计算机科学
COMPUTER SCIENCE

基于主动重心的青年高血压患者心肺运动时序数据增强

黄昉菀, 卢举鸿, 於志勇

引用本文

黄昉菀, 卢举鸿, 於志勇.基于主动重心的青年高血压患者心肺运动时序数据增强[J]. 计算机科学, 2023, 50(6A): 211200233-11.

HUANG Fangwan, LU Juhong, YU Zhiyong. Data Augmentation for Cardiopulmonary Exercise Time Series of Young HypertensivePatients Based on Active Barycenter [J]. Computer Science, 2023, 50(6A): 211200233-11.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.211200233

基于主动重心的青年高血压患者心肺运动时序数据增强

黄昉菀[1,2]　卢举鸿[1]　於志勇[1,2]
1 福州大学计算机与大数据学院　福州 350116
2 福建省网络计算与智能信息处理重点实验室　福州 350116
(hfw@fzu.edu.cn)

摘　要　精准医疗的逐步兴起,如挖掘青年高血压患者的心肺运动时序数据,可以了解不同个体对有氧运动训练的响应性,有助于提高患者高血压管理计划的制定效率,更有效地实现有氧运动干预的治疗。开展该研究的瓶颈之一在于难以获取充足的样本数据。为了解决获取数据难度大、成本高等问题,利用加权动态时间规整重心平均算法来进行时间序列数据增强,重点针对重心选择和权重分配进行了研究。针对重心选择问题,首次引入了主动重心的概念,提出了代表性重心与多样性重心选择策略,改善了数据增强的效果。此外,针对现有权重分配策略的不足,提出了随机权重距离递减分配策略,避免了合成重复样本,进一步提升了模型的泛化能力。实验结果表明,在该研究背景下同时考虑重心选择与权重分配进行数据增强,可以进一步提升青年高血压患者有氧运动干预疗效预测的准确性。

关键词:高血压;心肺运动实验;时序数据增强;动态时间规整重心平均;重心选择策略;权重分配策略
中图法分类号　TP391

Data Augmentation for Cardiopulmonary Exercise Time Series of Young Hypertensive Patients Based on Active Barycenter

HUANG Fangwan[1,2],LU Juhong[1] and YU Zhiyong[1,2]
1 College of Computer and Data Science,Fuzhou University,Fuzhou 350116,China
2 Fujian Key Laboratory of Network Computing and Intelligent Information Processing,Fuzhou 350116,China

Abstract　The gradual rise of precision medicine,such as mining cardiopulmonary exercise time series of young hypertensive patients,can understand the response of different individuals to aerobic exercise training. This helps to improve the efficiency of hypertension management plan and achieve aerobic exercise intervention more effectively. One of the bottlenecks in this study is that it is difficult to obtain sufficient sample data. To solve the above problem,this paper adopts the weighted dynamic-time-warping barycenter averaging algorithm(WDBA) to realize data augmentation of time series,focusing on the barycenter selection and the weight assignment. In this paper,the concept of active barycenter is introduced for the first time,and the selection strategies of representative barycenter and diversity barycenter are proposed to improve the effect of data augmentation. Furthermore,aiming at the shortcomings of the existing weight assignment strategies,a random strategy with decreasing distance is proposed to further improve the generalization ability of the model by avoiding the synthesis of duplicate samples. Experimental results show that the accuracy of predicting the efficacy of aerobic exercise intervention in young hypertensive patients can be further improved by considering both the barycenter selection and the weight assignment for data augmentation in the background of this study.

Keywords　Hypertension,Cardiopulmonary exercise test,Time series data augmentation,Dynamic-time-warping barycenter averaging,Barycenter selection strategy,Weight assignment strategy

1　引言

近年来,高血压作为一种常见的慢性疾病在年轻群体中的发病率显著上升[1-2]。医学研究表明生活方式的调整,包括控制体重、饮食和运动也可作为治疗高血压药物的替代品,用于降低血压[3-4]。其中,有氧运动不仅可以直接降低血压,还可以间接通过控制体重、减轻压力、改善血管内皮功能,从而达到类似的效果[5-7]。因此,有氧运动干预(Aerobic Exercise Intervention,AEI)已被广泛应用于治疗高血压[8-9]。然而,以治疗高血压为目的的有氧运动干预具体指南尚未被广泛接受。因为相同的运动计划、运动类型、时间、频率和持续时间在降低血压方面存在着显著的个体差异[10-12],所以在制定全面的高血压管理计划之前了解个体对 AEI 的反应将有助于提高血压管理的有效性和效率。

出于临床可行性和实用性的目的,本课题组前期的工作提出了基于患者治疗前心肺运动试验(Cardiopulmonary

基金项目:国家自然科学基金(61772136);福建省中青年教育科研项目(JAT210007)
This work was supported by the National Natural Science Foundation of China(61772136) and Fujian Province Young and Middle-aged Teachers Education Research Project(JAT210007).
通信作者:於志勇(yuzhiyong@fzu.edu.cn)

Exercise Test,CPET)的相关数据,利用机器学习技术预测AEI对年轻高血压患者的疗效[13]。CPET提供了多器官系统功能的综合生理评估,不仅包括心血管和肺,还包括肌肉骨骼和造血系统[14]。它可以帮助临床医生识别疾病的严重程度并评估对治疗的反应,从而在制定有氧运动训练处方和心脏康复方面发挥重要作用[15-16]。CPET利用一种带有许多传感器的电动自行车作为主要测力计,可以测量各种心肺代谢指标随时间的变化。由临床医生指导的具体测试方案通常包括4个阶段:1)休息1min,缓解患者的紧张情绪;2)空载骑行(踏板无阻力)3min进行热身;3)不断增加踏板的阻力(增量为20～35W/min),持续锻炼5～12min,直至精疲力竭;4)3min空载骑行和3min静坐,共计恢复时长为6min。为了提供个体对运动反应的最佳测量数据,个体的多个心肺代谢指标是通过连接到面罩的氧合分析仪在个体每次呼吸时进行收集。由于每一次呼吸代表一个采样点,因此每个代谢指标的信息可以视为一条时间序列。需要说明的是,CPET数据在收集过程中往往存在一定的噪声,因此数据挖掘方法的选择需要考虑对信噪比的鲁棒性。根据前期的工作,本文将采用基于解析字典学习的稀疏表示分类器(Sparse Representation Classifier based on Analytic Dictionary Learning,SRC-AL)来预测AEI对降低血压的功效[17]。

目前开展该研究的瓶颈之一是病人的CPET数据需要在专业医护人员的监督和患者的配合下进行规定的操作并通过专用设备采集,这使得数据的获取变得困难。在传统的机器学习和近年来备受青睐的深度学习中,训练样本对算法模型的重要性不言而喻。模型的性能提升依赖于大量训练数据,而如今面临的一个挑战是数据获取成本过大,难以获取充足的训练样本。为了应对这一挑战,本文提出通过数据增强技术,对高血压患者的CPET数据进行数据生成,旨在进一步提高患者AEI疗效的预测精度。本文的贡献如下:

(1)采用加权动态时间规整重心平均算法(Weighted Dynamic Time Warping Barycenter Averaging,WDBA)对病人的CPET时间序列进行数据增强,进一步提升了SRC-AL对高血压患者AEI疗效预测的准确性。

(2)首次提出了主动重心的概念,提出了代表性重心与多样性重心选择策略,并根据同类样本的数据分布对不同的CPET代谢指标提出了具体的重心选择策略建议。实验结果表明,主动重心挑选策略相比于重心无差别策略(随机重心策略和轮流重心策略)可以有效提升模型的分类精度。

(3)针对WDBA算法现有权重分配策略的不足,提出了随机权重距离递减分配策略。该策略的优点在于既遵循了按距离递减分配权重的原则,又结合了随机权重的灵活性,即便在初始重心相同的情况下,也能确保其他样本获得不一样的权重值,避免了重复样本的生成,进一步提升了模型的泛化能力。

2 CPET代谢指标

CPET提供了不同运动强度水平下与循环、呼吸和气体代谢相关的多个指标的时变信息。专业临床医生为预测高血压患者AEI疗效推荐了以下6个指标:

(1)氧脉冲(Oxygen Pulse,OP):每次心搏时的摄氧量。它被定义为每分钟身体摄入的氧气量(Oxygen volume,VO_2)与心率(Heart Rate,HR)的比值(即VO_2/HR),单位为毫升每跳(mL/beat)。更高的氧脉冲表明更好的心肺功能,可作为判断心肺功能的综合指标。

(2)每千克体重氧耗量(VO_2/kg):每千克体重每分钟消耗的氧气毫升数,单位为毫升每分钟(mL/min)。它反映了机体对氧气的利用能力,通常由最大心输出量、动脉含氧量、心输出量对运动肌肉的分布指数、肌肉氧容量等决定。

