DOI:10.19651/j.cnki.emt.2518102

基于 CEEMDAN 的脉搏波数据增强双层 SMOTE 算法*

李 辉 李振华 李瑞杰 张志东 薛晨阳
(中北大学仪器科学与动态测试教育部重点实验室 太原 030051)

摘 要:针对 SMOTE 算法在处理脉搏波数据不平衡问题中存在噪声干扰敏感及生成样本物理特性失真等问题。本文提出了一种基于 CEEMDAN 改进的 CP-SMOTE 算法,将预处理的脉搏波分解为主波层与次波层分别生成样本,可有效去除残余噪声。同时,在生成新样本时,本算法结合脉搏波信号特征,设计了自适应距离度量和约束监督机制,确保生成样本在保持数据物理特性的同时增强类间区分度。基于自建数据集和公开 PPG-BP 数据集,结合四种分类器对改进算法进行了全面实验。在自建数据集中,CP-SMOTE 在 AUC、G-mean、F1 等关键指标上全面超越 SMOTE 系列算法,最低提升 1.51%,最高提升 18.25%。在公开数据集中对比其他改进算法准确率、G-mean 和 AUC 分别提升 2.24%、1.47%和 1.43%以上。结果表明,该算法显著优于传统 SMOTE 及其他变种算法,研究结果验证了该算法生成的样本有效避免了物理特性失真问题与噪声干扰问题。

关键词:数据不平衡;CEEMDAN;自适应距离;约束监督机制;过采样算法
中图分类号:TN911.72 **文献标识码**:A **国家标准学科分类代码**:520.20;510.99

CEEMDAN-based pulse wave data augmentation with two-layer SMOTE

Li Hui Li Zhenhua Li Ruijie Zhang Zhidong Xue Chenyang
(Key Laboratory of Instrumentation Science and Dynamic Testing, Ministry of Education, North University of China, Taiyuan 030051, China)

Abstract: To address the SMOTE algorithm′s noise sensitivity and physical distortion in pulse wave imbalance processing, this study proposes a CEEMDAN-enhanced CP-SMOTE that decomposes preprocessed pulse waves into primary/secondary layers for stratified sample generation, effectively eliminating residual noise. By integrating adaptive distance metrics and constrained supervision mechanisms aligned with pulse wave characteristics, the algorithm ensures physiologically authentic sample generation while enhancing inter-class discriminability. Evaluations on proprietary and public PPG-BP datasets with four classifiers demonstrate CP-SMOTE′s superiority: 1.51%−18.25% improvements in AUC, G-mean, and F1 scores on proprietary data, CP-SMOTE consistently outperformed SMOTE-based algorithms across key metrics including AUC, G-mean, and F1-score, with improvements ranging from 1.51% to 18.25% on proprietary data, and minimum 1.43% gains in Accuracy(2.24%), G-mean (1.47%) and AUC (1.43%) on public data, confirming its effectiveness in mitigating physical distortion and noise interference compared to SMOTE variants.

Keywords: class imbalance; CEEMDAN; adaptive distance metric; constrained supervision mechanism; oversampling algorithm

0 引 言

脉搏信号可表征人体健康状态,其形态特征与多种疾病相关联,因此准确分析脉搏波数据对诊断与治疗疾病有着重要意义。在实际的数据采集过程中,经常面临样本失衡问题,若样本之间相差过大,会使模型更加偏向类别多的样本,使得总体模型性能下降。而少数类样本也包含重要信息,理应受到同样的关注[1]。合成少数过采样算法

收稿日期:2025-02-18
*基金项目:山西省重点研发计划(202102130501011)、海南省重点研发计划(ZDYF2022SHFZ304)项目资助

(synthetic minority oversampling technique,SMOTE)是解决不平衡分类的常用方法[2]。近年来,SMOTE 方法吸引了众多研究者的注意,广泛的应用到各个领域中,如异常检测[3]、回归问题[4]等。

目前众多学者也提出了许多通用 SMOTE 变种以解决 SMOTE 算法的各种限制。Chen 等[5]聚焦于噪声条件,通过相对密度将非噪声少数样本划分为边界样本和安全样本,显著的降低噪声对于 SMOTE 过程的干扰。梅大成等[6]提出了特征边界和密度适应的 SMOTE 算法,通过规划安全区域和基于密度调整参数,确保生成数据具有明确特征边界,有效防止过拟合。王曜等[7]通过 K 均值聚类算法(k-means clustering,K-means)聚类和正类安全水平指标确定最佳生成区域,有效缓解数据边缘化、噪声等问题。Maldonado 等[8]针对高维数据,使用加权闵可夫斯基距离定义少数类的每个样本的邻域,使得 SMOTE 能够更加贴合于高维数据。He 等[9]提出一种提出一种基于交叠最小化的算法,使得使合成样本尽可能远离类别交叠的区域。

