通常用于处理密集的数据矩阵。微生物数据极其稀疏,若使用SVD对微生物数据进行分解,首先要对矩阵中的0值进

行填充,这样会导致算法复杂度增加且数据可能失真。由于SVD对稀疏数据作用甚微,BasicSVD[17]、FunkSVD[18]、SVD++[19]等矩阵分解方法被相继提出,不再将矩阵分解为3个矩阵,而是分解为两个低阶矩阵,同时降低了计算复杂度。

1.2　代价敏感学习

代价敏感学习是目前解决不平衡问题的方法之一。由于不同的误分类结果会导致不同的代价,于是代价敏感学习主要考虑在分类模型中如何训练分类器使得最终分类结果代价总和最小[20]。代价敏感学习常用的方法有3种[21],分别是数据预处理、结果后处理和模型中的代价敏感学习。

数据预处理方法包括调整样本分布和样本加权。前者根据误分类代价改变样本分布,后者在不改变数据原始分布的情况下,根据误分类代价为数据集中的样本分配不同的权重。这两种方法都能使数据内的代价达到平衡。

结果后处理方法设置一个代价敏感的决策阈值。当单个样本预测为正样本的概率大于所设阈值时将其预测为正。

模型中的代价敏感学习是将误分类代价与模型的目标函数结合起来,通过最小化损失函数,构造一个基于代价敏感的学习模型。

本文所提出的MFCDA算法通过数据扩增改变数据中的样本分布,从而达到数据内的代价平衡。

2　算法描述与分析

本节将详细介绍基于代价敏感的MFCDA算法。算法由矩阵分解、数据扩增和数据过滤3项技术组成。

2.1　矩阵分解技术

对于微生物数据高稀疏性这一特点,使用矩阵分解技术将其转化为两个低阶矩阵。通过式(1),将原始数据矩阵$\boldsymbol{D}$分解为样本子空间$\boldsymbol{S}$和特征子空间$\boldsymbol{F}$,这样既可以降低数据的空间复杂度,又可以提取隐藏的k维属性。矩阵分解公式[22]为

$$\boldsymbol{D}_{m\times n}=\boldsymbol{S}_{m\times k}\boldsymbol{F}_{n\times k}^{\mathrm{T}} \tag{1}$$

式中:m为原始数据集中的样本数;n为特征数;k为隐含因子的数量,即样本子空间中特征数,$k\in\{1,2,\cdots,n\}$。

通过最小化平均绝对误差(MAE)得到更新后的$\boldsymbol{S}$和$\boldsymbol{F}$,即

$$\min_{S^*,F^*}\sum_{u=1}^{m}\sum_{i=1}^{n}\frac{\left|\boldsymbol{D}_{ui}-\boldsymbol{S}_u\cdot\boldsymbol{F}_i^{\mathrm{T}}\right|}{m\times n} \tag{2}$$

式中:$\boldsymbol{D}_{ui}$为矩阵$\boldsymbol{D}$的第u个样本的第i个特征值,$\boldsymbol{S}_u$为$\boldsymbol{S}$的第u个向量,$\boldsymbol{F}_i$为$\boldsymbol{F}$的第i个向量。

通过随机梯度下降来更新$\boldsymbol{S}_u$和$\boldsymbol{F}_i$,更新公式为

$$\boldsymbol{S}_u=\boldsymbol{S}_u+\alpha\cdot\left[\left(\boldsymbol{D}_{ui}-\boldsymbol{S}_u\cdot\boldsymbol{F}_i^{\mathrm{T}}\right)\cdot\boldsymbol{F}_i-\lambda\cdot\boldsymbol{S}_u\right] \tag{3}$$

$$\boldsymbol{F}_i=\boldsymbol{F}_i+\alpha\cdot\left[\left(\boldsymbol{D}_{ui}-\boldsymbol{S}_u\cdot\boldsymbol{F}_i^{\mathrm{T}}\right)\cdot\boldsymbol{S}_u-\lambda\cdot\boldsymbol{F}_i\right] \tag{4}$$

式中:α为学习率,λ为正则化系数。

2.2　数据扩增技术

对原始矩阵进行分解后,通过式(5)对$\boldsymbol{S}$进行扩增。扩增后的$\boldsymbol{S}$与$\boldsymbol{F}$相乘,得到扩增后的微生物数据集。对于$\boldsymbol{S}$中的每一个正向量$\boldsymbol{S}_i$,通过欧氏距离得到其5个近邻向量。若在这5个近邻向量中随机选择一个与原向量之间进行线性插值,将导致扩增后的样本更密集,无法增加样本的多样性。