(3)心输出量(Cardiac Output,CO):在给定的一段时间内流出心脏的血液量,通常表示为升每分钟(L/min)。它会随着肌肉运动、情绪激动、怀孕等而增加。

(4)每分钟通气量(Minute Ventilation Volume,VE):一分钟内从肺部吸入或呼出的空气量,可用每次呼吸时吸入或呼出的空气量乘以呼吸频率得到,静息时一般为6～8L/min,又称为肺通气量。

(5)呼吸气体交换率(Respiratory exchange ratio,R):每分钟二氧化碳(CO_2)排出量与氧气(O_2)摄取量的比值(VCO_2/VO_2)。它不仅反映了气体组织代谢的交换,还反映了储气过程中瞬态变化的影响。

(6)二氧化碳通气当量(VE/VCO_2):每分钟通气量与每分钟二氧化碳排出量的比值,反映了身体排出二氧化碳的能力。

3 相关工作

当某种应用场景只能提供有限的训练数据时,数据增强作为扩充训练数据量和提升模型鲁棒性的一种途径,已经被大量研究证明了其有效性。在计算机视觉领域,数据增强的相关研究已经十分成熟,常见的方法有空间几何变换、像素颜色变换和模糊等[18]。然而,时序数据却无法轻易地控制这种特殊转换对时间特性的影响。鉴于时间序列数据增强研究相对于其他领域尚未成熟,研究者们随之展开了一系列针对时序数据的数据增强研究工作。Le Guennec等提出从原始时间序列中随机提取连续切片进行拉伸或收缩,以生成新的时序样本[19]。此外,还可以从时间序列的成分出发进行数据增强。STL分解技术(Seasonal and Trend decomposition using Loess)可将时间序列分解成趋势、季节性和残差[20]。Bergmeir等建议将残差应用Boot Strapping算法来生成增强信号,然后将这些信号与趋势和季节性相加,以组合新的时间序列[21]。Kegel等系统地回顾了生成时序数据的方法,并且建议使用某种相似性度量来评估不同方法生成的时序数据的性能,提出了基于重组的生成时间序列新方法,根据度量规则衍生出的随机性规则重组STL分解成分生成新的时间序列[22]。目前,也有一些工作逐渐开始从时间序列频域角度研究时间序列的数据增强。Gao等提出利用时间序列频域中幅度谱和相位谱的扰动来增强卷积神经网络的时间序列异常检测中的数据。实验结果表明,基于幅度谱和相位谱的扰动的数据增强与上述方法的结合,对时间序列异常检测任务有明显的改进效果[23]。然而,数据增强方法与数据挖掘任务紧密相关,适用于时间序列异常检测的数据增强方法对于时间序列分类任务可能无效[24]。上述数据增强方法对时间序列分类任务的适用性是未知的。近年来,随着GAN框架在许多领域得到了广泛的关注,也有人开始研究使用GAN框架进行数据增强[24]。Esteban等提出了循环GAN(RGAN)和

循环条件 GAN（RCGAN）来生成逼真的多维时间序列数据[25]。RGAN 在生成器和判别器中采用 Recurrent Neural Network（RNN），而 RCGAN 采用以辅助信息为条件的两个 RNN。结合时间序列数据的特性，适用于生成不同领域时序数据的 TimeGAN 模型被提出[26]。该模型通过有监督和无监督损失的学习嵌入空间进行对抗性和联合训练。实验结果表明，TimeGAN 模型在生成时间序列的相似性方面相比于其他先进方法具有明显的优势。然而，训练数据较为匮乏的情况可能导致难以使用深度学习模型进行数据增强。

针对时间序列分类任务，基于动态时间规整（Dynamic Time Warping，DTW）的最邻近分类器（1NN-DTW）已经取得了良好的分类效果[27]。为了进一步提升该分类器在训练数据不足时的性能，Petitjean 等新颖地提出了加权动态时间规整重心平均算法 WDBA，先随机选取重心再利用不同的权重分配策略生成多条平均时间序列用于扩充训练集[28]。基于 UCR（University of California，Riverside）时间序列分类档案库的 85 个数据集的广泛实验证明了 WDBA 在训练样本数量有限（例如，每类只有 2 到 6 个训练样本）时特别有用[29]。针对本文的研究背景，SRC-AL 作为最邻近子空间分类器的泛化，已经被证明可以取得比 1NN-DTW 更好的分类效果[17]。由此可分析得到，在高血压患者 CPET 数据线性可分的情况下，WDBA 算法也将适用于本文背景下的时间序列数据增强。后续的实验也证明了该算法对本文应用背景的适用性。

4 CPET 数据增强

4.1 加权动态时间规整重心平均算法

定义 1 向量 $\boldsymbol{T}=[t_1,t_2,\cdots,t_l]^{\mathrm{T}}$ 表示一条长度为 l 的时间序列，其中 $t_j(j=1,2,\cdots,l)$ 为 j 时刻的值。

定义 2 $D=\{\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n\}$ 表示具有相同标签（即同类）的 n 条时间序列的集合。

定义 3 向量 $\bar{\boldsymbol{T}}$ 表示 D 中所有时间序列的平均时间序列，它可以通过在基于 D 形成的动态时间扭曲空间 E 中求解以下最小化问题得到：

$$\arg\min \sum_{i=1}^{n} DTW(\bar{\boldsymbol{T}},\boldsymbol{T}_i) \quad \text{s.t. } \bar{\boldsymbol{T}}\in E \tag{1}$$

本文采用 DTW 重心平均算法 DBA（DTW Barycenter Averaging）作为求解式（1）的方法[30]。DBA 采用期望最大化方案，迭代地更新初始（可能任意的）平均序列（即初始重心），计算其与平均序列 $\bar{\boldsymbol{T}}$ 的平方 DTW 距离。DBA 在每次迭代中都包含以下两个阶段：

（1）计算 D 中每个时间序列 $\boldsymbol{T}_i(i=1,2,\cdots,n)$ 与上一次迭代得到的平均序列之间的 DTW 距离，以找到它们的多重对齐 $\boldsymbol{A}_i=[a_1^i,a_2^i,\cdots,a_l^i]^{\mathrm{T}}$。该向量可用于表示两者的共性信息，以判断两者的相似性或同源性[31]，其求解方法详见文献[28]。

（2）基于所有的 $\boldsymbol{A}_i(i=1,2,\cdots,n)$，更新平均序列。

本文将利用 DBA 经过 k 次迭代得到的平均序列表示为 $\bar{\boldsymbol{T}}_k=[\overline{t_1^k},\overline{t_2^k},\cdots,\overline{t_l^k}]^{\mathrm{T}}$。需要说明的是，当给定一个初始平均序列和迭代次数后，利用 DBA 只能得到一条平均时间序列。改变这个初始时间序列不足以在合成数据集中创建足够的多样性。因此，加权平均时间序列的概念被提出用于解决该问题。

定义 4 加权平均时间序列 $\bar{\boldsymbol{T}}'$：给定一系列经过归一化处理的权重 $W=\{w_{T_1},w_{T_2},\cdots,w_{T_n}\}$，$\bar{\boldsymbol{T}}'$ 可以通过在基于 D 形成的动态时间扭曲空间 E 中求解以下最小化问题得到：

$$\arg\min \sum_{i=1}^{n} w_{T_i} DTW(\bar{\boldsymbol{T}},\boldsymbol{T}_i) \quad \text{s.t. } \bar{\boldsymbol{T}}\in E \tag{2}$$

求解式（2）可采用加权 DTW 重心平均算法 WDBA，它与 DBA 的区别在于其需要对每一个 $\boldsymbol{A}_i(i=1,2,\cdots,n)$ 分配权重，即每个 $\boldsymbol{A}_i$ 对更新平均序列的贡献是不同的，其伪代码如算法 1 所示。本文将利用 WDBA 经过 k 次迭代得到的平均序列表示为 $\bar{\boldsymbol{T}}_k'$。

算法 1 WDBA

输入：时间序列集 D=$\{\mathbf{T}_1,\mathbf{T}_2,\cdots,\mathbf{T}_n\}$，$\mathbf{T}_i$ 表示集合 D 中的一条时间序列数据，长度均为 l；经过归一化的权重集合 W=$\{w_1,w_2,\cdots,w_n\}$ 以及迭代次数 k