大多数方法都没有考虑脉搏波数据的时间信息(即连续变化的特征),导致生成的样本不符合实际的生理或物理规律。Zhao 等[10]提出的 T-SMOTE 充分利用时间序列数据的时态信息生成边界样本并进行加权采样,利用少数类边界附近的样本合成更多的样本。但并未关注信号本身的特性。并且,如果原始信号包含噪声,SMOTE 方法会将噪声也归入到新样本数据的生成过程中,这个过程会放大噪声的影响,影响训练的效果。Yang 等[11]将动态时间调整(dynamic time warping,DTW)引入 SMOTE 中获得不同时间序列之间更为合理的相似性,但并未考虑动态噪声造成的影响。此外,在寻找 k 近邻时,使用欧氏距离匹配不同数据之间的相似性也不符合脉搏波的特性[12],大量的 SMOTE 变种算法并未对此不足进行改进。因此,针对上诉问题,本文提出了将一种新型的 SMOTE 方法,主要针对于脉搏波信号的不平衡问题,引入自适应噪声完备集合经验模态分解(complete ensemble EMD with adaptive noise,CEEMDAN)到 SMOTE 方法中,将脉搏波数据分解为主波层与次波层,同时削减信号自身残余噪声,然后分别在两个层通过自适应距离衡量样本相似度,然后通过监督约束机制生成新的主波层与次波层样本,通过加权重构出新的脉搏波样本。实验表明,所提出的改进型 SMOTE 算法能更好地处理脉搏波信号,既有效地平衡了数据分布,又较好地保持了合成样本与原始样本的一致性,有效解决数据不平衡问题。

1　CP-SMOTE 算法

1.1　SMOTE 算法

SMOTE 方法的核心思想是通过少量样本与其相邻样本之间的随机线性差值来获得新样本,SMOTE 的具体流程如下:

1)对于每个少量样本 $\boldsymbol{X}_i(i=1,2,\cdots,n)$,以欧氏距离为标准计算它到少数类样本的距离,得到其 k 近邻。

2)根据样本不平衡比例设置一个采样比例以确定采样倍率 N,对于每一个少数类样本 $\boldsymbol{x}$,从其 k 近邻中随机选择若干个样本,假设选择的近邻为 $\boldsymbol{x}_n$。

3)对于每一个随机选出的近邻 $\boldsymbol{x}_n$,分别与原样本按照如下的公式构建新的样本,其中 $\boldsymbol{X}_i$ 是当前样本,$\boldsymbol{X}_j$ 是与 $\boldsymbol{X}_i$ 最相似的一个邻居样本,α 是一个随机值,通常在[0,1]之间,控制差值的程度。合成示例如图 1 所示。

$$\boldsymbol{X}_{new}=\boldsymbol{X}_i+\alpha(\boldsymbol{X}_j-\boldsymbol{X}_i) \tag{1}$$

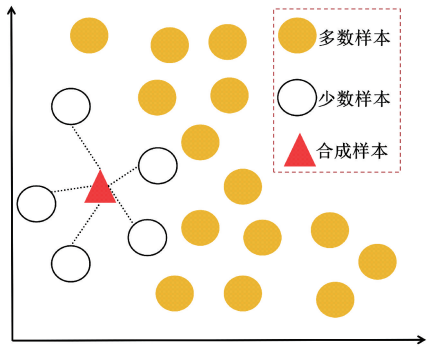

图 1　SMOTE 合成新样本示例

Fig. 1　SMOTE synthesizes new sample examples

1.2　基于 CEEMDAN 改进 SMOTE

通过 CEEMDAN 将脉搏波数据进行分割,得到多个 IMF 分量,每个分量代表不同时间尺度上的特征,这有助于更好的保留时间序列结构。然后对每层 IMF 计算能量占比,找到能量占比最大的分层作为主波层,因为脉搏波是一种低频信号,能量集中在低频部分,一般在 0.5～5 Hz 之间[13]。因此能量占比最大的 IMF 层对信号的主要形状起着主导作用。主波层包含着信号的全局特征,对应着每个脉搏波周期的关键点,比如主波、潮波点、重搏波点等。然后将剩余的所有 IMF 分量作为次波层,次波层为在主波基础上叠加的小幅度波动与细节变化。所有信号均通过高斯滤波器与小波变换进行了去噪处理,但信号中仍然可能存在噪声,CEEMDAN 通过自适应白噪声添加,可以有效消除原始信号残余噪声,减少噪声对合成样本的影响。将提取出来的主波层与次波层分别应用改良 SMOTE 方法扩充信号,随后按照分解时顺序进行加权叠合,得到新生成的脉搏波数据。本文对 SMOTE 方法的细节部分进行了一定的改良以适应 CEEMDAN 分解过程,具体的改动如下:

1)在分解出成主波层与次波层后,进行 SMOTE 之前,对主波层提取出峰值位置以及数量,信号均值、标准差、波形轮廓特征向量。次波层计算能量值,过零点数量以及细节特征。这些特征用于保证生成新的脉搏波数据与原始数据在形状上不会出现太大的偏差,保持脉搏波的物理意义,使得生成的新样本能够很好的保留类别特征信息。

2)自适应距离计算:尽管 SMOTE 方法简单且拥有良

好的性能,但它同样存在一个明显的缺点,用于计算少数样本的 k 个最近邻的欧几里得距离是假设所有的变量都是同等相关的[14],这个假设不适合与脉搏波数据,因此,本文使用自适应距离替代 SMOTE 过采样中的欧几里得距离,在自适应距离中考虑了脉波信号的多个特征,而不仅仅是信号的原始值,对于主波层与次波层采用不同策略的自适应距离,这种多特征距离计算方法使得 SMOTE 能够生成更为合理的合成样本,尤其是对于脉波信号这样的时序数据来说,形态、频率等方面的特性是很重要的。同时,自适应距离可以更好的区分类别。防止类别之间信息发生混叠。

对于主波层,首先计算主波层的形状特征差异 sd,假设 $s_1(i)$和 $s_2(i)$分别为形状轮廓 1 和形状轮廓 2 在点 i 的值,N 为形状轮廓的长度,公式为:

$$sd = \frac{1}{N}\sum_{i=1}^{N} | s_1(i) - s_2(i) | \tag{2}$$

计算样本 $\boldsymbol{x}_1$ 与 $\boldsymbol{x}_2$ 之间的皮尔逊相关系数:

$$\rho = \frac{cov(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\sigma \boldsymbol{x}_1 \sigma \boldsymbol{x}_2} \tag{3}$$

其中,$cov(\boldsymbol{x}_1, x_2)$表示两个样本之间的协方差,而 $\sigma \boldsymbol{x}_1$ 和 $\sigma \boldsymbol{x}_2$ 代表 $\boldsymbol{x}_1$ 与 $\boldsymbol{x}_2$ 的标准差。

和原始 SMOTE 方法一样计算欧式距离 d_E,$\boldsymbol{x}_{1,norm}$ 和 $\boldsymbol{x}_{2,norm}$ 表示两个归一化之后的序列,公式为:

$$d_E = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{x}_{1,norm}(i) - \boldsymbol{x}_{2,norm}(i))^2} \tag{4}$$

因此,最终主波层的自适应距离 D_{main} 可以计算为这 3 个参数的加权求和,其中的加权值是计算每个度量对最终结果的影响程度,根据其贡献大小动态调整权重计算出来的合适值。

$$D_{main} = 0.4 \times sd + 0.3 \times (1 - \rho) + 0.3 \times d_E \tag{5}$$

次波层与主波层所蕴含的原始信号信息不同,因此根据步骤 1)对不同层提取的特征向量不用,次波层与主波层的自适应距离计算公式也不同,这种不同层采用不同方法的距离计算方法能够使得 SMOTE 能够生成更加合理的目标样本。次波层的自适应距离首先计算不同次波层的能量差异 ed,能量差异是两个次波层 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 能量差的绝对值,能量定义为信号平方和。

$$ed = | \sum_{i=1}^{N} x_1(i)^2 - \sum_{i=1}^{N} x_2(i)^2 | \tag{6}$$

根据提取出来的过零点信息,计算两个样本之间的零点穿越次数差值的绝对值,零点穿越次数是指信号从正值变为负值或者从负值变为正值的次数。设 $zeros_1$ 和 $zero_2$ 分别表示样本 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 的零点穿越次数,则零点穿越次数的计算方法与零点穿越差异的计算公式为:

$$zero_{1,2} = len(where(diff(sgn(\boldsymbol{x}_{1,2})))) \tag{7}$$

$$zd = | zero_1 - zero_2 | \tag{8}$$

接着计算两个样本间的频谱差异,同自适应距离主波层距离计算两个之间的欧几里得距离,假设 $F(x)$表示快速傅里叶变换过程(fast Fourier transform,FFT),则频谱差异的计算公式为:

$$fd = \sqrt{\sum_{k=0}^{M-1} | \sum_{n=0}^{N-1} x_1[n] - \sum_{n=0}^{N-1} x_2[n] | \mathrm{e}^{-\frac{j2\pi kn}{N}}} \tag{9}$$

