文章编号：2095-2163(2020)11-0031-07 中图分类号：TP391 文献标志码：A

针对缺失不平衡数据的自适应马氏距离双权重过采样方法研究

周迪豪，宋 燕
（上海理工大学 光电信息与计算机工程学院，上海 200093）

摘 要：具有缺失信息的不平衡数据，是如今分类问题面临的一个巨大挑战。针对此问题，本文提出一种基于马氏距离的自适应双权重过采样技术（Adaptive Double-weighted Mahalanobis Oversampling Technique，MAWOTE）。MAWOTE 的主要思想是：(1)考虑到全局特征信息中更大的最优解空间，提出了一种基于小批量梯度下降（Mini-Batch Gradient Descent，MBGD）规则的非负潜在因子矩阵分解方法（Non-negative Latent Factor Matrix Factorizations，NLFs），对缺失信息进行填补，并使其满足原始数据分布；(2)引入马氏距离作为距离度量，统一样本特征量纲；(3)提出一种自适应的双权重分配方法，有效提高了合成少类样本的安全性和可靠性；(4)为了保持样本的原始信息分布，利用 k 近邻思想进行过采样插值。最后，在 6 个不同缺失率的公共数据集上进行对比实验，实验结果表明本文提出的方法在处理具有缺失值的不平衡数据集的分类问题时，明显优于其它先进算法。

关键词：分类；不平衡数据；缺失值；潜在因子；马氏距离；双权重；过采样

An adaptive double-weighted oversampling technique based on Mahalanobis distance for imbalanced classes with missing values

ZHOU Dihao, SONG Yan
(School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 Imbalanced problem with missing values is a huge challenge in the classification task. However, most available techniques now to deal with the imbalanced problems are based on the completed datasets. To solve this problem, an adaptive double-weighted oversampling technique based on Mahalanobis distance for imbalanced classes with missing values (MAWOTE) is proposed in this paper. The main idea of MAWOTE is 4-fold as follows: (1) considering the larger optimal solution space in the information of global features, an non-negative latent factor matrix factorizations method (NLFs) with MBGD rules is put forward to impute missing values precisely according to the original distribution; (2) Mahalanobis distance is introduced to be a measure which can eliminate the different standard among features; (3) an adaptive double-weighted assignment strategy is proposed to synthesize the new samples more safely; (4) in order to maintain the original distribution information of dataset, k-nearest neighbor idea is introduced in oversampling process. Finally, experiments on six public datasets with different missing rates show that the proposed method is capable of dealing the imbalanced datasets with missing values and out-performs other 5 advanced algorithms.

【Key words】 Classification; imbalanced data; missing values; latent factor; Mahala Nobis distance; double-weighted; oversampling

0 引 言

分类问题是模式识别等众多计算机领域的一个重要分支。分类问题往往存在类别不平衡现象。在不平衡情况下，若又存在信息缺失现象，则分类器性能会受到巨大影响[1]。在现实世界中，由于受数据复杂性及测量设备局限性等影响，数据不平衡和数据缺失现象普遍存在。在二类不平衡数据集中，样本较少的类称为少类，样本较多的类称为多类，若直接对数据集进行分类，则少类样本容易被忽略，而多类样本会被过度关注，从而导致分类器向多类倾斜[2]。而少类样本通常包含有重要信息，例如在诊断样本中，少类样本的误分类可能会造成严重的危害[3]。因此，解决缺失数据的不平衡问题就显得尤为重要。

通常，处理不平衡类的方法主要有两种：基于模型（算法）方法和基于数据方法[4]。基于模型方法，如 cost-sensitive boosting 算法[5]考虑了错误分类的代价，而非类与类之间的关系。然而，当发生类别分

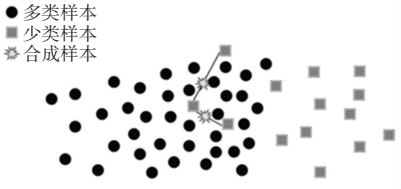
基金项目：上海市自然科学基金（18ZR1427100）。

作者简介：周迪豪（1996-），男，硕士研究生，主要研究方向：大数据模型应用；宋 燕（1979-），女，博士，副教授，博士生导师，主要研究方向：大数据算法、图像处理、预测控制。