输出：一条加权平均时间序列 $\bar{\mathbf{T}}'_k$

步骤 1 选取初始平均时间序列 $\bar{\mathbf{T}}_0'$（$\bar{\mathbf{T}}_0'$ 由重心选择策略得到）；

步骤 2 根据权重分配策略和 $\bar{\mathbf{T}}_0'$，利用集合 W 对 D 中每条时间序列分配权重，可得加权后的时间序列集合：

D′=$\{(\mathbf{T}_1,w_{T_1}),\cdots,(\mathbf{T}_i,w_{T_i}),\cdots,(\mathbf{T}_n,w_{T_n})\}$

$w_{T_i}(i=1,2,\cdots,n)\in W$；

步骤 3 利用 D′对 $\bar{\mathbf{T}}_0'$ 进行 k 次迭代更新；

```
for p=1 to k do
  初始化向量：
  T′=[t1′,…,tj′,…,tl′]^T=[0,…,0,…,0]^T；
  初始化向量：
  S=[s1,…,sj,…,sl]^T=[0,…,0,…,0]^T；
  for i=1 to n do
    计算平均时间序列 T̄′p−1 与 D 中时间序列 Ti 的多重对齐 Ai；
      for j=1 to l do / * 多重对齐叠加 * /
          tj′=ti′+a_j^i · wTi；
          si=si+|a_j^i| · wTi；
      end
  end
  for j=1 to l do / * 归一化得到最终序列 * /
    tj′=tj′/sj；
  end
  T̄′p=T′；
end
```

步骤 4 返回经过 k 次更新的加权平均时间序列 $\bar{\mathbf{T}}'_k$。

对 WDBA 算法而言，步骤 1 的重心选择和步骤 2 的权重分配对数据增强效果至关重要。现有的研究工作中重心选择均采用随机重心策略，即从时间序列集合 D 中随机挑选一条作为初始重心。这种做法的缺点在于数据增强效果受随机性影响较大。当初始重心选择不当（如选择到离群点）时，有限次数的迭代更新无法保证能得到较好的平均时间序列。而在权重分配问题上，Forestier 等提出了 3 种权重分配策略，分别是全局平均策略（Average All，AA）、局部平均策略（Average Selected，AS）和局部平均距离递减策略（Average Selected with Distance，ASD）[29]。上述策略要么未考虑样本间的距离

差异，要么容易生成重复样本。针对现有策略的不足，本文对重心选择问题和权重分配问题进行了深入的研究。

4.2 重心选择

4.2.1 随机重心策略

Forestier 等通过计算同类时间序列集合 D 中每个样本与其余样本的加权 DTW 距离总和，找到具有最小距离总和的样本作为初始重心，其伪代码详见算法 2。由于权重集合 W 是随机产生的，此做法的效果相当于从 D 中任选一条时间序列作为初始重心。此方法的效率不高，因为所有的样本被选中的概率都是相同的。这可能会造成一些质量较差的样本被选中，甚至是重复选中，从而影响新生成样本的质量。

算法 2 随机重心策略

输入：同类时间序列集合 $D=\{\mathbf{T}_1,\mathbf{T}_2,\cdots,\mathbf{T}_n\}$，$T_i$ 表示一条时间序列数据；随机权重集合 $W=\{w_1,w_2,\cdots,w_n\}$

输出：初始重心 $\bar{\mathbf{T}}_0'$

步骤 1 初始化最小距离 minDistance=+∞；

步骤 2 更新最小距离 minDistance：

```
for i=1 to n do
  tmp=0;/*临时变量*/
  for j=1 to n do
    tmp+=w_i*(DTW(T_i,T_j))^2;
  end
  if tmp<minDistance then
    T̄0'=T_i;
    minDistance=tmp;
  end
end
```

步骤 3 返回初始重心 $\bar{\mathbf{T}}_0'$。

4.2.2 轮流重心策略

在数据增强过程中，若采用随机重心策略的方式生成新样本，有可能出现某个样本被重复选中，而有些样本从未被选中的情况。为了避免此现象，本文首先提出了轮流重心策略，即 D 中的每条样本轮流作为初始重心用于数据增强。这种策略仍然没有区分样本之间的质量差异，只是避免了质量较差样本被重复选中的情况，但其代价是质量较好的样本同样无法被重复选中。如何能够在避免选中质量较差样本的同时，又能重复选中质量较好的样本呢？针对此问题，本文借鉴主动学习技术来主动挑选合适的样本作为初始重心。

主动学习技术最早是应用于训练样本的主动挑选，研究的问题是在成本固定的情况下，如何挑选出信息量最高的样本进行标记，以最大化提升模型的学习能力[32]。衡量样本信息量的指标主要包括不确定性、代表性、多样性等。针对数据增强的应用背景，本文提出了代表性重心策略和多样性重心策略，旨在寻找最有利于数据增强的样本作为初始重心。

4.2.3 代表性重心策略

在数据集中，样本的代表性可以根据与之相似或接近的样本数量来评估。显然，具有高代表性的样本不太可能是异常样本。将其作为初始重心，能够确保生成的样本在保留随机性的同时，与同类别样本更相似。将这些基于代表性生成的样本添加到训练集中，避免了数据增强过程中产生具有迷惑性的样本，让分类器无法对其做出正确的判断。

根据上述思想，本文提出的代表性重心策略首先对 D 中的样本进行聚类操作，然后将拥有最多成员的簇作为基于代表性的初始重心集合 C_R，最后将 C_R 中的样本轮流作为初始重心进行数据增强。该策略的实现过程详见算法 3，难点之一在于如何度量样本间的相似度。对于拥有高维度、时间顺序特性的数据来说，采用欧氏距离的方式进行相似度的度量往往存在着较大的偏差。因此，该策略使用 DTW 距离作为质心计算以及迭代更新的依据。在质心计算上，当数据集存在孤立点或者数据集分布差异较大时，采用简单的取均值的方式得到样本的质心将大大地影响质心的准确性。同时，随着算法 3 中步骤 4 迭代次数增多，反映某类样本特性的质心的偏差以及相似性度量带来的误差也将不断增大。因此，该策略采用 K-Medoids 算法而不是 K-Means 算法进行聚类，即不采用簇中样本的平均值作为质心，而选用簇中的样本与其他样本相似性最大的某个样本作为簇的质心。最后需要说明的是，基于代表性的初始重心集合 C_R 处于整个训练集同类样本分布最密集的地方，是最能代表这一类训练样本集合的样本。但是，C_R 的大小也受 k 的影响，k 的范围应在 $[1,|D|)$ 之间，其中 $|D|$ 表示集合 D 的大小。当 $k=1$ 时，C_R 即为 D 本身，该策略将退化为轮流重心策略。

算法 3 代表性重心策略

输入：同类时间序列集合 $D=\{\mathbf{T}_1,\mathbf{T}_2,\cdots,\mathbf{T}_n\}$，$T_i$ 表示一条时间序列数据；簇的个数 k

输出：基于代表性的初始重心集合 C_R

步骤 1 随机从集合 D 中选 k 个样本作为初始质心；

步骤 2 计算集合 D 中的每个样本与 k 个质心的相似度，并将其归为最相似质心所属簇中；

步骤 3 更新每个簇的质心：计算每个簇内样本与其他样本的总相似度，总相似度最大的样本为簇的新质心；

步骤 4 重复步骤 2 和步骤 3 直至每个簇的质心不再变化；

步骤 5 将拥有最多成员的簇以及簇内样本作为基于代表性的初始重心集合 C_R。

4.2.4 多样性重心策略

与代表性重心不同，对于同类的样本集合 D，多样性重心选择更偏向于样本之间差异较大的情况，即同类样本之间较为分散，存在着簇内成员数量较为均衡的几个簇。这种情形下，重心的选择方式如果考虑代表性，重心样本的选择只偏向于某个簇，那么合成的样本也会和该簇的样本更为相似，从而陷入样本的局部生成。这对于其他簇来说也许是“不公平的”。因此，出于数据集的全局考虑，若同类样本中存在着几个分散的簇，也就意味着该类样本存在的特征空间较为广泛。从多样性重心的角度出发，选择同类数据集中属于不同特征子空间的样本作为重心，对于该类样本在机器学习模型中的训练会有更大的帮助。根据上述思想，本文提出的多样性重心策略同样首先对 D 中的样本进行聚类操作，然后在每个簇中挑选最接近质心（含质心）的部分样本加入基于多样性的初始重心集合 C_D 中，最后将 C_D 中的样本轮流作为初始重心进行数据增强。该策略的实现过程详见算法 4。需要注意的是，C_D 的大小也受 k 的影响，此时 k 的范围应在 $(1,|D|]$ 之间。当 $k=|D|$ 时，C_D 即为 D 本身，该策略将退化为轮流重心策略。

算法 4 多样性重心策略

输入：同类时间序列集合 $D=\{\mathbf{T}_1,\mathbf{T}_2,\cdots,\mathbf{T}_n\}$，$\mathbf{T}_i$ 表示一条时间序列数据；簇的个数 k