其中,M 表示 FFT 结果的长度,通常等同于原始序列长度 N。

因此,最终次波层的自适应距离 $D_{residual}$ 可以计算为这 3 个参数的加权求和,次波层的加权参数也是同主波层一致,计算每个度量对最终结果的影响程度,根据其贡献大小动态调整权重计算出来的合适值。

$$D_{residual} = 0.3 \times (ed + zd) + 0.4 \times fd \tag{10}$$

3)多重监督约束机制:对主波层和次波层计算自适应距离之后,与原始 SMOTE 方法一致,选择 k 个近邻进行进行样本插值生产新样本,针对生产的每一个新样本,设计了一种多重监督约束机制,根据主波层与次波层的不同特征信息,分别对新样本进行约束,对超过监督约束值的新样本应用不同的调整机制以满足要求并重新验证约束条件。

对于主波层:(1)峰值数量与位置监督约束:对旧样本与新样本对比峰值位置和数量,若出现较大误差则调整波形,确保生成波形的主要特征点数量与源波形相近,保留脉搏波信号的基本形态特征,防止出现虚假峰值。(2)幅值范围监督约束:检查新样本的幅值范围,控制生成样本的幅值范围在原始信号的 0.5～1.5 倍之间,防止出现虚假的波动,并保持信号的能量水平。(3)形状相似性监督约束:计算新样本与源数据的相关系数,确保相关系数不低于 0.75,维持波形的基本特征。

对于次波层:(1)能量监督约束:确保新样本的能量与原始信号的能量变化不超过 50%,防止能量过度衰减或放大,维持信号的强度特征。(2)过零点监督约束:限制新波形中的过零点数量不超过原始波形的 1.5 倍,可以控制信号的震荡特性,保持信号的频率特征,防止过度平滑或者噪声干扰。(3)频谱监督约束:控制新样本的频域特征与原始信号变化在 40%以内,维持频域特征分布,保持信号的频率成分。

最后按照 CEEMDAN 分解的次序将主波层与次波层重构回脉搏波数据,算法的总体流程如下:

算法 1 CP-SMOTE

Input:原始信号集合 signals,标签集合 labels,目标样本数 target_num

Output:合成信号 synthetic_signals

```
For each 信号 in signals do:
    imfs = CEEMDAN 分解(信号)
    For i = 0 to 信号分量 do:
        energies[i] = 计算能量(imfs[i])
    End For
```

```
6       主波层 = 信号分量[获取最大能量索引(energies)]
7       次波层 = Sum(imfs[除最大能量索引外])
8       End For
9       For each class in labels do:
10          While 样本数 < target_num do:
11              features = 提取特征(主波)
12              距离 d=0.4×形状特征+0.3×(1−相关系数)+0.3×归一化欧氏距离
13              neighbor = 获取近邻(主波, features)
14              new_main = 插值生成(主波, neighbor)
15              features = 提取特征(次波)
16      距离 d2=0.3×能量差异 +0.3×过零点差异 +0.4×频谱差异
17              neighbor = 获取近邻(次波, features)
18              new_sub = 插值生成(次波, neighbor)
19              If 监督约束(new_main) and 监督约束(new_sub) Then
20                  结果.Add(new_main + new_sub)
21      End If
22      End While
23      End For
24      Return 结果
```

2　实验分析

2.1　数据集

为了验证本文所提出方法的有效性，从山西中医药大学附属医院以及山西广誉远国医馆于 2023 年 12 月～2024 年 6 月接受治疗的患者，患者数据由中医医生进行数据标注。数据采样率为 1 kHz，采集患者桡动脉处脉搏波数据，经过人工筛选剔除无用数据后，采集到弦脉数据 142 个，沉脉数据 330 个，样本失衡较大。采用小波变换滤除高频噪声与基线漂移[15]。

本文同时选取 PPG-BP 脉搏波公开数据集进行实验，包含来自 219 名受试者的 657 条脉搏波数据集。数据集涵盖的年龄范围为 20～89 岁，疾病记录包括高血压和糖尿病以及每名受试者的参考血压值[16]，其中糖尿病样本 38 人，正常样本 180 人，对数据集采用高斯滤波与小波变换滤波进行处理得到无噪声干扰数据。