通讯作者：宋 燕 Email：sonya@usst.edu.cn

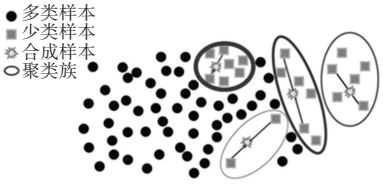
收稿日期：2020-10-09

布重叠时,模型的性能可能会大大下降[6]。基于数据方法的主要目标,是利用采样技术调整样本数,从而达到类别平衡。采样方法大致可分为欠采样和过采样[7]。前者是删除多类中多余的样本,而后者则是为少类合成新样本。Tomek links[8]是最具代表性的欠采样算法,它可以消除在类边界附近的噪声样本。然而,由于样本的删除,欠采样可能导致数据集关键信息的丢失。相反,通过合成新的少类样本来平衡类别规模,过采样方法得以保留大部分数据集信息[3]。目前,人们在过采样方面已经做了大量的工作并体现在文献中。如 SMOTE、ADASYN、K-means SMOTE 和 MWMOTE[9]。SMOTE 针对少类合成新样本,但在边界处易产生错误样本,如图 1(a)所示。



(a) MOTE 示意图

(a) MOTE schematic diagram



(b)K-means SMOTE 示意图

(b) Schematic diagram of K-means SMOTE

图 1 SMOTE 与 K-means SMOTE 合成策略示意图

Fig. 1 The schematic diagram of SMOTE and K-means SMOTE

ADASYN 在 SMOTE 的基础上考虑加强了边界点权重,分类超平面更显著。K-means SMOTE 利用 K 均值聚类形成簇,并分配簇权重以合成新样本,如图 1(b)所示。图中的椭圆轮廓越粗代表该簇的权重分配得越大。MWMOTE 不仅考虑了少类信息,其将多类信息也纳入考量范围,并作为分配少类样本权重的依据。尽管上述方法考虑到了类别分布,但却没有充分考虑数据少类的类内分布。

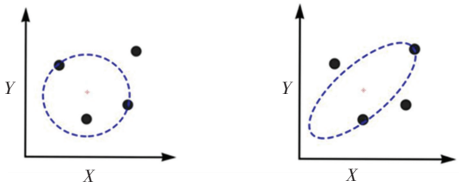
另一方面,数据缺失又是不平衡数据分类中另一大困难,数据丢失可能会严重影响模型的性能,因此,在分类前对缺失值进行填补是非常重要的。目前,缺失值插补的代表性研究成果包括:回归插补(RI)、经验最大化(EM)、多重插补(MI)和 k-NN 插补法[10]。此外,文献[11]提出一种基于非负潜在因子的矩阵分解(LMF)模型,用于填补缺失值的方法。该方法能够更好地利用样本的全局信息,改善了回填精度。文献[12]表明,传统过采样算法在回填缺失值后通常不能保证完全符合原始样本分布,而传统过采样方法的插值公式使得合成样本的不确定性进一步加大。对此,一种基于模糊的信息分解(FID)[13]被提出,该算法能够同时解决不平衡问题和信息缺失问题。然而,由于模糊隶属度的存在,FID 对局部信息的依赖程度很高,可能导致过拟合现象。因此,探究一种新的算法来有效地处理带有缺失值的不平衡数据集成为一个急需解决的问题。根据文献[14]的启发,马氏距离能够有效解决样本集中普遍存在特征量纲不一的问题。

本文根据上述难题,提出一种针对含有缺失信息的不平衡数据的自适应马氏距离多权重过采样方法(MAWOTE)。该方法首先通过基于 MGBD 更新规则的 NLFs 模型,将不完整的数据集填补完整,其次基于马氏距离将数据集中样本使用模糊 C 均值聚类(FCM),随后对各样本簇自适应分配双权重,并根据 k 近邻信息进行过采样,最后使用 SVM,验证算法的有效性。

1 相关技术

1.1 马氏距离

本文所提出的过采样算法是依据马氏距离[14]。马氏距离在计算时考虑了数据集的相关关系,与欧氏距离不同的是它考虑到各种特性之间的联系。如图 2 所示,在欧式距离中,A 点与 B 点到原点的距离相同,然而在马氏距离中,A 点与 B 点到原点的距离则不相同,因为马氏距离表示的是数据的协方差距离。



(a) 欧式距离 (b) 马氏距离

(a) Euclidean distance (b) Mahalanobis distance

图 2 欧式距离与马氏距离示意图

Fig. 2 Schematic diagram of Euclidean distance and Mahalanobis distance

假定样本向量 $x=(x_1,x_2,\cdots,x_n)^T$ 和 $y=(y_1,y_2,\cdots,y_n)^T$,两点间欧式距离为:

$$D_E(x)=\sqrt{(x-y)^T(x-y)}. \tag{1}$$

样本 x 到样本均值 μ 的马氏距离表示为:

$$D(x)=\sqrt{(x-\mu)^T S^{-1}(x-\mu)}. \tag{2}$$

在此，可以考虑一个协方差矩阵 S，协方差计算如式(3)所示：

$$S(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}. \tag{3}$$

因此，马氏距离可以通过式(4)求解：

$$D(x)=\sqrt{(x-y)^T S^{-1}(x-y)}. \tag{4}$$

当协方差矩阵为单位矩阵时，马氏距离可等同于欧氏距离。

1.2　模糊 C 均值聚类

FCM 算法在 K-means 算法的基础上增加了模糊思想。不同于 K-means 的硬聚类，FCM 提供了一个概率性的灵活簇划分结果。FCM 通过最小化目标函数将数据集划分为 c 个簇：

$$J(\mathbb{U},c_1,c_2,\cdots,c_c)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\|x_j-c_i\|^2. \tag{5}$$