输出：基于多样性的初始重心集合 C_D

步骤 1　随机从集合 D 中选 k 个样本作为初始质心；

步骤 2　计算集合 D 中的每个样本与 k 个质心的相似度，并将其归为最相似质心所属簇中；

步骤 3　更新每个簇的质心：计算每个簇内样本与其他样本的总相似度，总相似度最大的样本为簇的新质心；

步骤 4　重复步骤 2 和步骤 3 直至每个簇的质心不再变化；

步骤 5　将每个簇的质心或每个簇中最近似质心的一些样本加入 C_D。

4.2.5　CPET 代谢指标的重心策略选择建议

综上所述，本文提出的两种初始重心主动选择策略与同类样本的数据分布存在密切的联系。当同类样本的数据分布较为集中时，应采用代表性重心策略进行初始重心选择，避免生成新的孤立样本。当同类样本的数据分布较为分散时，应采用多样性重心策略，避免仅选择某个簇而忽略了其他簇的信息。由于不同的 CPET 代谢指标具有不同的数据分布特点，本文首先利用可视化手段直观地展示了 24 名高血压患者基于治疗前某个 CPET 指标在二维空间的样本分布情况（见图 1）。这些代谢指标的一个显著特点是维度高，因此本文采用了 T 分布随机近邻嵌入（T-Distribution Stochastic Neighbor Embedding，T-SNE）算法[33]对患者的代谢指标数据进行降维处理。T-SNE 具有保持局部结构的能力，这意味着高维数据空间中距离相近的点投影到低维中仍然相近。对于每个高血压患者，可将其分为 AEI 疗效反应强或疗效反应弱两个类别，具体内容详见第 5 节的数据集描述。图中橙色的“圆”表示疗效反应强的样本，蓝色的“三角形”则表示疗效反应弱的样本。需要注意的是，在数据增强的过程中，本文采用的方式是按类别生成新样本。因此在进行数据增强之前，更应该关注同类样本之间的分布情况。

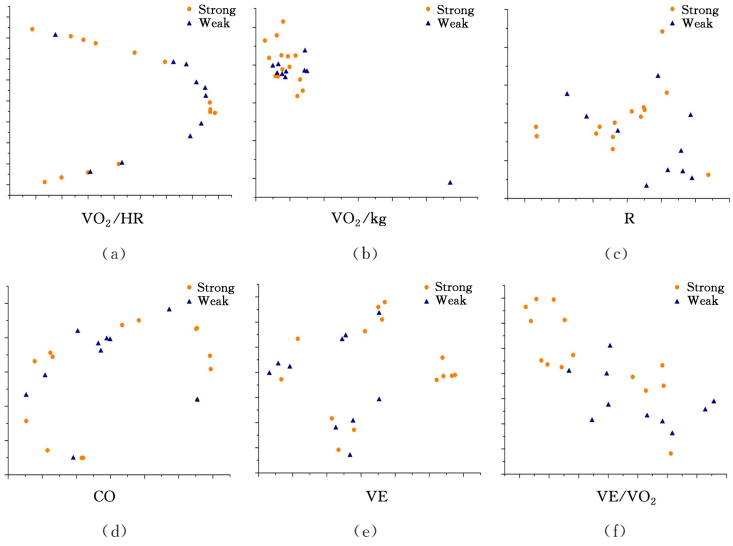

图 1　24 名高血压患者基于治疗前某个 CPET 代谢指标在二维空间的样本分布情况（电子版为彩图）

Fig. 1　Sample distribution of 24 hypertensive patients in two-dimensional space based on CPET metabolic index before treatment

与图 1(a)－图 1(c)不同，图 1(d)－图 1(f)呈现出截然不同的数据分布特点。从同类样本的分布可以看出，不论是 AEI 疗效反应强还是反应弱的高血压患者，其样本之间都较为分散，并且明显地聚集成了几个簇，这些簇的样本数量相对均衡。此时再考虑代表性重心选择策略可能会陷入难以确定重心集的局面。无论选择哪个簇作为代表性重心集都将导致“不公平”现象的产生。对于拥有分散簇的代谢指标 CO，VE 和 VE/VCO_2 而言，在对它们进行数据增强时，本文建议采用多样性重心策略。

4.3　权重分配

进行数据增强时，WDBA 算法最关键的两个问题是重心选择与权重分配。在根据重心选择策略得到初始重心后，再根据权重分配策略对其和剩余同类样本进行权重分配，从而生成多种多样的新样本。目前已有的权重分配策略包括：随机权重策略、最近邻优先策略以及固定权重距离递减策略。本文在分析上述策略的不足后，提出了一种新的权重分配策略——随机权重距离递减策略。

4.3.1　现有的权重分配策略

(1)随机权重策略

当要生成一定数量的样本时，可以通过对权重的调整来实现生成样本的多样性。一种最直接的方式是对参与生成样本的训练集 D 中的样本赋予随机的权重。Forestier 等建议参照单位浓度参数为 $\bar{w}\sim Dir(\bar{1})$ 的平坦狄利克雷分布对权重进行采样，并为狄利克雷分布使用了 Gamma 分布随机变量的形状参数较低的值（设置为 0.2），以便赋予时间序列更多的权重[29]。全局平均策略（AA）使用的就是狄利克雷分布权重采样策略。AA 策略中，除了随机初始重心被赋予最大权重外，同类样本训练集 D 中的剩余序列都被随机赋予不同的权重，这使得在很多情况下生成的新时间序列不受原数据分布的约束。如图 1(a)所示的高血压患者 CPET 代谢指标 VO_2/HR 样本在特征空间的分布呈“U”型时，全局平均将很可能填满原始数据的凹槽部分（即“U”型的底部）。若两类数据分别集中在“U”型的两侧，出现在底部的样本则均属于难以区分的样本。因此，在样本分布为“流型”时，将不建议采用随机权重策略对样本进行权重分配。

(2)最近邻优先策略

为了应对样本的“流型”分布情况，在进行权重分配时应考虑初始重心与其余样本的距离远近。因此，可以采用最近邻的思想，将大部分权重集中于初始重心的周围的样本，对初始重心较远的样本赋予较小的权重，从而避免它们为生成样本带来较大的影响。最近邻优先策略的一种具体做法是局部平均策略（AS）。首先为重心分配 0.5 的权重值，其次从

该重心最近邻的5个样本中随机取2个样本各赋予0.15的权重值，剩余的0.2权重均分至剩余的时序数据。最近邻优先权重分配策略较为合理地将大部分权重值集中在重心周围的样本，避免了“流型”的样本集合因为随机权重分配策略出现的不合理填补情况。但是其对剩余的0.2权重的分配并没有考虑样本与初始重心之间的距离。

(3)固定权重距离递减策略

在最近邻优先的基础上，考虑数据分布的同时，还可以进一步兼顾邻居而言与重心之间的相对距离。对于初始重心的众多邻居而言，它们与重心之间的距离有远有近，对于距离越近的邻居应当具有更大的权重值，反之，对应的样本权重应该越小。局部平均距离递减策略(ASD)通过重心与其最近邻样本的距离评估重心与其他样本的疏密情况。具体的做法是为初始重心赋予1的权重，然后为该重心的最近邻居分配0.5的权重值，并为其余样本定义了权重分配的指数衰减函数：

$$W_{T_i}=e^{\ln(0.5)*\frac{DTW(\boldsymbol{T}_i,\boldsymbol{T}_0)}{d_n}} \tag{3}$$

其中，d_n 表示初始重心 $\bar{\boldsymbol{T}}_0'$ 与最近邻居的DTW距离。由式(3)可知，当重心确定后，d_n 是一个固定的值，样本得到的权重只与样本到重心的距离有关。由于其他样本到该重心的距离是固定的，所以当初始重心被重复选择时，利用该策略生成的时间序列必然出现重复。原始的训练数据越少，时间序列的重复生成概率越大，这将降低WDBA算法的效率，因而可能导致数据增强后的效果受到限制。