2.2　评价指标

对于分类任务，常规的分类的指标为准确率(accuracy)、精确率(precision)、召回率(recall)、F1 系数[17]，为更好的分析不平衡数据集的分类效果，本文增加选用几何平均(geometric mean，G-mean)、ROC 曲线下面积(area under curve，AUC)作为评价指标，这些指标均由混淆矩阵计算所得。混淆矩阵可以直观表现出分类器的分类效果，表 1 为二分类混淆矩阵。

表 1　混淆矩阵

Table 1　Confusion matrix

类别	预测正类	预测负类
实际正类	TP	FN
实际负类	FP	TN

G-mean 可以综合评价两类类别的分类正确率。G-mean 值较高时，说明模性性能较好。AUC 能够同时评估分类器对于多数类样本和少数类样本的分类能力，AUC 越高，也证明分类器分类效果越好。各个评价指标的计算方法如下：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{14}$$

$$G - mean = sq(\frac{TP}{TP + FP} \times \frac{TP}{TP + FN}) \tag{15}$$

$$AUC = \frac{(1 + \frac{TP}{TP + FN} - \frac{FP}{FP + TN})}{2} \tag{16}$$

2.3　实验设置

为了评估本文方法的有效性，在使用主成分分析(principal component analysis，PCA)对数据进行降维之后，本文采用了随机森林(random forest，RF)和支持向量机(support vector machine，SVM)进行分类[18-19]，对原始脉搏波数据，本文采用包含两个隐藏层的卷积神经网络(convolutional neural networks，CNN)模型和长短期记忆网络(long short-term memory LSTM)模型进行分类[20-21]。同时与其他 4 种 SMOTE 变种在分类器上进行对比，这 4 种变体分别是 SMOTE-kTLNN[22]、H-SMOTE-FAM[23]、Time series-SMOTE[10](简为 T-SMOTE)、RSMOTE[5]。

本文的所有 SMOTE 算法的过采样率都设置为 1，表示将少数类样本生成至多类样本数量。因 PPG-BP 数据集正常人类别与糖尿病患者类别数量相差过大，失衡比例大于 4，故删除部分正常人原始样本，使得正常人样本/糖尿病样本=3。

2.4　结果分析

为客观的比较不同算法对于分类器的提升，采用 5 折交叉验证的方法消除随机性对实验的干扰。将数据集按照 7∶3 的比例进行划分成训练集、测试集。并且确保测试集中不包含采用算法合成的新样本，全部采用原始数据进行测试，以获得最真实的实验结果。实验结果如表 2～3 所示。

从表 2 与 3 中可以看出，本文提出的算法在自建数据集与公开的 PPG-BP 数据集上相较于 SMOTE 算法均表现出显著的性能提升。在自建数据集上，本算法在 SVM 和

表 2　各算法在自建数据集上的表现

Table 2　The performance of each algorithm on self-built dataset

算法	指标	SMOTE	T-SMOTE	RSMOTE	SMOTE-kTLNN	H-SMOTE-FAM	Ours
SVM	accuracy	0.534 9	0.581 4	0.550 4	0.651 2	0.674 4	**0.765 2**
	precision	0.537 9	0.589 9	0.558 6	0.757 6	0.753 1	**0.768 6**
	recall	0.534 9	0.581 4	0.550 4	0.651 2	0.674 4	**0.765 2**
	F1	0.535 9	0.582 2	0.551 2	0.589 0	0.629 5	**0.765 8**
	G-mean	0.533 0	0.583 9	0.552 7	0.487 8	0.549 2	**0.766 4**
	AUC	0.576 5	0.604 7	0.623 8	0.640 8	0.643 5	**0.791 5**
RF	accuracy	0.742 2	0.666 7	0.635 7	0.713 2	0.759 7	**0.825 8**
	precision	0.749 9	0.692 0	0.679 2	0.718 8	0.759 1	**0.826 2**
	recall	0.744 2	0.666 7	0.635 7	0.713 2	0.759 7	**0.825 8**
	F1	0.744 8	0.664 4	0.627 3	0.713 9	0.759 0	**0.824 8**
	G-mean	0.746 5	0.668 4	0.629 5	0.715 3	0.753 3	**0.817 7**
	AUC	0.833 3	0.758 1	0.741 7	0.765 9	0.811 3	**0.919 7**
CNN	accuracy	0.511 6	0.627 9	**0.635 7**	0.612 4	0.620 2	0.628 8
	precision	0.518 2	0.708 5	**0.728 0**	0.631 2	0.628 3	0.635 8
	recall	0.513 6	0.630 3	**0.638 1**	0.613 8	0.621 5	0.628 8
	F1	0.479 6	0.590 5	0.596 9	0.599 7	0.608 8	**0.623 9**
	G-mean	0.447 5	0.550 3	0.554 9	0.585 9	0.586 3	**0.618 4**
	AUC	0.549 0	**0.687 3**	0.567 8	0.613 0	0.630 9	0.652 0
LSTM	accuracy	0.542 6	0.496 1	0.527 1	0.542 6	0.534 9	**0.569 3**
	precision	0.549 0	0.492 8	0.522 4	0.534 1	0.530 1	**0.589 7**
	recall	0.437 5	0.492 8	0.546 9	**0.623 1**	0.676 9	0.630 1
	F1	0.487 0	0.511 3	0.534 4	0.604 4	0.594 6	**0.609 3**
	G-mean	0.490 1	0.511 6	0.534 5	0.587 5	0.599 0	**0.609 6**
	AUC	0.522 1	0.480 5	**0.627 4**	0.532 5	0.609 3	0.611 3