其中，$x_j(j=1,2,\cdots,n)$ 表示数据集 $\{x_1,x_2,\cdots,x_j\}$ 中第 j 个样本；$\mathbb{U}\triangleq\{u_{ij}\}(1=1,2,\cdots,c;j=1,2\cdots,n)$ 为隶属度矩阵；元素 u_{ij} 代表样本 x_j 的隶属度；参数 $m>1$ 表示模糊因子。

基于式(5)，c_i 和 u_{ij} 的更新公式为：

$$c_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j}{\sum_{j=1}^{n}u_{ij}^{m}}, \tag{6}$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_j-c_i\|}{\|x_j-c_k\|}\right)^{\frac{2}{m-1}}}. \tag{7}$$

FCM 聚类算法交替更新聚类中心 c_i 和隶属度矩阵 $\mathbb{U}$，直至收敛或迭代次数达到设定值。

2　缺失值填补及过采样方法

为了使缺失的数据能够得到精确填补，并使过采样后新生成的样本即可靠又重要，本文使用小批量梯度下降法求解 NLFs 后，在马氏距离度量的基础上，对少类样本簇使用自适应的双权重分配策略进行过采样，最终得到一个平衡数据集。基于此，本文提出一种针对缺失不平衡数据的自适应马氏距离双权重过采样方法(MAWOTE)。

2.1　非负矩阵分解插补方法

为了在解决缺失值现象的同时，避免过拟合现象。根据非负矩阵分解(NMF)模型[11]，假设存在一个含有 n 个样本 m 个特征的数据集矩阵 $R_{n\times m}$，则定义存在 3 个矩阵能够构成 $R_{n\times m}$，即 NLFs 模型：

$$R=(X+Y)Z. \tag{8}$$

与文献[11]相似，根据上式可构造用以最小化的损失函数：

$$\arg\min_{X,Y,Z} f(X,Y,Z)=\frac{1}{2}\sum_{(i,j)\in R_k}\left(r_{ij}-\sum_{k=1}^{d}(x_{i,k}+y_{i,k})z_{kj}\right)^2+\lambda_X\sum_{k=1}^{d}x_{i,k}{}^2+\lambda_Y\sum_{k=1}^{d}y_{i,k}{}^2+\lambda_Z\sum_{k=1}^{d}z_{kj}{}^2$$
$$\text{s.t. } x_{i,k},y_{i,k},z_{k,j}\geqslant 0. \tag{9}$$

式中，$r_{i,j}$、$x_{i,k}$、$y_{i,k}$ 和 $z_{k,j}$ 分别对应 R、X、Y 和 Z 中的项，d 为潜在因子数。为了避免过度拟合，在模型中考虑了正则化项，λ_X、λ_Y、λ_Z 是非负正则化系数。第三矩阵使得模型拥有了更多的解空间，以此可得到更好的精度[16]。

通常情况下，批梯度下降(BGD)算法，是解决最小化问题时运用最广泛的更新规则之一[17]。然而，由于每次迭代都需要使用所有样本来寻求最优解，会造成很大的计算开销。因此，在 BGD 的基础上，随机梯度下降法(SGD)[18]被提出。SGD 每次迭代时只更新一个样本，以加快训练速度。然而，SGD 可能导致算法收敛于局部最优。相比之下，小批量梯度下降(MBGD)算法[19]中和了 BGD 算法和 SGD 算法的优点，即每次训练时从所有样本中随机选择一定数量的样本用以更新。基于 MBGD 的更新规则，矩阵 X、Y 和 Z 的更新规则可用式(10) - 式(12)表示：

$$x_{i,k}\leftarrow x_{i,k}+\eta_{i,k}\Big(\sum_{j\in R(i)}z_{k,j}\Big(r_{i,j}-\sum_{k=1}^{d}(x_{i,k}+y_{i,k})z_{k,j}\Big)-2\lambda_X x_{i,k}\Big), \tag{10}$$

$$y_{i,k}\leftarrow y_{i,k}+\gamma_{i,k}\Big(\sum_{j\in R(i)}z_{k,j}\Big(r_{i,j}-\sum_{k=1}^{d}(x_{i,k}+y_{i,k})z_{k,j}\Big)-2\lambda_Y y_{i,k}\Big), \tag{11}$$

$$z_{k,j}\leftarrow z_{k,j}+\eta_{k,j}\Big(\sum_{i\in R(i)}(x_{i,k}+y_{i,k})\Big(r_{i,j}-\sum_{k=1}^{d}(x_{i,k}+y_{i,k})z_{k,j}\Big)-2\lambda_Z z_{k,j}\Big). \tag{12}$$