4.3.2 随机权重距离递减策略

由于AS与ASD的权重都是基于距离进行分配，在初始重心选定后，其他时间序列与重心的距离是固定的，使得分配的权重也是固定值。同类样本集合 D 的大小与生成样本的数量决定了初始重心的重复率，从而决定了重复生成样本的概率。当重心可选的集合缩小时，这两种策略重复生成样本的概率将大大增加。在选择具有代表性或多样性的样本作为重心的同时，为了克服兼顾重心与其余样本的相对距离带来的重复生成样本问题，本文提出了利用随机权重按距离递减分配(Random Selected with Distance，RSD)的新策略。对于同类样本集合 D，该策略的做法是首先对平坦狄利克雷分布进行随机采样得到 n 个权重，归一化后将它们从大到小排序，构成权重集 $W'=\{w_1',w_2',\cdots,w_n'\}$，其中 $w_1'>w_2'>\cdots>w_n'$。其次，对初始重心 $\bar{\boldsymbol{T}}_0'$ 分配最大权重 w_1'。最后，对剩余样本按照其与 $\bar{\boldsymbol{T}}_0'$ 的DTW距离远近分配剩余权重，距离越近分配越大权重，距离越远分配越小权重。该策略的实现过程详见算法5。该策略的优点在于既遵循了按距离递减分配权重的原则，又结合了随机权重的灵活性，即便在初始重心相同的情况下，也能确保其他样本获得不一样的权重值，避免了重复样本的生成。

算法5 随机权重距离递减策略

输入：同类时间序列集合 $D=\{\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n\}$，$\boldsymbol{T}_i$ 表示一条时间序列数据

输出：加权后的时间序列集合：

$D'=\{(\boldsymbol{T}_1,w_{T_1}),\cdots,(\boldsymbol{T}_i,w_{T_i}),\cdots,(\boldsymbol{T}_n,w_{T_n})\}$；

步骤1 利用狄利克雷分布进行随机采样得到n个权重，归一化后将它们降序排列构成权重集 $W'=\{w_1',w_2',\cdots,w_n'\}$ $w_1'>w_2'>\cdots>w_n'$。

步骤2 计算初始重心 $\bar{\boldsymbol{T}}_0'$ 到D中所有样本的相对距离 $DTW(\bar{\boldsymbol{T}}_0',\boldsymbol{T}_i)$。

步骤3 对n个 $DTW(\bar{\boldsymbol{T}}_0',\boldsymbol{T}_i)$ 进行从小到大排序，记 u_i 为排序序号。

步骤4 按照排序序号分配权重

```
for i=1 to n do
    w_Ti = w'_ui
end
```

步骤5 得到所有时刻的序列与权重的映射对 $(\boldsymbol{T}_i,w_{T_i})(i=1,2,\cdots,n)$。

为了直观地体现上述4种权重分配策略的差异，本文以代谢指标 VO_2/HR 为例，首先采用统一的重心策略(如随机重心)选定初始重心，然后分别利用上述4种权重分配策略进行1:1的数据增强，最后将真实样本和生成样本仍然利用T-SNE可视化技术将它们投影到二维空间(见图2)。图中的“空心圆”是基于“橙圆”(强反应者的真实样本)的生成样本；而“空心三角”是基于“蓝三角”(弱反应者的真实样本)的生成样本。选择 VO_2/HR 的原因是基于该指标的真实样本在二维空间的数据分配呈现比较特殊的“U”型结构(见图1(a))，该形状对权重分配策略较为敏感[29]。图2(a)是利用AA策略得到的数据分布图，可以发现生成样本较多地出现在“U”型的底部，这个区域属于分类器较难识别的样本，因此按该样本进行数据增强，对分类器性能的提升帮助较小。图2(c)是利用ASD策略得到的数据分布图，可以发现该图中“空心圆”和“空心三角形”的个数最少，说明利用该策略，当初始重心出现重复选择时，会生成重复样本。图2(b)是利用AS策略得到的数据分布，可以发现生成的数据较为密集，这是最近邻权重分配策略的特点。图2(d)是利用RSD策略得到的数据分布，可以发现该策略不仅解决了AA策略带来的问题，而且生成的数据在合理的范围内相比AS策略更加分散，这也意味着RSD策略为生成样本带来了更多的随机性。

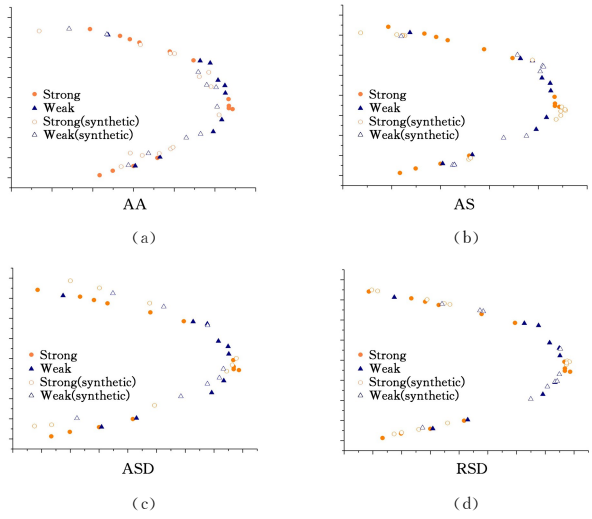

图2 基于VO2/HR的真实样本和生成样本在4种权重分配策略下的数据分布情况(以随机重心选择策略为例)

Fig. 2 Data distribution of real samples and generated samples based on VO_2/HR under four weight allocation strategies (taking random barycenter selection strategy as an example)

4.4 数据增强方案汇总

根据算法1描述的WDBA执行流程不难发现，重心选择和权重分配采用何种策略将会直接影响生成的平均时间序列的质量。在重心选择问题上，本文在现有随机重心策略

(Random Barycenter，RAN)的基础上，提出了轮流重心策略(Alternate Barycenter，ALT)和两个主动重心策略(代表性重心策略(Representative Barycenter，REP)和多样性重心策略(Diversiform Barycenter，DIV))。而在权重分配上，本文在现有的随机权重策略(AA)、最近邻优先策略(AS)以及固定权重距离递减策略(ASD)的基础上，提出了随机权重距离递减策略(RSD)。由于权重分配依赖重心选择所得的重心，因此根据重心选择策略和权重分配策略的不同组合，可得到 16 种基于 WDBA 的数据增强方案，其命名和说明如表 1 所列。

表 1　基于 WDBA 算法的 16 种数据增强方案

Table 1　16 Data augmentation schemes based on WDBA algorithm

编号	方案名	重心选择策略	权重分配策略
1	RAN-AA[23]	随机	随机
2	RAN-AS[23]	随机	最近邻优先
3	RAN-ASD[23]	随机	固定权重距离递减
4	RAN-RSD	随机	随机权重距离递减
5	ALT-AA	轮流重心	随机
6	ALT-AS	轮流重心	最近邻优先
7	ALT-ASD	轮流重心	固定权重距离递减
8	ALT-RSD	轮流重心	随机权重距离递减
9	REP-AA	代表性重心	随机
10	REP-AS	代表性重心	最近邻优先
11	REP-ASD	代表性重心	固定权重距离递减
12	REP-RSD	代表性重心	随机权重距离递减
13	DIV-AA	多样性重心	随机
14	DIV-AS	多样性重心	最近邻优先
15	DIV-ASD	多样性重心	固定权重距离递减
16	DIV-RSD	多样性重心	随机权重距离递减

5　CPET 数据增强

本文的实验采用了来自 24 名 I 期高血压年轻患者(收缩压 SBP：140～160 mmHg，舒张压 DBP：90～100 mmHg)AEI 治疗前的 CPET 数据，该数据由中山大学第一附属医院心内科提供。全体患者的整个锻炼过程都是在医院专业医护人员的监督下完成的。使用动态血压监测和运动血压结果评估运动前后的血压。

5.1　数据集描述

(1)入组标准：年龄 18-45 岁；未经药物治疗或用降压药 2 周以上的 I 期高血压患者；入组前 4 个月无定期运动；能坚持随访半年以上。

(2) 治疗处方：患者使用意大利 COSMED K4 电动自行车进行有氧运动。训练强度对应的任务代谢当量(Metabolic Equivalent，MET)为患者最大耗氧量(VO_{2max})的 70%。每周进行有氧运动 5 次，每次 45 分钟(运动强度相当于每周 2000～3000 千卡)，持续 12 周。

(3) 分类标准：根据治疗效果将患者分为 AEI 治疗的强反应者或弱反应者。分类过程如下：

1)所有患者在 AEI 前后均进行 24 小时动态血压监测以获得其平均血压。

2)计算每个患者治疗前后的血压变化率：

$$r_i=\frac{\|MBPB-MBPA\|}{MBPB} \tag{4}$$

其中，$MBPB$ 表示治疗前 24 小时的平均血压(SBP＋DBP)，$MBPA$ 则表示治疗后 24 小时的平均血压。

3)对患者的血压变化率进行 Z-SCORE 标准化：

$$Z_i=\frac{r_i-\mu}{\sigma} \tag{5}$$

其中，μ 和 σ 分别是平均值和标准差。

4)依据 Z_i 判断第 i 位患者的降压疗效。当 $Z_i>0$，则第 i 位患者被认定为 AEI 强反应者；反之 $Z_i<0$，则患者被视为 AEI 弱反应者。24 位患者的真实标签详情如表 2 所列。从表 2 中可以发现，除最后 1 例患者外，其余患者在经过为期 12 周的有氧运动训练后均显示出一定的降压效果，平均降压变化率为 7.582%。降压效果最好的患者(编号 1)在 AEI 治疗后血压下降达到 40 mmHg。但部分个体血压无明显变化，也证明 AEI 在高血压患者中的疗效存在显著差异，因此本文的研究是非常有意义的。