表 3　各算法在 PPG-BP 数据集上的表现

Table 3　The performance of each algorithm on PPG-BP dataset

算法	指标	SMOTE	T-SMOTE	RSMOTE	SMOTE-kTLNN	H-SMOTE-FAM	Ours
SVM	accuracy	0.825 0	0.812 5	0.525 0	0.840 6	0.836 1	**0.863 0**
	precision	**0.868 8**	0.816 0	0.532 8	0.851 4	0.874 2	0.863 0
	recall	0.825 0	0.812 5	0.525 0	0.840 6	0.836 1	**0.863 0**
	F1	0.817 7	0.811 0	0.520 5	0.841 6	0.829 0	**0.863 0**
	G-mean	0.794 7	0.804 4	0.520 3	0.846 9	0.801 8	**0.861 6**
	AUC	0.771 9	0.870 6	0.562 7	0.878 4	0.880 4	**0.894 7**
RF	accuracy	0.800 0	0.805 6	0.750 0	0.804 9	0.833 3	**0.849 3**
	precision	0.802 0	0.856 0	0.778 0	0.835 4	0.857 0	**0.860 3**
	recall	0.800 0	0.805 6	0.750 0	0.804 9	0.833 3	**0.849 3**
	F1	0.799 0	0.792 8	0.746 2	0.818 4	0.838 0	**0.846 4**
	G-mean	0.794 7	0.750 0	0.744 2	0.586 9	**0.845 8**	0.831 2
	AUC	0.852 1	0.871 9	**0.884 7**	0.608 3	0.863 3	0.872 3

表 3（续）

Table 3（continued）

算法	指标	SMOTE	T-SMOTE	RSMOTE	SMOTE-kTLNN	H-SMOTE-FAM	Ours
CNN	accuracy	0.625 0	0.625 0	**0.736 8**	0.637 7	0.655 7	0.698 6
	precision	0.686 5	0.627 0	**0.833 3**	0.676 9	0.697 1	0.700 8
	recall	0.625 0	0.625 0	**0.722 2**	0.626 3	0.633 1	0.697 8
	F1	0.580 7	0.623 5	**0.707 7**	0.604 4	0.614 3	0.697 2
	G-mean	0.533 9	0.621 9	0.666 7	0.568 5	0.569 8	**0.695 3**
	AUC	0.772 5	0.730 5	0.700 0	0.674 2	0.768 4	**0.802 6**
LSTM	accuracy	0.655 0	0.774 2	0.685 0	0.594 2	0.786 9	**0.823 8**
	precision	0.685 3	0.812 5	0.707 6	0.600 0	0.743 6	**0.854 4**
	recall	0.757 1	0.764 7	0.755 7	0.454 5	**0.806 2**	0.796 8
	F1	0.681 6	0.787 9	0.699 8	0.517 2	0.816 9	**0.821 9**
	G-mean	0.701 0	0.788 2	0.715 4	0.522 2	0.810 9	**0.823 7**
	AUC	0.646 4	0.806 7	0.751 7	0.541 2	0.790 4	**0.854 1**

RF 中的各项指标提升尤为显著，其中准确率提升均超过 2.67%。而在 PPG-BP 数据集上，准确率的提升幅度均超过 3.8%。针对精确率和召回率两个聚焦于正类样本的指标，本算法在自建数据集上的提升分别达到 4.07% 和 8.16%。这表明，改进的算法在处理正类样本时表现出更强的识别能力。此外，F1 作为衡量类别重叠现象的重要指标，其数值越高，表明数据之间的类别重叠越少，分类器性能越优。本算法在两个数据集上的 F1 均取得较大提升，其中自建数据集提升幅度超过 8%，PPG-BP 数据集提升幅度为 4.53%。结果表明，CP-SMOTE 能够有效减弱原始信号残余噪声的影响，同时降低数据类别间重叠现象。在两个数据集中，CP-SMOTE 的 G-mean 分别提升了 7.12%和 3.65%，证明对少数类和多数类样本均具较强识别能力。AUC 作为二分类的重要指标，在自建数据集和 PPG-BP 数据集上的提升分别为 8.7%和 3.1%以上，进一步表明算法的优越性。