考虑到在更新过程中学习率可能为负，为了保证样本矩阵的非负性，因此设置：

$$\eta_{i,k}=\frac{x_{i,k}}{\sum_{j\in R(i)}\left(z_{k,j}\sum_{k=1}^{d}(x_{i,k}+y_{i,k})z_{k,j}+2\lambda_X x_{i,k}\right)}, \tag{13}$$

$$\gamma_{i,k}=\frac{y_{i,k}}{\sum_{j\in R(i)}\left(z_{k,j}\sum_{k=1}^{d}(x_{i,k}+y_{i,k})z_{k,j}+2\lambda_Y y_{i,k}\right)}, \tag{14}$$

$$\eta_{k,j}=\frac{z_{k,j}}{\sum_{i\in R(i)}\left((x_{i,k}+y_{i,k})\sum_{k=1}^{d}(x_{i,k}+y_{i,k})z_{k,j}+2\lambda_Z z_{k,j}\right)}. \tag{15}$$

基于此,更新公式可重写为:

$$x_{i,k}\leftarrow\frac{x_{i,k}\sum_{j\in R(i)}z_{k,j}r_{i,j}}{\sum_{j\in R(i)}\left(z_{k,j}\sum_{k=1}^{d}(x_{i,k}+y_{i,k})z_{k,j}+2\lambda_X x_{i,k}\right)}, \tag{16}$$

$$y_{i,k}\leftarrow\frac{y_{i,k}\sum_{j\in R(i)}z_{k,j}r_{i,j}}{\sum_{j\in R(i)}\left(z_{k,j}\sum_{k=1}^{d}(x_{i,k}+y_{i,k})z_{k,j}+2\lambda_Y y_{i,k}\right)}, \tag{17}$$

$$z_{k,j}\leftarrow\frac{z_{k,j}\sum_{i\in R(i)}(x_{i,k}+y_{i,k})r_{i,j}}{\sum_{i\in R(i)}\left((x_{i,k}+y_{i,k})\sum_{k=1}^{d}(x_{i,k}+y_{i,k})z_{k,j}+2\lambda_Z z_{k,j}\right)}. \tag{18}$$

上述公式提供了基于 MBGD 的 NLFs 更新规则。通过 MGBD 更新迭代后,所有缺失值将通过 NLFs 填补,并得到与原始数据集相似的信息,有助于后续过采样时对分布信息的利用。

2.2 样本聚类

传统的随机插值过采样生成新样本的方法,容易导致错误样本的产生。即该类样本位于多类样本中,从而导致新的噪声点产生。MAWOTE 基于马氏距离,采用 FCM 算法对少类样本以及多类样本分别进行聚类,在簇中过采样可保证新样本的可靠性。聚类后形成的少类样本簇可以定义为 S,多类样本簇可以定义为 L。则二者的簇样本数量可分别表示为 $|S|$ 和 $|L|$,将少类簇及多类簇的簇中心分别表示为 $s_j(1,2,\dots,|S|)$ 和 $l_j(1,2,\dots,|L|)$。

2.3 自适应分配权重

为了在对少类簇过采样时充分考虑每个簇内部的数据分布和簇间的差异,提出了一种双权重分配方法来确定每个少类簇应当生成新的少类样本的数量。在马氏距离度量基础上,双权重要素包括:类簇间距离要素和簇稀疏度要素。

类簇间距离为首要因素,它衡量了不同簇到分类边界的远近从而判断其对分类的重要性,通过下式计算:

$$r_i=e^{-\frac{1}{|L|}\sum_{j=1}^{|L|}\|s_i-l_j\|}, \tag{19}$$

式中,$s_i, i=1,2,\dots,|S|$ 和 $l_j, j=1,2,\dots,|L|$ 分别表示第 i 个少类簇中心以及第 j 个多类簇中心。由于越靠近边界的样本信息在分类中越重要,且不易学习。因此,当第 i 个簇距离多类样本越远时,类间距离因子 r_i 变小。