表 2　24 名 I 期高血压年轻患者的真实标签

Table 2　Real labels of 24 young patients with stage I hypertension

样本编号	MBPB/mmHg	MBPA/mmHg	r_i/%	Z_i	标签
1	242	202	16.529	2.209	Strong
2	229	195	14.847	1.792	Strong
3	221	192	13.122	1.365	Strong
4	241	212	12.033	1.094	Strong
5	235	213	9.362	0.432	Strong
6	253	230	9.091	0.365	Strong
7	223	203	8.969	0.334	Strong
8	249	227	8.835	0.301	Strong
9	209	191	8.612	0.246	Strong
10	244	223	8.607	0.245	Strong
11	214	196	8.411	0.196	Strong
12	246	226	8.130	0.127	Strong
13	204	188	7.843	0.055	Strong
14	244	225	7.787	0.041	Strong
15	244	226	7.377	−0.060	Weak
16	231	214	7.359	−0.065	Weak
17	240	223	7.083	−0.133	Weak
18	231	215	6.926	−0.172	Weak
19	214	207	3.271	−1.079	Weak
20	221	214	3.167	−1.104	Weak
21	222	216	2.703	−1.220	Weak
22	211	208	1.422	−1.537	Weak
23	207	205	0.966	−1.650	Weak
24	207	208	0.483	−1.770	Weak

5.2　实验结果分析

本课题组的前期工作表明基于解析字典学习的稀疏表示分类器(SRC-AL)得益于其良好的特征提取能力和抗噪能力，非常适合作为 CPET 代谢指标时间序列的分类器[17]。因此，本文的实验将基于该分类器分析不同的数据增强方案对其分类精度的影响。实验采用四折交叉验证法，即将样本随机分成 4 份，其中任意 3 份作为训练样本并进行 1 倍的数据增强(即生成样本个数＝原始样本个数)，剩余 1 份作为测试样本。依次记录测试样本在 16 种数据增强方案下的结果，最后的实验结果是多次实验的综合表现。重复进行 6 次实验，以避免数据集划分的随机性对实验结果的影响。

(1)重心选择策略对数据增强效果的影响

为了评估重心选择对数据增强效果的影响，本文首先在统一权重分配策略的基础上，对高血压患者的 6 个代谢指标分别按照不同的重心选择策略进行实验。实验结果如表 3－表 6 所列。当权重分配采用 AA 策略时，由表 3 可知，采用随机重心进行数据增强后，除了 VO_2/HR 之外，其余指标的分类精度都比无数据增强时有所提升。轮流重心的表现不如随机重心，但代表性重心和多样性重心的表现都要优于随机重心，其中多样性重心的分类精度相比于无数据增强时平均提升了 10.5%，代表性重心平均提升了 9.19%，而随机重心

平均提升了 5.60%，轮流重心则仅提升了 3.42%。

当权重分配采用 AS 策略时，由表 4 可知，轮流重心的表现优于随机重心，但仍然不如代表性重心和多样性重心。此时，代表性重心的表现最优，其分类精度相比于无数据增强时平均提升了 8.72%，代表性重心平均提升了 7.90%，轮流重心平均提升了 6.05%，随机重心则仅提升了 4.99%。

当权重分配采用 ASD 策略时，由表 5 可知，随机重心和轮流重心的表现相当，分类精度相比于无数据增强时分别平均提升了 5.05%和 5.01%，而基于主动挑选的代表性重心和多样性重心则分别平均提升了 7.07%和 7.56%。

当权重分配采用 RSD 策略时，由表 6 可知，随机重心的表现略优于轮流重心，但远不如代表性重心和多样性重心。其中多样性重心的分类精度相比于无数据增强时平均提升了 9.99%，代表性重心平均提升了 9.28%，而随机重心平均提升了 5.33%，轮流重心则仅提升了 4.19%。

综上所述，可以得到以下结论：1）基于主动挑选的重心策略明显优于随机重心或轮流重心策略，这说明并非所有的训练样本都适合作为初始重心；2）正如第 4.2.5 节所建议的，具有不同数据分布特点的代谢指标应采用不同的主动重心策略。图 1(a)－图 1(c)所给出的 VO_2/HR，VO_2/kg 和 R 的同类样本分布较为集中，所以应采用代表性重心策略。实验结果表明这 3 个指标在 4 种权重分配策略下采用代表性重心策略的分类精度平均提升率分别为 7.40%，9.89%和 13.76%，优于采用多样性重心策略的分类精度平均提升率(6.03%，8.16%，12.18%)。而图 1(d)－图 1(f)所给出 CO，VE 和 VE/VCO_2 的同类样本分布较为分散，应采用多样性重心策略。实验结果表明这 3 个指标在 4 种权重分配策略下采用多样性重心策略的分类精度平均提升率分别为 8.51%，10.63%和 8.40%，优于采用代表性重心策略的分类精度平均提升率(7.98%，7.17%，5.19%)。

(2)权重分配策略对数据增强效果的影响

为了衡量权重分配策略的优劣，本文根据表 3－表 6 的实验结果统计了不同权重分配策略下所有指标(无论何种重心选择策略)的平均分类精度。其中，基于 AA 权重策略的平均分类精度为 0.751，基于 AS 权重策略的平均分类精度为 0.749，基于 ASD 策略的平均分类精度为 0.744，而基于本文提出的 RSD 权重策略的平均分类精度最高达到了 0.752，相比于无数据增强时的平均分类精度(0.701)提升了 7.15%。从实验结果可以发现，AA 策略和 RSD 策略的表现要优于 AS 策略和 ASD 策略。造成这一结果的主要原因是 AS 和 ASD 策略都是根据样本与初始重心的 DTW 距离来分配权重，特别是 ASD 策略的权重随着距离从近到远递减。当初始重心选定后，ASD 策略极易生成重复样本。而 AS 策略比 ASD 策略表现略好的原因是 AS 策略是从 5 个最近邻中随机挑选两个各分配 0.15 权重值，因此生成重复样本的概率小于 ASD 策略。

为了进一步讨论生成重复样本对模型性能的影响，本文设计了另一个实验。该实验采用随机重心策略选定初始重心后，对比了随距离递减的两种权重分配策略 ASD(容易生成重复样本)和 RSD(不易生成重复样本)在生成样本数量不断增多(即生成重复样本的概率不断增大)时的表现差异，实验结果如图 3 所示。由于本文研究的问题属于二分类问题，因此本实验对每类样本的生成数量采取 5 的倍数依次递增。例如，图中的纵坐标生成样本数量为 10，表示每类训练样本各生成 5 个新样本。如图 3 所示，不论是哪种代谢指标，随着生成样本数量增多，分类精度也大体呈现随之提升的趋势，其中 RSD 策略的上升趋势比 ASD 的更为稳定。特别地，代谢指标 VE，QT，VE/VCO_2 和 R 在生成样本数量超过 50 后，RSD 策略的表现要优于 ASD 策略。但是对 VO_2/HR 而言，ASD 策略与 RSD 策略的优劣难以区分；对 VO_2/kg 而言，ASD 策略甚至略好于 RSD 策略。但总体而言，RSD 策略仍优于 ASD 策略。

表 3 AA 权重策略下不同代谢指标基于不同重心选择策略的分类精度

Table 3 Classification accuracy of different metabolic indicators based on different barycenter selection strategies under AA weighting strategy

数据增强方案	CPET 代谢指标					
	VO_2/HR	VO_2/kg	R	VE/VCO_2	VE	CO
无	0.750	0.708	0.667	0.667	0.708	0.708
随机重心	0.729	0.757	**0.757**	0.701	0.736	0.757
轮流重心	0.736	0.736	0.736	0.701	0.729	0.708
代表性重心	0.792	**0.785**	**0.757**	0.736	0.771	0.750
多样性重心	**0.799**	0.778	**0.757**	**0.736**	**0.792**	**0.785**

表 4 AS 权重策略下不同代谢指标基于不同重心选择策略的分类精度

Table 4 Classification accuracy of different metabolic indicators based on different barycenter selection strategies under AS weighting strategy