CP-SMOTE 算法与其他 SMOTE 变种的对比分析。实验结果表明，在自建数据集中，在 SVM 中，CP-SMOTE 的准确率高出 9.08%，F1 提升 13.63%以上；在 RF 中，各项指标均高出 6.58%以上。在 LSTM 和 CNN 中，尽管部分指标略低于其他算法，但综合所有评估指标后，CP-SMOTE 展现了更强的整体分类能力。在 PPG-BP 数据集中，使用 SVM 时，CP-SMOTE 的准确率高出 2.24%，F1 提升 2.14%～24.25%；在 RF 中，各项指标均高出 3.01%以上。在 LSTM 和 CNN 中，尽管部分指标略低于其他算法，但 F1、G-mean，AUC 比其他模型高出 1.28%，CP-SMOTE 展现了更强的整体分类能力，表现出卓越的性能优势。这些结果表明，将 CEEMDAN 引入 SMOTE 并分层处理脉搏波信号效果是显著的。CEEMDAN 方法减弱了数据中的残余噪声影响，自适应距离度量增强了不同类别脉搏波数据的区分能力，而约束监督机制则在数据生成过程中保留了脉搏波的物理特性和临床意义。CP-SMOTE 算法因其在鲁棒性和实用性上的显著表现，为脉搏波数据的处理和分类提供了全新的解决方案。

3　结　　论

在脉搏波数据出现类不平衡问题时，解决类不平衡的有效办法是过采样算法，但目前的过采样算法都没有针对脉搏波数据的有效改进。本文针对脉搏波数据不平衡问题，提出了一种新的过采样算法 CP-SMOTE。该算法将 CEEMDAN 引入 SMOTE 中将信号分为主波层与次波层将信号分为主波层与次波层的同时减弱信号残余噪声的影响。并采用自适应距离和约束监督机制，在合成新样本的同时保持了数据的物理特性和类间区分度。在自建数据集和公开数据集上的实验表明，CP-SMOTE 在 SVM、RF、CNN 和 LSTM 等多种分类器上均取得了优于 SMOTE 的性能表现，准确率、精确率、召回率、F1 值、G-mean 和 AUC 等评估指标全面领先，并在总体指标上优于 RSMOTE 等算法，证实了算法的有效性和泛化能力。研究结果为解决人体脉搏波数据的类别不平衡问题提供了新的思路和方法。

参考文献

[1] LI Y J, ZHANG J, ZHANG S, et al. Multi-objective optimization-based adaptive class-specific cost extreme learning machine for imbalanced classification[J]. Neurocomputing, 2022, 496: 107-120.

[2] 王晓霞，李雷孝，林浩. SMOTE 类算法研究综述[J]. 计算机科学与探索，2024，18(5)：1135-1159.
WANG X X, LI L X, LIN H. Review of SMOTE algorithm[J]. Computer Science and Exploration, 2024, 18(5): 1135-1159.

[3] 李文悦，何怡刚，邢致恺，等. 基于双输入残差图卷积网络的电力变压器健康状态评估方法[J]. 电子测量与仪器学报，2024，38(11)：15-24.

LI W Y, HE Y G, XING ZH K, et al. Health evaluation of power transformer based on double input residual graph convolutional network[J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(11):15-24.

[4] CAMACHO L, DOUZAS G, BACAO F. Geometric SMOTE for regression[J]. Expert Systems with Applications, 2022, 193: 116387.

[5] CHEN B Y, XIA SH Y, CHEN Z ZH, et al. RSMOTE: A self-adaptive robust SMOTE for imbalanced problems with label noise[J]. Information Sciences, 2021, 553: 397-428.

[6] 梅大成,陈江,郑涛. 边界与密度适应的 SMOTE 算法研究[J]. 计算机应用研究,2022,39(5):1478-1482.
MEI D CH, CHEN J, ZHENG T. Research on SMOTE algorithm based on boundary and density adaptation[J]. Computer Application Research, 2022, 39(5):1478-1482.