另一因素是少类簇的稀疏度,其可以描述少类簇中样本的密度分布情况,如式(20):

$$p_i=\frac{1}{m_i}\sum_{l=1}^{m_i}\|x_{il}-s_i\|,\ i=1,2,\dots,|S|. \tag{20}$$

其中 m_i 表示第 i 个少类簇的样本数,x_{il} 表示第 i 个少类簇第 l 个样本。由此可见,当第 i 个少类簇分布得越紧密时,p_i 越大。

考虑到上述两个因素,确定每个少类簇需要合成的新样本数,用以下公式对距离因素和密度因素进行归一化:

$$\hat{r}_i=\frac{r_i-R_{\min}}{R_{\max}-R_{\min}}, \tag{21}$$

$$\hat{p}_i=\frac{p_i-P_{\min}}{P_{\max}-P_{\min}}. \tag{22}$$

其中,$R_{\max}$ 为 $\max\{r_i\}$;$R_{\min}$ 为 $\min\{r_i\}$;$P_{\max}$ 为 $\max\{p_i\}$;$P_{\min}$ 为 $\min\{p_i\}$;$i=1,2,\dots,|S|$。

综合上述因素,考虑边界点信息与簇密度信息,自适应双权重分配因子可表示为:

$$w_i=\hat{r}_i^{\hat{p}_i}+\hat{p}_i. \tag{23}$$

将其标准化为:

$$\hat{w}_i=\frac{w_i}{\sum_{i=1}^{|S|}w_i}. \tag{24}$$

根据上式易得 $\sum_{i=1}^{|S|}\hat{w}_i=1$,由此 $\hat{w}_i$ 表示每个少类簇用于生成的新样本数的综合权重参数。

最后,若给定 G 作为 MAWOTE 需要在少类中新合成的总样本数,则其中第 i 个少类样本簇需要合成的样本数量 g_i 可以通过下式求得:

$$g_i=\hat{w}_i\times G. \tag{25}$$

2.4 过采样插值

有别于传统的欧式距离过采样方法,MAWOTE 算法在过采样时使用马氏距离作为度量标准,该方法能够消除特征间量纲不一问题。此外,为了减少随机采样中产生的不确定性,MAWOTE 算法引入 k 近邻概念,保证了过采样后的数据集服从原始数据分布。因此,基于马氏距离,从第 i 个少类簇中随机抽取样本 $N_{i,s}$,并通过下式过采样新样本:

$$N_{i,new}=N_{i,s}+\frac{1}{k_i}\sum_{j=1}^{k_i}(N_{i,s}-N_{i,sj})\times\theta_i. \tag{26}$$

式中，θ_i 为一个[0,1] 之间的随机数；k_i 为预先给定的最近邻参数；$N_{i,sj}$ 为 k_i 个最近邻样本的其中之一。在后面的实验中，为了计算简单，在少类簇中不同 k_i 被定义为相同的，即 $k_i = k$。需要注意的是，上述公式中距离计算公式皆使用马氏距离。

为了更清晰地展示 MAWOTE 算法，其具体步骤如下：

(1)输入含有缺失值的不平衡数据集；

(2)运用 NLFs 模型以及 MBGD 更新规则，对数据集中的缺失值进行填补；

(3)针对填补后完整的数据集，基于马氏距离，分别对少类与多类样本聚类；

(4)自适应计算每个少类样本簇对应分配的权重；

(5)对少类进行过采样，合成新的少类样本以平衡数据集。

3　实验结果与分析

3.1　数据集与评价标准

为了验证本文算法的有效性和适用性，实验从 UCI 机器学习库[20]中选取了 6 个公共数据集，以此来研究 MAWOTE 在各种场景下的性能。表 1 为本文中使用的数据集信息。

表 1　UCI 数据集信息表

Tab. 1　Information of the UCI Datasets

数据集	样本数	少类数	特征数	不平衡比%
letter	20 000	734	16	3.8
pageblocks	5 473	231	10	4.4
libras	360	72	90	25.0
ecoil	336	77	7	29.7
ilpd	583	167	10	40.1
spambase	4 601	1 813	57	60.5

本文以不同的比率{0.1, 0.3, 0.5}，随机删除特征值。通过不同大小、属性和不平衡比率的数据集，验证 MAWOTE 算法的泛化性能。

对于应用过采样算法的分类器性能，根据文献[13]，引入 4 个主要的评价指标，其中包括总体准确度、查准率、查全率和 G-Mean：

$$\text{OversallAccuracy} = \frac{TP + TN}{TP + FN + FP + TN}, \tag{27}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{28}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{29}$$

$$\text{G - Mean} = \sqrt{\frac{TP}{TP + FN}} \times \sqrt{\frac{TN}{TN + FP}}. \tag{30}$$

OverallAccuracy(ACC)反映了正确分类的样本数量。然而，在不平衡学习条件下，ACC 无法准确判断少数样本的分类情况。式中 TP 和 FN，分别表示少类中正确预测和错误预测的样本数，TN 和 FP 分别表示多类中正确预测和错误预测的样本数。