数据增强方案	CPET 代谢指标					
	VO_2/HR	VO_2/kg	R	VE/VCO_2	VE	CO
无	0.750	0.708	0.667	0.667	0.708	0.708
随机重心	0.771	0.750	0.701	0.715	0.736	0.743
轮流重心	0.743	0.778	0.743	0.701	0.736	**0.757**
代表性重心	**0.792**	**0.806**	**0.778**	**0.729**	0.757	0.708
多样性重心	0.757	0.771	0.757	0.722	**0.785**	0.743

表 5 ASD 权重策略下不同代谢指标基于不同重心选择策略的分类精度

Table 5 Classification accuracy of different metabolic indicators based on different barycenter selection strategies under ASD weighting strategy

数据增强方案	CPET 代谢指标					
	VO_2/HR	VO_2/kg	R	VE/VCO_2	VE	CO
无	0.750	**0.708**	0.667	0.667	0.708	0.708
随机重心	0.764	0.715	0.743	**0.708**	0.743	0.743
轮流重心	0.792	0.667	**0.750**	0.667	**0.792**	0.750
代表性重心	**0.819**	**0.764**	0.736	0.701	0.743	0.743
多样性重心	**0.819**	0.757	0.722	**0.708**	0.764	**0.757**

表 6 RSD 权重策略下不同代谢指标基于不同重心选择策略的分类精度

Table 6 Classification accuracy of different metabolic indicators based on different barycenter selection strategies under RSD weighting strategy

数据增强方案	CPET 代谢指标					
	VO_2/HR	VO_2/kg	R	VE/VCO_2	VE	CO
无	0.750	0.708	0.667	0.667	0.708	0.708
随机重心	0.799	0.688	0.729	0.688	0.778	0.750
轮流重心	0.757	0.736	0.715	0.708	0.743	0.722
代表性重心	**0.819**	**0.757**	**0.764**	0.715	0.764	0.778
多样性重心	0.806	**0.757**	0.757	**0.729**	**0.792**	**0.785**

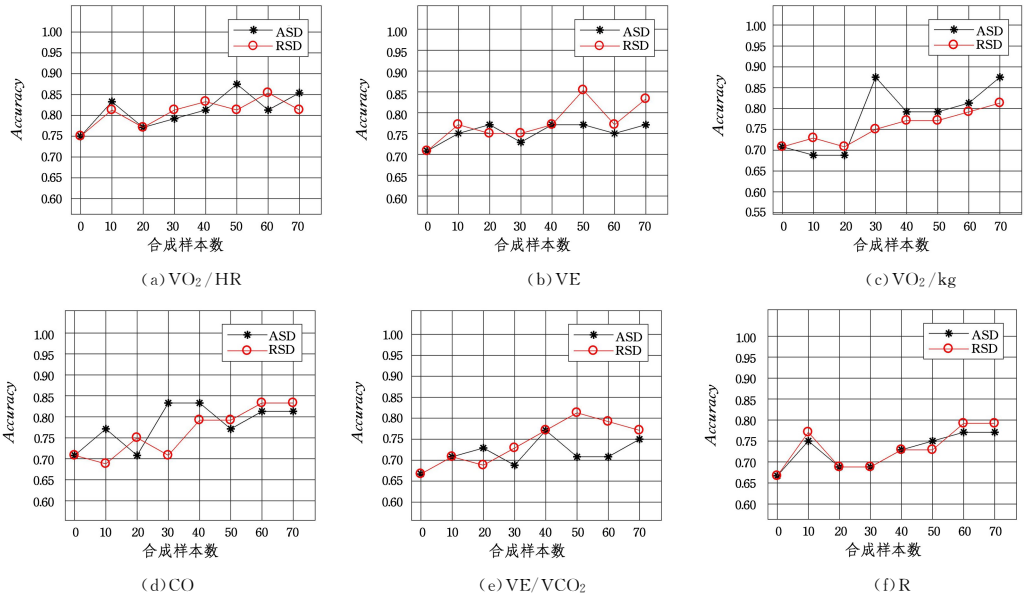

图 3　随机重心策略下 6 个代谢指标基于 ASD 策略和 RSD 策略的分类精度

Fig. 3　Classification accuracy of six metabolic indicators based on ASD and RSD strategies under random barycenter strategy

6　UCR 数据集补充实验

鉴于本文所使用的高血压患者 CPET 数据集相对较小，为了进一步验证本文所提出的重心策略和权重策略对 WDBA 算法的影响，本文选用 UCR 档案库[34]中来自于不同应用类型的若干数据集进行了补充实验，表 7 列出了所选数据集的一些基本信息，这些数据集的共同特点是样本长度较大，但训练样本较少。

本文将表 7 的数据集进行了与 CPET 数据集类似的数据增强实验。实验同样首先利用原始样本生成相同数量的合成样本，然后将其加入训练集共同训练 SRC-AL 分类器，最后利用测试样本评估不同数据增强方案的优劣。实验结果如表 8 所列，其中若某个方案的分类精度低于无数据增强时，则该方案的分类精度用下划线标识。

表 7　来自于 UCR 档案库中部分数据集的基本情况

Table 7　Information from some data sets in UCR archive

类型	数据集名	类别数	样本长度	训练样本数	测试样本数
ECG	TwoLeadECG	2	82	23	1 139
ECG	ECGFiveDays	2	136	23	861
Sensor	MoteStrain	2	84	20	1 252
Image	ArrowHead	3	251	36	175
Image	Herring	2	512	64	64
Motion	GunPoint	2	150	50	150

表 8　不同数据增强方案在 6 个 UCR 数据集下的分类精度

Table 8　Classification accuracy of different data augmentation schemes under 6 UCR data sets

方案	TwoLeadECG	ECGFiveDays	MoteStrain	ArrowHead	Herring	GunPoint
无数据增强	0.918	0.974	0.885	0.829	0.672	0.933
RAN-AA[23]	0.933	0.974	0.892	0.806	0.688	0.947
RAN-AS[23]	0.911	0.976	0.895	0.829	0.672	0.940
RAN-ASD[23]	0.902	0.967	0.884	0.829	0.672	0.900
RAN-RSD	0.931	0.969	0.883	0.806	0.688	0.933
ALT-AA	0.926	0.976	0.876	0.794	0.719	0.933
ALT-AS	0.928	0.972	0.896	0.800	0.688	0.927
ALT-ASD	0.929	0.972	0.881	0.857	0.703	0.940
ALT-RSD	0.932	0.973	0.879	0.794	0.672	0.940
REP-AA	0.946	0.976	0.888	0.840	**0.750**	0.933
REP-AS	0.931	0.973	**0.900**	0.840	0.719	0.940
REP-ASD	0.912	0.942	**0.900**	0.834	0.719	0.913
REP-RSD	**0.948**	0.977	0.898	0.817	0.703	**0.947**
DIV-AA	0.940	0.979	0.899	0.851	0.734	0.933
DIV-AS	0.934	0.980	0.887	0.829	0.688	0.947
DIV-ASD	0.904	0.945	0.887	**0.869**	0.703	0.913
DIV-RSD	0.933	**0.984**	0.894	**0.869**	**0.750**	**0.947**

实验结果表明两个基于主动重心挑选的策略（代表性重心 REP 和多样性重心 DIV）的表现要优于重心无差别策略（随机重心 RAN 和轮流重心 ALT）。具体体现在：一方面，第 10－17 行的结果中，分类精度降低的方案个数要远远少于第 2－9 行的个数；而另一方面，取得最好结果的方案个数却远远多于第 2－9 行的个数。当确定了主动重心挑选策略的优势后，再来评估权重分配策略的好坏。与代表性重心 REP 结合的 4 种权重分配方案中，本文提出的 RSD 策略在两个数据集上取得了最优结果，而其他 3 种现有策略仅在 1 个数据集上表现最优。类似的，与多样性重心 DIV 结合的 4 种权重分配方案中，RSD 策略在 4 个数据集上表现最优，与其他 3 个策略的优势进一步拉大。这充分说明了本文提出的两个主动