本文选择S_i的近邻向量中与其距离最近的$S_{near_{i1}}$和最远的$S_{near_{i5}}$进行扩增，如图1所示。扩增公式为

$$S_{new_a}=\beta\times S_{near_{i1}}+\gamma\times S_{near_{i5}}+(1-\beta-\gamma)\times S_i \tag{5}$$

式中：S_{new_a}为第a个合成向量，β和γ为介于0和0.5之间的随机值。合成向量位于这3个向量构成的空间内部，分布较为稀疏，增加了数据集中正样本的多样性。

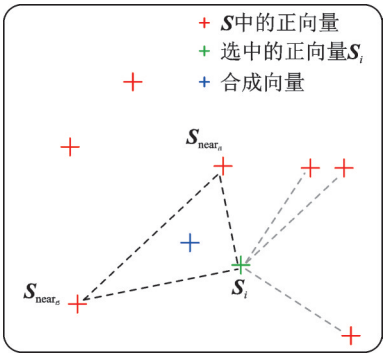

图1 数据扩增过程

Fig.1 Data augmentation process

对于微生物数据集不平衡且正样本价值高的特点，在扩增中融入代价敏感思想。在训练集的扩增倍数中引入代价因子τ，使得数据内部能够达到代价平衡。对于不同的数据集，引入了不同的τ。每个数据集扩增的数量为$1.5\times\tau\times l$，其中l为D中正样本的数量。

2.3 数据过滤技术

决策边界附近的样本容易被错误分类。靠近负向量的正向量与特征子空间相乘所得到的正样本被认为是质量较差的样本，难以正确分类。为了降低该问题造成的影响，提出对扩增后的合成向量进行过滤。首先计算每个合成向量S_{new_a}到所有负向量的距离H_{new_a}，然后对距离进行排序，将距离最大的$\tau\times l$个合成向量标记为正。距离公式为

$$H_{new_a}=\sum_{j=1}^{b}\left|S_{new_a}-S_j\right| \tag{6}$$

式中：S_j为S中第j个负向量，b为S中所有负向量的个数。通过数据过滤技术，能够丰富数据矩阵的样本多样性。

2.4 MFCDA算法

算法1描述了MFCDA算法。第1行是初始化，S和F是随机赋值的两个矩阵。第2～8行是矩阵分解过程，更新S和F两个低阶矩阵，并使他们相乘所得的矩阵接近原始矩阵D。第9～16行进行数据扩增。第17行选择与所有负向量距离最大的$\tau\times l$个合成向量添加到S中，形成S'。第18行将S'和F相乘得到扩增后的数据矩阵D'。第19行返回扩增后的数据矩阵D'。

算法1 MFCDA算法

输入：原始数据集D，学习率α，正则化参数λ，代价因子τ，隐含因子k，正样本数量l。

输出：扩增后的数据集D'。

① Initialize S, F, MAE_{former}, $MAE_{current}$, H, a, l;/*初始化*/

② $S, F\leftarrow$updateMatrix(S, F, α, λ, k);/*式(3)，式(4)*/

③ $MAE_{current}\leftarrow$computeError(D, S, F);/*式(2)*/

④ While($MAE_{former}>MAE_{current}$)do

⑤ $MAE_{former}\leftarrow MAE_{current}$;

⑥ $S, F\leftarrow$updateMatrix(S, F, α, λ, k);/*式(3)，式(4)*/

⑦ $MAE_{current}\leftarrow$computeError(D, S, F);/*式(2)*/

⑧ end while

⑨ while($a<1.5\times\tau\times l$)do

⑩ for ($i=1$ to l) do

⑪ $S_{new_a} \leftarrow$ generateVector(S_i);/*式(5)*/

⑫ $H_{new_a} \leftarrow$ computeDistance(S_{new_a});/*式(6)*/

⑬ $H \leftarrow H \cup H_{new_a}$;

⑭ $a \leftarrow a+1$;

⑮ end for

⑯ end while

⑰ $S' \leftarrow$ selectVector(H);

⑱ $D' \leftarrow$ computeMatrix(S', F);/*式(1)*/

⑲ return D'

3　实验与结果

本节将评估MFCDA方法的有效性。将MFCDA与ROS、SMOTE、Borderline-SMOTE1、Borderline-SMOTE2和ADASYN五种算法在3个分类器下的分类效果进行比较。实验结果表明,MFCDA能够正确识别更多的正样本,得到更低的分类代价。

3.1　实验数据集

在8个微生物数据集上进行实验,以验证MFCDA算法的有效性。数据集来自mAML网站(http://39.100.246.211:8050/Dataset)。使用的8个数据集具体信息见表1。