此外，在数据不平衡的情况下，ROC 曲线下面积(AUC)是一个评估分类器性能的重要指标。因此，AUC 指标也将被纳入衡量指标中。AUC 越接近 1，分类性能越好。

3.2　对比实验分析

本文所有实验都是在一台 2.2 GHz CPU、16 GB 内存、Ubuntu 操作系统的电脑上完成，软件环境为 python3.7。

为了验证本文算法的优越性，选用支持向量机(SVM)作为基本分类器，在{1,10,100}范围内，对松弛变量进行优化。为了验证所提出的算法与传统方法相比是否能达到有效水平，本文对比了各种过采样及缺失值填补方法，见表 2。参数设置参考文献[13]。

表 2　对比实验方法表

Tab. 2　Methods for Comparative Experiments

缺失值填补	过采样方法	结合方法
MI	SMOTE	MISM
RI	MWMOTE	RIMW
KNN	Borderline SMOTE	KNBS
EM	ADASYN	EMAD
/	/	FID

在 MAWOTE 算法中，为了方便计算，设置 $\lambda_X = \lambda_Y = \lambda_Z = 1$。FCM 中模糊因子 m 根据经验值通常取 2[15]，k_i 从{3,5,8,10}中根据性能表现选取最优值。

结合上述评价指标，表 3 和表 4 展示了两种缺失率状态下的算法最佳性能，其中粗体字代表该组中最佳结果。

表 3 及表 4 的实验结果表明，在大缺失、大规模的不平衡数据集下，MAWOTE 的性能波动更小。尽管 MAWOTE 并不在所有数据集中最优，但其拥有更好的泛化性能。FID 是 MAWOTE 着重对比的算法，FID 的“查准率”指标明显大于“查全率”指标，这说明分类超平面仍然倾斜。而 MAWOTE 在马氏距离的基础上考虑了更多的样本类内信息，因此在面对不同数据集时，性能稳定性更具优势。

表 3 0.1 缺失率下的比较结果表

Tab. 3 Comparison results on 0.1 missing rate

数据集	评价指标	MISM	RIMW	KNBS	EMAD	FID	MAWOTE
letter	ACC	0.921	0.722	0.862	0.905	0.937	**0.945**
	Precision	0.278	0.538	**0.875**	0.478	0.722	0.559
	Recall	0.864	0.291	0.578	0.866	0.375	0.852
	G-Mean	0.793	0.539	0.714	**0.917**	0.611	0.806
	AUC	0.837	0.601	0.868	0.916	0.736	**0.918**
pageblocks	ACC	0.900	0.893	0.876	0.925	**0.967**	0.934
	Precision	0.415	0.256	0.525	0.436	**0.824**	0.554
	Recall	0.875	0.858	0.904	0.928	0.390	**0.930**
	G-Mean	0.829	0.864	0.914	0.921	0.628	**0.927**
	AUC	0.869	0.852	0.908	0.905	0.702	**0.925**
libras	ACC	0.916	0.930	**0.930**	**0.930**	0.887	0.895
	Precision	0.833	**0.923**	0.730	0.714	0.771	0.829
	Recall	0.833	0.750	**0.916**	0.909	0.744	0.738
	G-Mean	0.877	0.858	0.924	0.926	0.750	**0.928**
	AUC	**0.899**	0.866	0.857	0.848	0.798	0.833
ecoli	ACC	0.835	0.881	0.836	0.850	0.924	**0.936**
	Precision	0.867	**1.000**	0.928	0.894	0.973	0.904
	Recall	0.591	0.667	0.565	0.680	**0.918**	0.902
	G-Mean	0.751	0.774	0.743	0.804	0.931	**0.945**
	AUC	0.846	0.927	0.870	**0.964**	0.959	0.951
ilpd	ACC	0.620	0.703	0.686	**0.794**	0.793	0.794
	Precision	0.749	0.700	0.831	**0.867**	0.656	0.781
	Recall	0.500	0.667	0.669	0.476	0.720	**0.793**
	G-Mean	**0.745**	0.653	0.693	0.687	0.703	0.707
	AUC	0.773	0.734	0.810	0.783	**0.868**	0.844
spambase	ACC	0.916	0.903	0.884	0.913	0.847	**0.932**
	Precision	**0.909**	0.884	0.850	0.856	0.881	0.867
	Recall	0.882	0.823	0.832	0.933	0.696	**0.936**
	G-Mean	0.901	0.873	0.867	**0.917**	0.810	**0.917**
	AUC	0.912	0.874	0.858	**0.908**	0.819	0.863