重心挑选策略与 RSD 权重分配策略的优越性。

结束语 本文对青年高血压患者心肺运动试验(CPET)时序数据进行数据挖掘,旨在了解不同个体对有氧运动训练的响应性。由于获取青年高血压患者心肺运动时序数据需要患者配合专业医护人员的指导并通过专用仪器采集,所以开展该研究的瓶颈之一在于难以获取充足的样本数据。对于该问题,较为可行的方法是数据增强。本文针对加权动态时间规整重心平均数据增强算法 WDBA,分析其不足,针对重心选择和权重分配策略进行了深入的研究。首先,针对重心选择问题,本文在现有随机重心策略的基础上,提出了轮流重心策略和两个主动重心挑选策略(代表性重心策略和多样性重心策略)。实验结果表明,利用主动学习的思想,考虑适合的样本作为重心能够提升数据增强在 SRC-AL 分类模型的预测精度。此外,对于不同的 CPET 指标,本文建议应根据真实训练样本的数据分布进行代表性重心或多样性重心策略的选择。例如:对代谢指标 VO_2/HR, VO_2/kg 和 R 而言,考虑代表性重心的数据增强效果更加明显;对代谢指标 CO, VE 和 VE/VCO_2 而言,考虑多样性重心更有利于数据增强效果的提升。最后,针对现有权重分配策略的不足,提出了随机权重距离递减分配策略。实验结果表明,该策略可有效避免生成重复样本,从而进一步提升模型的泛化能力。除了本文的应用之外,UCR 数据集的补充实验也证明了本文提出的数据增强方案能够有效地提升分类模型的分类精度。

在后续的工作中,作者将致力于以下 4 个方面的研究:1)研究基于其他信息量衡量标准的主动重心选择策略,如不确定性或多标准组合;2)研究重心选择与权重分配策略之间的潜在关系,例如不同的重心选择策略,是应该更注重权重分配的灵活性,还是更关注距离的影响;3)研究是否根据生成样本的数量动态调整重心选择与权重分配策略;4)研究本文提出的数据增强方案在其他应用场景中的适用性。

参 考 文 献

[1] YANO Y, REIS J P, COLANGELO L A, et al. Association of blood pressure classification in young adults using the 2017 American College of Cardiology/American Heart Association blood pressure guideline with cardiovascular events later in life [J]. Jama, 2018, 320(17): 1774-1782.

[2] WU S, SONG Y, CHEN S, et al. Blood pressure classification of 2017 associated with cardiovascular disease and mortality in young Chinese adults[J]. Hypertension, 2020, 76(1): 251-258.

[3] BROOK R D, APPEL L J, RUBENFIRE M, et al. Beyond medications and diet: alternative approaches to lowering blood pressure: a scientific statement from the American Heart Association[J]. Hypertension, 2013, 61(6): 1360-1383.

[4] WEN H, WANG L. Reducing effect of aerobic exercise on blood pressure of essential hypertensive patients: A meta-analysis[J]. Medicine, 2017, 96(11)e6150.

[5] CAO L, LI X, YAN P, et al. The effectiveness of aerobic exercise for hypertensive population: a systematic review and meta-analysis[J]. The Journal of Clinical Hypertension, 2019, 21(7): 868-876.

[6] PEDRALLI M L, EIBEL B, WACLAWOVSKY G, et al. Effects of exercise training on endothelial function in individuals with hypertension: a systematic review with meta-analysis[J]. Journal of the American Society of Hypertension, 2018, 12(12): e65-e75.

[7] GOROSTEGI-ANDUAGA I, CORRES P, MARTINEZAGUIRRE-BETOLAZA A, et al. Effects of different aerobic exercise programmes with nutritional intervention in sedentary adults with overweight/obesity and hypertension: EXERDIET-HTA study[J]. European Journal of Preventive Cardiology, 2018, 25(4): 343-353.

[8] PEDERSEN B K, SALTIN B. Exercise as medicine-evidence for prescribing exercise as therapy in 26 different chronic diseases [J]. Scandinavian Journal of Medicine & Science in Sports, 2015, 25: 1-72.

[9] LOPES S, MESQUITA-BASTOS J, ALVES A J, et al. Exercise as a tool for hypertension and resistant hypertension management: current insights[J]. Integrated Blood Pressure Control, 2018, 11: 65-71.

[10] HACKE C, NUNAN D, WEISSER B. Do exercise trials for hypertension adequately report interventions? A reporting quality study[J]. International Journal of Sports Medicine, 2018, 39(12): 902-908.

[11] OZEMEK C, ARENA R. Precision in promoting physical activity and exercise with the overarching goal of moving more[J]. Progress in Cardiovascular Diseases, 2019, 62(1): 3-8.

[12] ROSS R, GOODPASTER B H, KOCH L G, et al. Precision exercise medicine: understanding exercise response variability[J]. British journal of sports medicine, 2019, 53(18): 1141-1153.

[13] YANG G, LENG X, HUANG F, et al. Use CPET data to predict the intervention effect of aerobic exercise on young hypertensive patients[C]// 2019 IEEE International Conference on Bioinformatics and Biomedicine(BIBM). IEEE, 2019: 1699-1702.

[14] ALBOUAINI K, EGRED M, ALAHMAR A, et al. Cardiopulmonary exercise testing and its application[J]. Postgraduate medical journal, 2007, 83(985): 675-682.

[15] BALADY G J, ARENA R, SIETSEMA K, et al. Clinician's guide to cardiopulmonary exercise testing in adults: a scientific statement from the American Heart Association[J]. Circulation, 2010, 122(2): 191-225.

[16] YOUN J C, KANG S M. Cardiopulmonary exercise test in patients with hypertension: focused on hypertensive response to exercise[J]. Pulse, 2015, 3(2): 114-117.

[17] HUANG F, LENG X, KASUKURTHI M V, et al. Utilizing Machine Learning Techniques to Predict the Efficacy of Aerobic Exercise Intervention on Young Hypertensive Patients Based on Cardiopulmonary Exercise Testing[J]. Journal of Healthcare Engineering, 2021(1): 6633832.

[18] SHORTEN C, KHOSHGOFTAAR T M. A survey on image data augmentation for deep learning[J]. Journal of Big Data, 2019, 6(1): 1-48.

[19] LE GUENNEC A, MALINOWSKI S, TAVENARD R. Data augmentation for time series classification using convolutional neural networks[C]// ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data. 2016.

[20] WEN Q, GAO J, SONG X, et al. RobustSTL: A robust seasonal-

trend decomposition algorithm for long time series[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:5409-5416.

[21] BERGMEIR C, HYNDMAN R J, BENITEZ J M. Bagging exponential smoothing methods using STL decomposition and Box-Cox transformation[J]. International Journal of Forecasting, 2016, 32(2):303-312.

[22] KEGEL L, HAHMANN M, LEHNER W. Feature-based comparison and generation of time series[C]// Proceedings of the 30th International Conference on Scientific and Statistical Database Management. 2018:1-12.

[23] GAO J, SONG X, WEN Q, et al. RobustTAD: Robust time series anomaly detection via decomposition and convolutional neural networks[J]. arXiv:2002.09545, 2020.

[24] WEN Q, SUN L, YANG F, et al. Time series data augmentation for deep learning: A survey[J]. arXiv:2002.12478, 2020.

[25] ESTEBAN C, HYLAND S L, RÄTSCH G. Real-valued (medical) time series generation with recurrent conditional gans[J]. arXiv:1706.02633, 2017.

[26] YOON J, JARRETT D, VAN DER SCHAAR M. Time-series generative adversarial networks[C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019:5508-5518.

[27] BAGNALL A, LINES J, BOSTROM A, et al. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances[J]. Data Mining and Knowledge Discovery, 2017, 31(3):606-660.

[28] PETITJEAN F, FORESTIER G, WEBB G I, et al. Faster and more accurate classification of time series by exploiting a novel dynamic time warping averaging algorithm[J]. Knowledge and Information Systems, 2016, 47(1):1-26.

[29] FORESTIER G, PETITJEAN F, DAU H A, et al. Generating synthetic time series to augment sparse datasets[C]// 2017 IEEE International Conference on Data Mining (ICDM). IEEE, 2017:865-870.

[30] PETITJEAN F, KETTERLIN A, GANÇARSKI P. A global averaging method for dynamic time warping, with applications to clustering[J]. Pattern Recognition, 2011, 44(3):678-693.

[31] PETITJEAN F, GANÇARSKI P. Summarizing a set of time series by averaging: From Steiner sequence to compact multiple alignment[J]. Theoretical Computer Science, 2012, 414(1):76-91.

[32] GILYAZEV R A, TURDAKOV D Y. Active learning and crowdsourcing: A survey of optimization methods for data labeling[J]. Programming and Computer Software, 2018, 44(6):476-491.

[33] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of machine learning research, 2008, 9(11):2579-2605.

[34] DAU H A, BAGNALL A, KAMGAR K, et al. The UCR time series archive[J]. IEEE/CAA Journal of Automatica Sinica, 2019, 6(6):1293-1305.

HUANG Fangwan, born in 1980, Ph.D, senior lecturer, is a member of China Computer Federation. Her main research interests include computational intelligence, machine learning and big data analysis.

YU Zhiyong, born in 1982, Ph.D, professor, is a member of China Computer Federation. His main research interests include pervasive computing, mobile social networks, and crowd sensing.