表1　不平衡微生物数据集

Table 1　Imbalanced microbial datasets

数据集	特征数	少数类	多数类	不平衡比	稀疏度/%
D001289	142	162	1 170	7.22	94
D001327	153	462	1 170	2.53	94
D002318	140	116	1 170	10.09	94
D002446	137	58	1 170	20.17	93
D003863	145	228	1 170	5.13	94
D003920	148	90	1 180	13.11	94
D007674	136	59	1 170	19.83	93
D008107	136	62	1 170	18.87	93

3.2　比较方法

将MFCDA与5种传统的过采样算法进行比较,以说明所提出方法的优越性能。

比较方法的详细信息介绍如下:

(1)ROS是一种随机复制和重复少数类样本的方法。它使少数类和多数类的数量相同,从而产生了一个新的平衡数据集。

(2)SMOTE算法的基本思想是对少数样本进行分析,并在少数样本的基础上人工合成新的样本。它随机选择两个最近邻少数样本的直线上的一点作为新的正样本。

(3)ADASYN根据少数数据样本的分布自适应地生成新样本,不仅可以减少由于原始数据分布不平衡而导致的学习偏差,而且可以将决策边界自适应地转移到难以学习的样本上。

(4)Borderline-SMOTE将少数类样本分为3类,即安全、危险和噪声。仅对属于危险的样本进行过采样。Borderline-SMOTE1随机选择危险样本附近的正样本生成新样本。Borderline-SMOTE2选择k最近邻中的任何样本来生成新样本,并不考虑所选近邻样本的类别。

3.3　评价指标

在二分类问题中,通常使用混淆矩阵来评估分类算法的性能。在混淆矩阵中,正确预测的正样本数为TP,正确预测的负样本数为TN,预测为正的负样本数为FP,预测为负的正样本数为FN。通过这4项指标,能够得到Accuracy、Precision、Recall和TNR的值,计算公式分别为

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \tag{7}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{8}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{9}$$

$$\text{TNR}=\frac{\text{TN}}{\text{TN}+\text{FP}} \tag{10}$$

由上述公式可以得到F_β分数,计算公式为

$$F_\beta=(1+\beta^2)\times\frac{\text{Precision}\times\text{Recall}}{(\beta^2\times\text{Precision})+\text{Recall}} \tag{11}$$

式中β用来调节Precision和Recall的相对重要性。

Cost是分类结果的总代价,计算公式为

$$\text{Cost}=\tau\times\text{FN}+\text{FP} \tag{12}$$

由于微生物数据集不平衡,在不对数据集做任何处理的情况下,分类结果将所有的正样本预测为负样本,也能够得到高Accuracy,这是不合理的。Precision表示在所有预测为正例的样本中预测正确的比率。Recall表示在所有实际为正例的样本中预测正确的比率。TNR表示在所有实际为负例的样本中预测正确的比率。F_β是Precision和Recall的加权平均,当$\beta=1$时,认为Precision和Recall同样重要。但是在微生物数据集中,正确识别患病样本更为重要,Recall权重应高于Precision,取$\beta=2$,使用F_2分数。Cost用来计算分类结果的总代价,是最为重要的评价指标。本文采用Precision、Recall、F_2和Cost这4个指标来评估算法的分类性能。

3.4 参数设置

实验代价因子设置见表2,其余的参数设置见表3。本实验将数据集按照7∶3的比例分为训练集和测试集。

表2 代价因子设置

Table 2 Cost factor setting

数据集	τ	数据集	τ
D001289	10	D003863	10
D001327	5	D003920	20
D002318	20	D007674	30
D002446	25	D008107	30

表3 实验参数

Table 3 Experimental parameters

参数	参数值
隐含因子k	10
学习率α	0.000 1
正则化参数λ	0.005

3.5 实验结果与分析

本节将给出不同过采样算法的实验结果以及消融实验的结果。

3.5.1 不同过采样算法的数据分布图

选取不平衡比为2.53、20.17和19.83的微生物数据集D001327、D002446和D007674,对原始数据和利用ROS、SMOTE、Borderline-SMOTE1、Borderline-SMOTE2、ADASYN以及MFCDA进行过采样后的扩增数据使用t-SNE将其降维并绘出分布图,分别如图2~4所示。

图2展示了D001327原始数据的分布图以及使用不同过采样方法后的数据扩增图。从图中可以看出,D001327中的正样本略少于负样本,使用ROS扩增只是简单地复制粘贴,数据的分布没有发生变化,多样性没有增加。使用SMOTE、Borderline-SMOTE1、Borderline-SMOTE2和ADASYN方法扩增后,在原始样本附近生成了新样本,一定程度上增加了正样本的多样性,但正负样本仍然重叠。使用MFCDA方法扩增后,正样本的分布具有一定的规律,但仍难与负样本区分。