[7] 王曜,郑列. 一种新的基于聚类的试探性 SMOTE 算法[J]. 重庆理工大学学报(自然科学),2022,36(4):187-195.
WANG Y, ZHENG L. New tentative SMOTE algorithm based on clustering[J]. Journal of Chongqing University of Technology (Natural Sciences),2022,36(4):187-195.

[8] MALDONADO S, VAIRETTI C, FERNANDEZ A, et al. FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification[J]. Pattern Recognition, 2022, 124: 108511.

[9] HE Y L, LU X, FOURNIER-VIGER P, et al. A novel overlapping minimization SMOTE algorithm for imbalanced classification[J]. Frontiers of Information Technology & Electronic Engineering, 2024, 25(9): 1266-1281.

[10] ZHAO P, LUO CH, QIAO B, et al. T-SMOTE: Temporal-oriented synthetic minority oversampling technique for imbalanced time series classification[C]. International Joint Conferences on Artificial Intelligence, 2022: 2406-2412.

[11] YANG X Y, ZHANG ZH G, CUI X, et al. A time series data augmentation method based on dynamic time warping[C]. 2021 International Conference on Computer Communication and Artificial Intelligence. IEEE, 2021: 116-120.

[12] QU H CH, ZHANG ZH. A time series data augmentation method based on SMOTE[C]. 2024 36th Chinese Control and Decision Conference. IEEE, 2024.

[13] BI R Y, ZHAO Y L, MA Y H, et al. Research of pulse position based on gradient pressure method[J]. Biomedical Signal Processing and Control, 2023, 80: 104372.

[14] MOHAMMED R. FCM-CSMOTE: Fuzzy C-means center-smote[J]. Expert Systems with Applications, 2024, 248: 123406.

[15] 赵云龙,李家炜,马宇航,等. 基于三通道脉搏采集系统的血管弹性研究[J]. 电子测量技术,2021,44(24):141-146.
ZHAO Y L, LI J W, MA Y H, et al. Research on arterial stiffness based on three-channel pulse acquisition system[J]. Electronic Measurement Technology,2021,44(24):141-146.

[16] LIANG Y B, CHEN ZH CH, LIU G Y, et al. A new, short-recorded photoplethysmogram dataset for blood pressure monitoring in China[J]. Scientific Data, 2018, 5(1): 1-7.

[17] LUQUE A, CARRASCO A, MARTÍN A, et al. The impact of class imbalance in classification performance metrics based on the binary confusion matrix[J]. Pattern Recognition, 2019, 91: 216-231.

[18] 王宽田,姚江云,唐永忠,等. 基于 SVM 的液压机械驱动齿轮组故障诊断研究[J]. 电子测量技术,2024,47(13):10-17.
WANG K T, YAO J Y, TANG Y ZH, et al. Research on fault diagnosis of hydraulic mechanical drive gear group based on SVM[J]. Electronic Measurement Technology,2024,47(13):10-17.

[19] 魏新园,周京欢,钱牧云,等. 随机森林算法在超声缺陷识别中的应用研究[J]. 电子测量与仪器学报,2024,38(5):47-55.
WEI X Y, ZHOU J H, QIAN M Y, et al. Application of random forest algorithm in ultrasonic defect recognition[J]. Journal of Electronic Measurement and Instrumentation,2024,38(5):47-55.

[20] SHERSTINSKY A. Fundamentals of recurrent neural network(RNN) and long short-term memory(LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.

[21] SADOUK L. CNN approaches for time series classification[J]. Chapters, 2019, DOI: 10.5772/intechopen.81170.

[22] SUN P F, WANG ZH P, JIA L Y, et al. SMOTE-kTLNN: A hybrid re-sampling method based on SMOTE and a two-layer nearest neighbor classifier[J]. Expert Systems with Applications, 2024, 238: 121848.

[23] LIAW L C M, TAN S C, GOH P Y, et al. A histogram SMOTE-based sampling algorithm with incremental learning for imbalanced data classification[J]. Information Sciences, 2025, 686: 121193.

作者简介

李辉,硕士研究生,主要研究方向为中医的客观化与数字化、深度学习。
E-mail:2500782140@qq.com

李振华,博士,讲师,主要研究方向为智慧医疗、数据分析等。
E-mail:lzh734007968@163.com

李瑞杰,硕士研究生,主要研究方向为智慧医疗、中医诊断数字化。
E-mail:2469663626@qq.com

张志东(通信作者),博士,教授,主要研究方向为微纳测试技术、智慧医疗等。
E-mail:zdzhang@nuc.edu.cn

薛晨阳,博士,教授,主要研究方向为新型微纳器件、固体光谱学、智慧医疗等。
E-mail:xuechenyang@nuc.edu.cn