表 4 0.5 缺失率下的比较结果表

Tab. 4 Comparison results on 0.5 missing rate

数据集	评价指标	MISM	RIMW	KNBS	EMAD	FID	MAWOTE
letter	ACC	0.762	0.734	0.803	0.767	**0.957**	0.955
	Precision	0.203	0.124	0.213	0.182	**0.804**	0.267
	Recall	0.750	0.708	0.708	0.800	0.432	**0.804**
	G-Mean	0.766	0.802	0.836	0.782	0.656	**0.889**
	AUC	0.843	0.779	0.788	0.883	0.815	**0.903**
pageblocks	ACC	0.925	0.888	0.896	**0.952**	0.945	0.926
	Precision	0.382	0.590	0.443	**0.688**	0.633	0.627
	Recall	0.863	0.762	0.695	0.871	0.544	**0.909**
	G-Mean	0.900	0.853	0.729	**0.916**	0.713	0.882
	AUC	0.894	0.766	0.747	0.905	0.759	**0.912**
libras	ACC	0.902	0.930	**0.958**	0.944	0.915	0.936
	Precision	0.882	0.700	0.857	0.727	**0.867**	0.884
	Recall	0.750	0.778	0.923	0.889	0.765	**0.941**
	G-Mean	0.849	0.860	**0.932**	0.920	0.858	0.923
	AUC	0.895	0.865	0.929	0.855	0.864	**0.933**
ecoli	ACC	0.955	0.955	0.761	0.895	0.939	**0.958**
	Precision	0.889	0.952	0.647	**1.000**	0.875	0.958
	Recall	0.941	0.909	0.523	0.611	0.875	**0.942**
	G-Mean	**0.950**	0.930	0.675	0.741	0.916	**0.950**
	AUC	0.934	**0.942**	0.723	0.919	0.917	0.931
ilpd	ACC	0.703	0.786	0.720	0.762	0.853	**0.863**
	Precision	**0.781**	0.583	0.767	0.735	0.758	0.776
	Recall	0.490	0.689	0.683	0.681	0.730	**0.754**
	G-Mean	0.681	0.642	0.657	**0.814**	0.743	0.808
	AUC	0.758	0.685	0.768	**0.885**	0.847	0.860
spambase	ACC	0.874	0.888	0.926	0.952	0.899	**0.961**
	Precision	0.808	0.902	0.747	0.917	**0.926**	0.882
	Recall	0.887	0.822	0.815	**0.966**	0.751	0.889
	G-Mean	0.877	0.876	0.907	0.917	0.852	**0.936**
	AUC	0.886	0.855	0.904	**0.973**	0.858	0.965

4 结束语

本文提出了一种针对含带缺失信息的不平衡数据的自适应马氏距离双权重过采样方法,即 MAWOTE 算法。MAWOTE 的基本思想概括为:考虑样本中全局已知特征值的信息,利用扩展了解空间后的 NLFs 模型,结合 MGBD 更新规则,对数据集进行精确的填补;引入了马氏距离消除特征间的量纲不一问题;在使用 FCM 算法生成类别簇的基础上,综合类

间距离、类内密度因素，对少类簇自适应分配双权重，提高合成样本质量；采用 k 近邻思想在少类簇内合成新的少类样本，保证数据集在新样本添加后，依然能够维持原始信息分布。实验结果表明，MAWOTE 在 6 个不同大小、维度、缺失率的数据集上取得了稳定的性能表现，与过往的算法相比，性能更为突出。

参考文献

[1] ZARATE L E, NOGUEIRA B M, SANTOS T R A, et al. Techniques for Missing Value Recovering in Imbalanced Databases: Application in a Marketing Database with Massive Missing Data[C]// IEEE International Conference on Systems. 2007.

[2] LIU Haoyue, ZHOU Mengchu, LIU Qing. An Embedded Feature Selection Method for Imbalanced Data Classification[J]. IEEE/CAA Journal of Automatica Sinica, 2019, 6(3): 703-715.

[3] 胡峰, 王蕾, 周耀. 基于三支决策的不平衡数据过采样方法[J]. 电子学报, 2018, 46(1): 135-144.

[4] ABDI L, HASHEMI S. To combat multi-class imbalanced problems by means of over-sampling and boosting techniques[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(1): 238-251.

[5] GALAR M. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches [J]. IEEE Transactions on Systems Man & Cybernetics Part C Applications & Reviews, 2012, 42(4): 463-484.

[6] WEISS G M, MCCARTHY K, ZABAR B. Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs[C]//International Conference on Data Mining. 2007.

[7] 张冬雪. 基于欠采样不平衡数据 SVM 算法与应用[D]. 哈尔滨:哈尔滨工程大学, 2013.

[8] TOMEK I. Two Modifications of CNN[J]. IEEE Trans. Syst., Man, Cybern., 1976, 6(11):769-772.

[9] 李艳霞, 柴毅, 胡友强, 等. 不平衡数据分类方法综述[J]. 控制与决策, 2019, 34(4): 4-19.