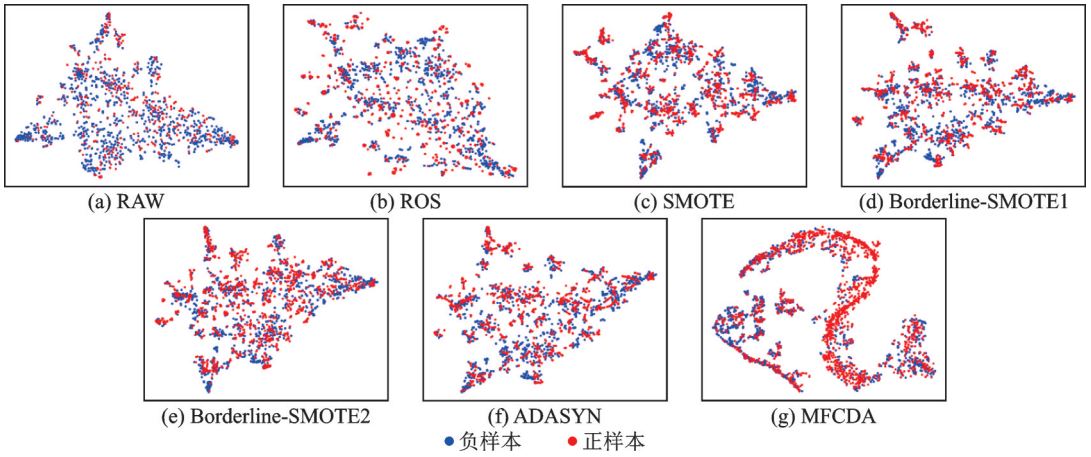

图 2　D001327 使用不同过采样方法的数据分布图

Fig.2　Data distribution diagrams for D001327 using different oversampling methods

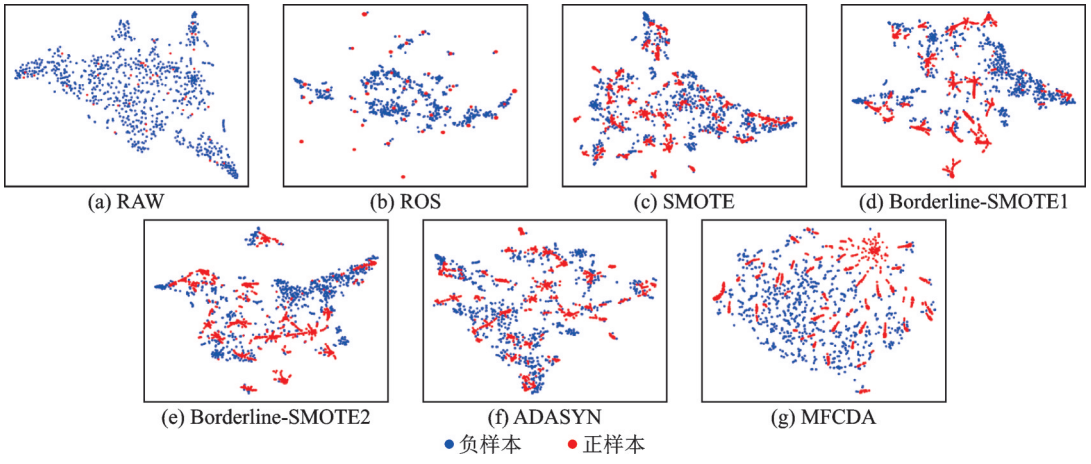

图 3　D002446 使用不同过采样方法的数据分布图

Fig.3　Data distribution diagrams for D002446 using different oversampling methods

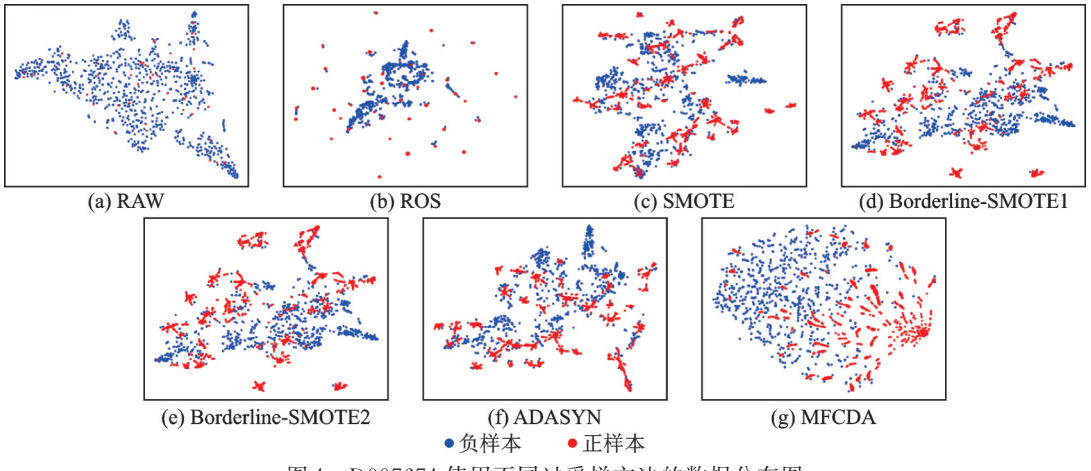

图 4　D007674 使用不同过采样方法的数据分布图

Fig.4　Data distribution diagrams for D007674 using different oversampling methods

图3、4分别展示了D002446和D007674数据集使用不同过采样方法后的数据扩增图。从图中可以看出，使用SMOTE、Borderline-SMOTE1、Borderline-SMOTE2和ADASYN方法扩增后，合成样本的分布有一定的规律，与原始正样本联系紧密，但是与负样本仍然难以区分。使用MFCDA方法扩增后，极大丰富了正样本的多样性，可以很好地与负样本区分开。

从分布图对比结果可以看出，MFCDA扩增的正样本具有明显的类间分离，并且随着数据集不平衡比的增加，正负样本之间的区分度提高。这表明MFCDA更适合于具有高不平衡比的数据集。

3.5.2 不同过采样方法后的数据稀疏度

统计了使用不同过采样方法后的数据稀疏度，如图5所示。从图中可以看出，其他的过采样方法轻微地减少了原始数据的稀疏度，但MFCDA能够有效地对数据进行稠密化。这是由于其他过采样方法都是在原始数据空间进行处理，而MFCDA是通过提取的隐藏特征进行扩增，所以大大降低了数据的稀疏性。结合数据分布图来看，数据的高稀疏度使得正负样本区分度降低，而降稀疏度过程能够提升样本的类间区分性。

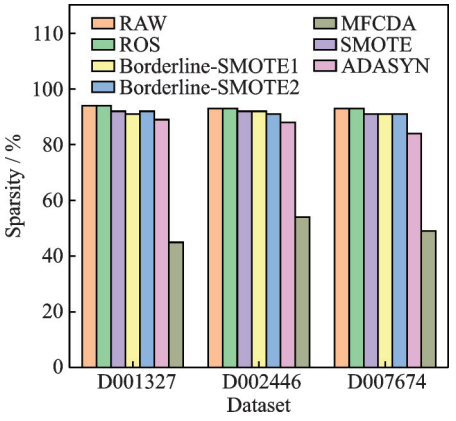

图5 不同过采样方法扩增后的数据稀疏度

Fig.5 Data sparsity augmented by different oversampling methods

3.5.3 不同过采样方法的指标对比

使用了decision tree、random forest和GBDT三个分类器验证不同扩增算法的有效性。在相同的分类器下，将在原始数据集上训练的模型性能与在基于ROS、SMOTE、Borderline-SMOTE1、Borderline-SMOTE2、ADASYN和MFCDA方法扩增后的数据集上训练的模型性能进行比较。每个实验重复10次记录平均值，以减轻随机性对结果的影响。

图6～8分别给出了不同方法在各个数据集上的平均指标值。图6展示了在decision tree分类器下的实验结果。在Recall评价指标下，MFCDA在所有数据集中均优于其余5种过采样方法。在F_2评价指标下，MFCDA在8个测试集中的7个数据集上优于其他过采样方法，在D008107数据集上排名第二。在Cost评价指标下，MFCDA在6个数据集上优于其他方法。MFCDA在Precision评价指标下表现相对较差，仅在D002446数据集上表现最好。

图7展示了在random forest分类器下的实验结果。在Recall、F_2和Cost评价指标下，MFCDA在所有数据集中均优于其余5种过采样方法。MFCDA在Precision评价指标下表现相对较差。