[10] PAN R, YANG T S, CAO J H, et al. Missing data imputation by K nearest neighbours based on grey relational structure and mutual information[J]. Appl. Intell., 2015, 43(3):614-632.

[11] LUO X, ZHOU M C, XIA Y N, et al. An Efficient Non-Negative Matrix-Factorization-Based Approach to Collaborative Filtering for Recommender Systems[J]. IEEE Trans. Ind. Inf., 2014, 10(2):1273-1284.

[12] SHANG M S, LUO X, LIU Z G, et al. Randomized Latent Factor Model for High-dimensional and Sparse Matrices from Industrial Applications[J]. IEEE/CAA J. Autom. Sinica, 2019, 6(1):131-141.

[13] LIU S G, ZHANG J, XIANG Y, et al. Fuzzy-Based Information Decomposition for Incomplete and Imbalanced Data Learning[J]. IEEE Trans. Fuzzy Syst., 2017, 25(6):1476-1490.

[14] ABDI L. HASHEMI S. To combat multi-class imbalanced problems by means of over-sampling and boosting techniques[J]. IEEE Trans. Knowl. Data Eng., 2016, 28(1):238-251.

[15] 王燕, 亓祥惠, 段亚西. 基于核函数与马氏距离的 FCM 图像分割算法[J]. 计算机应用研究, 2020, 37(2): 611-614, 624.

[16] LUO X, ZHOU M, SHANG M, et al. A novel approach to extracting non-negative latent factors from non-negative big sparse matrices[J]. IEEE access, 2016,4:2649-2655.

[17] 阴法明, 赵晓铃. 用于运动估计的基于梯度下降搜索扩展算法[J]. 计算机工程与应用, 2010(33): 139-141.

[18] 鲁淑霞, 周谧, 金钊. 非均衡加权随机梯度下降 SVM 在线算法[J]. 计算机科学与探索, 2017, 11(10): 1662-1671.

[19] 毕小然, 闫绍山, 高迎. 基于小批量梯度下降法的个性化推荐模型[J]. 计算机科学与应用, 2019, 9(4): 695-702.

[20] M. Lichman, UCI Machine Learning Repository. (2016). [Online]. Available: http://archive.ics.uci.edu/ml.

(上接第 30 页)

或侧摔、仰摔时，会被误判为摔倒事件；文中方法融合了 5 个摔倒特征参数，经过各参数之间的相互修正并进行最终判定，能更加精确识别摔倒事件，降低误判率。

4　结束语

本文针对独居老人在室内发生摔倒事件的情形，提出了基于多特征融合的摔倒检测方法。该方法通过背景减除法来分割运动目标，使用混合高斯模型算法对背景进行更新，在获取完整的目标后，再依次使用人体宽高比、人体有效面积比、人体质心到底边距离、中心变化率、高度变化率 5 个特征参数判断是否有摔倒事件发生。该方法简单、易于实现、误判率低，能更为准确地区分摔倒行为和日常活动行为。

参考文献

[1] 秦敏花.中国人口老龄化发展现状、成因与对策研究[J]. 企业科技与发展,2019 (9):219-220

[2] FISH R F A, MESSENGER H, BARYUDIN L, et al. Falldetection system using a combination of accelerometer, audio inputand magnetometer: U.S.Pand Application 14/465,489[P]. 2014-8-21.

[3] 陈伟,周晴,曹桂涛. 基于 svm 和阈值分析法的摔倒检测系统[J]. 计算机应用与软件,2017,34(7):182-187.

[4] ALWAN M, RAJENDRAN P J, KELL S, et al. A smart and passive floor-vibration based fall detector for elderly[C]// Information and Communication Technologies, 2006. ICTTA'06. 2nd, 2006: 1003-1007.

[5] VAIDEHI V, GANAPATHY K, MOHAN K, et al. Video Based Automatic Fall Detection in Indoor Environment [C]// International Conferenceon Recent Trends in Information Technology, 2011: 1016-1020.

[6] NADI M, EL-BENDARY N, HASSANIEN A E, et al. Falling detection system based on machine learning[C]//Proceedings of the 4th International Conference on Advanced Information Technology and Sensor Application. Harbin: IEEE,2015.

[7] ADRI'AN NUM~EZ-MARCOS, AZKUNE G, ARGANDA-CARRERAS I. Vision-Based Detection with Convolutional Neural Networks[J]. Wireless Communications and Mobile Computing, 2017,(1):1-16.

[8] LOPEZ-FUENTESL, JOOST V D W, GONZALEZ-HIDALGO M, et al. Review on Computer Vision Techniques in Emergency Situation [J].Multimedia Tools & Applications,2017 (1):1-39.