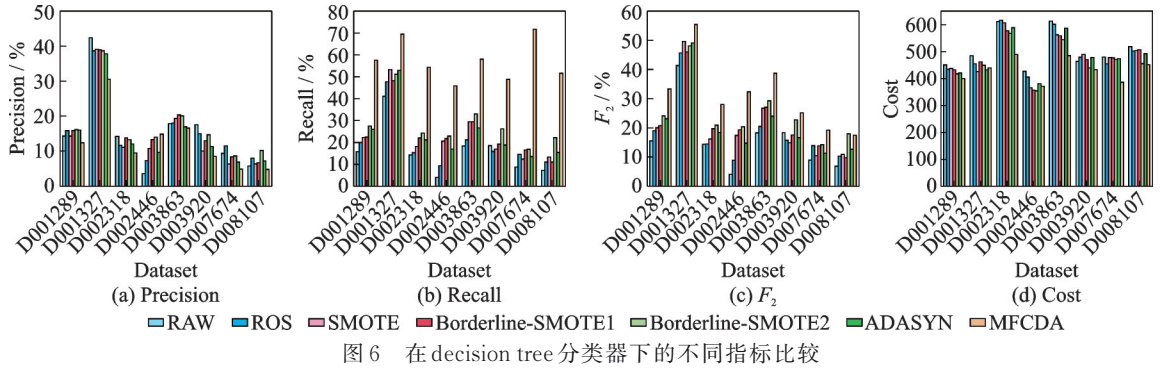

图6 在decision tree分类器下的不同指标比较

Fig.6 Comparison of different indicators under decision tree classifier

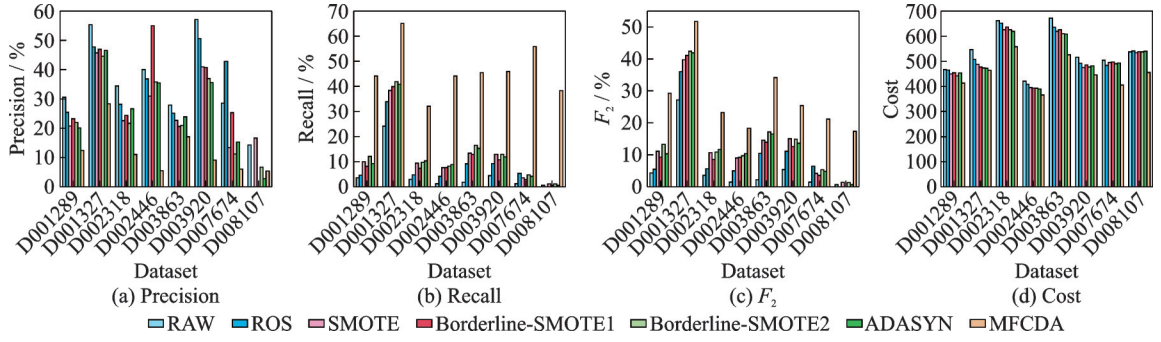

图7　在random forest分类器下的不同指标比较

Fig.7　Comparison of different indicators under random forest classifier

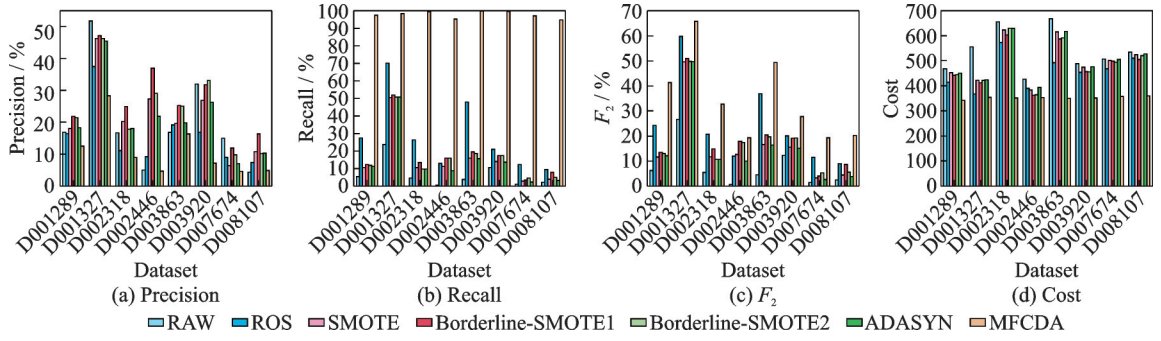

图8　在GBDT分类器下的不同指标比较

Fig.8　Comparison of different indicators under GBDT classifier

图8展示了在GBDT分类器下的实验结果。在Recall、F_2和Cost评价指标下,MFCDA在所有数据集中均优于其余5种过采样方法。MFCDA在Precision评价指标下表现相对较差。但是在GBDT分类器下,大部分数据集的Recall评价指标高于0.95,因此认为该分类器将MFCDA算法扩增后的绝大部分样本都预测为正类,适用于目的是预测正确更多正样本的领域。

3.5.4　消融实验

为证明代价敏感和数据过滤的有效性,在3个数据集上进行了消融实验,实验结果如表4所示。表中,Non是不处理的原始数据集。Cost factor+Data filtering是不经过矩阵分解,根据代价因子控制扩增数量并进行数据过滤。MF+Cost factor是在矩阵分解后仅根据代价因子控制扩增数量而不进行数据过滤。MF+Data filtering采用了数据过滤技术,扩增至正负样本同一数量。MFCDA即在矩阵分解后,通过代价因子控制扩增数量,又对扩增的数据进行过滤。从表4的实验结果可知:

(1)相较于原始数据集,MF+Cost factor和MF+Data filtering分类性能有所提升,都能够识别出更多的正样本,获得更低的分类结果代价。Cost factor+Data filtering的分类效果表现较差,说明了矩阵分解过程使实验结果有所改进,高稀疏性被解决是MFCDA分类性能提升的重要原因。

(2)MF+Data filtering模块的分类效果优于MF+Cost factor模块,这是由于MF+Data filtering模块能够增加扩增样本的多样性,过滤掉决策边界附近的合成样本。

(3)MFCDA模型取得了最好的效果,原因在于根据矩阵分解后的样本子空间进行扩增,可有效地

表4 消融实验

Table 4 Ablation experiment

数据集	模型	稀疏度/%	Precision	Recall	F_2	Cost
D001327	Non	94	0.50	0.30	0.32	529
	Cost factor+Data filtering	92	0.16	0.40	0.20	528
	MF+Cost factor	46	0.27	0.57	0.46	505
	MF+Data filtering	46	0.31	0.52	0.46	490
	MFCDA	45	0.29	0.78	0.58	419
D002446	Non	93	0.16	0.02	0.02	424
	Cost factor+ Data filtering	91	—	0.00	—	425
	MF+Cost factor	55	0.04	0.02	0.05	425
	MF+Data filtering	56	0.08	0.04	0.11	416
	MFCDA	54	0.08	0.62	0.23	364
D007674	Non	93	0.18	0.04	0.04	497
	Cost factor+Data filtering	90	0.29	0.04	0.07	492
	MF+Cost factor	50	0.08	0.33	0.15	437
	MF+Data filtering	49	0.07	0.43	0.14	435
	MFCDA	49	0.05	0.75	0.20	383

降低数据的高稀疏性。通过代价因子控制合成样本的数量，使得数据集内部能够达到更为合理的代价平衡，而不是单纯的数量平衡。最后对合成向量进行过滤，在保证样本多样性的同时避免了合成噪声样本对最终分类结果的干扰。

3.5.5 超参数k的敏感性

k是本实验中关键的超参数，为了找到合适的k值，在D001327数据集中对3个分类器的实验结果取平均值，$k\in\{5, 10, \cdots, 50\}$，如图9所示。实验结果证明，在decision tree和random forest下，当k取值从5到30时，Cost的变化较为稳定，在小范围内浮动。因此，将k设置在这个范围内能获得更好的效果。在GBDT分类器下，不同的k对分类性能影响较大，在k=10时Cost达到最小值，所以将k设置为10会得到更好的性能。

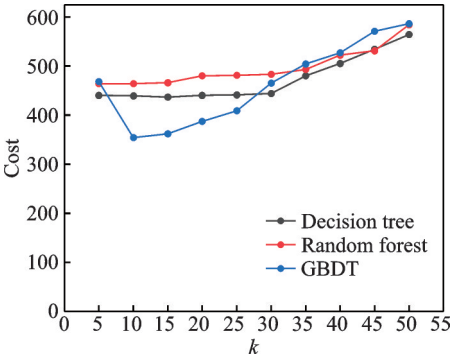

图9 不同的k对Cost的影响

Fig.9 Effect of different k on Cost

从实验对比结果可以看出，相比传统的数据过采样方法，本文提出的MFCDA算法在Recall、F_2和Cost指标上都具有明显的优势。传统的过采样方法在这3个指标上表现不佳是因为扩增的新样本仍然具有较强的稀疏性，样本的多样性不够丰富。MFCDA是在矩阵分解后的样本子空间内进行扩增，更好地把握住了数据的本质，在一定程度上缓解了微生物数据的高稀疏性。同时通过代价因子来控制合成样本的数量，让数据集内部达到了更为合理的代价平衡。数据过滤技术还可以增强样本的多样性，减少噪声样本对分类结果的干扰。

4 结束语

本文针对微生物数据，提出了一种融合矩阵分解和代价敏感的数据扩增算法，即MFCDA算法。

该算法针对微生物数据的不平衡和高稀疏性,对矩阵分解后的对象子空间进行扩增,在数据内部达到代价平衡的同时缓解稀疏性。设计了数据过滤技术,移除位于决策边界附近的样本,避免了此类噪声样本对分类结果的干扰。实验结果表明,MFCDA算法在不平衡分类性能方面表现优异,能有效地降低分类结果的总代价;在3个分类器下,MFCDA算法相比其他5种过采样方法,具有较好的分类性能。

参考文献:

[1] LAPIERRE N, JU C J T, ZHOU G, et al. MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction[J]. Methods, 2019, 166: 74-82.

[2] ABD ELRAHMAN S M, ABRAHAM A. A review of class imbalance problem[J]. Journal of Network and Innovative Computing, 2013, 1: 332-340.

[3] HAI T N, BAO T, QUAN M, et al. Enhancing disease prediction on imbalanced metagenomic dataset by cost-sensitive[J]. International Journal of Advanced Computer Science and Applications, 2020, 11(7): 651-657.

[4] 杨平安, 林亚平, 祝团飞. AdaBoostRS: 高维不平衡数据学习的集成整合[J]. 计算机科学, 2019, 46(12): 8-12.
YANG Ping'an, LIN Yaping, ZHU Tuanfei. AdaBoostRS: Ensemble integration for high-dimensional imbalanced data learning[J]. Computer Science, 2019, 46(12): 8-12.

[5] KAUR P , GOSAIN A. Empirical assessment of ensemble based approaches to classify imbalanced data in binary classification [J]. International Journal of Advanced Computer Science and Applications, 2019, 10(3): 48-58.

[6] DHAR S, CHERKASSKY V. Development and evaluation of cost-sensitive universum-SVM[J]. IEEE Transactions on Cybernetics, 2014, 45(4): 806-818.

[7] GALAR M, FERNANDEZ A, BARRENECHEA E, et al. A review on ensembles for the class imbalance problem: Bagging- , boosting- , and hybrid-based approaches[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2011, 42(4): 463-484.

[8] KOTSIANTIS S, KANELLOPOULOS D, PINTELAS P. Handling imbalanced datasets: A review[J]. GESTS International Transactions on Computer Science and Engineering, 2006, 30(1): 25-36.

[9] LIU X Y, WU J, ZHOU Z H. Exploratory undersampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2008, 39(2): 539-550.

[10] TOMEK I. Two modifications of CNN[J]. IEEE Transactions on Systems, Man and Cybernetics, 1976, 6: 769-772.

[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[C]// Proceedings of International Conference on Intelligent Computing.Berlin: Springer, 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[C]//Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2009: 475-482.

[14] YOUNG W A, NYKL S L, WECKMAN G R, et al. Using Voronoi diagrams to improve classification performances when modeling imbalanced datasets[J]. Neural Computing and Applications, 2015, 26(5): 1041-1054.

[15] LAST F, DOUZAS G, BACAO F. Oversampling for imbalanced learning based on k-means and smote[J]. arXiv preprint arXiv:1711.00837, 2017.

[16] HOECKER A, KARTVELISHVILI V. SVD approach to data unfolding[J]. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 1996, 372(3): 469-481.

[17] LI W, LIANG Z, CAO Y, et al. Estimating intrafraction tumor motion during fiducial-based liver stereotactic radiotherapy via

an iterative closest point (ICP) algorithm[J]. Radiation Oncology, 2019, 14(1): 1-8.

[18] YUE X, LIU Q. Parallel algorithm of improved FunkSVD based on spark[J]. KSII Transactions on Internet and Information Systems (TIIS), 2021, 15(5): 1649-1665.

[19] KUMAR R, VERMA B K, RASTOGI S S. Social popularity based SVD++ recommender system[J]. International Journal of Computer Applications, 2014, 87(14): 33-36.

[20] 万建武, 杨明. 代价敏感学习方法综述[J]. 软件学报, 2020, 31(1): 113-136.
WAN Jianwu, YANG Ming. A review of cost-sensitive learning methods[J]. Journal of Software, 2020, 31(1): 113-136.

[21] SHENG V S, LING C X. Thresholding for making classifiers cost-sensitive[C]//Proceedings of the 21st AAAI Conference on Artificial Intelligence. Boston:AAAI, 2006: 476-481.

[22] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30-37.

作者简介:

王曦(1997-),女,硕士研究生,研究方向:代价敏感学习、不平衡学习等,E-mail:wangxiii1997@163.com。

温柳英(1983-),**通信作者**,女,副教授,研究方向:机器学习、不平衡学习、微生物信息学等,E-mail:Wenliuying1983@163.com。

闵帆(1973-),男,教授,研究方向:粒计算、主动学习、推荐系统、多标签学习等。

(编辑:夏道家)