计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220295
61(1): 156−171, 2024

多项正则化约束的伪标签传播优化脑电信号聚类

代成龙[1] 李光辉[1] 李 栋[1] 申佳华[1] 皮德常[2]
[1](江南大学人工智能与计算机学院 江苏无锡 214122)
[2](南京航空航天大学计算机科学与技术学院 南京 211106)
(chenglongdai@jiangnan.edu.cn)

Electroencephalogram Clustering with Multiple Regularization Constrained Pseudo Label Propagation Optimization

Dai Chenglong[1], Li Guanghui[1], Li Dong[1], Shen Jiahua[1], and Pi Dechang[2]
[1](*School of Artificial Intelligence and Computer Science*, *Jiangnan University*, *Wuxi*, *Jiangsu* 214122)
[2](*College of Computer Science and Technology*, *Nanjing University of Aeronautics and Astronautics*, *Nanjing* 211106)

Abstract As the non-invasive analyzing media, electroencephalogram (EEG) signals are widely applied in brain-computer interfaces, dysfunctional disorder diagnosis and rehabilitation. However, the techniques used in such applications are supervised and completely require EEG labels, like classification. Meanwhile, with the ever-increasing of unlabeled EEG emerged in these applications, traditional supervised techniques are becoming inapplicable, which probably degrades the development of this new-type unlabeled EEG in the emerging potential fields. To deal with the issue of unsupervised analysis for unlabeled EEG signals, we propose a multiple regularization constrained pseudo label propagation optimization model, which integrates the pseudo label propagation learning, EEG similarity adjacency matrix approximation, and label classifier learning. Subsequently, to pursuit the goal of EEG clustering with the proposed model, we transform the model to a multi-objective optimization function and propose a gradient descent-based algorithm named EEGapc (electroencephalogram clustering with pseudo label propagation) to solve it. EEGapc not only can make best use of messages passing through pairwise EEG signals in EEG-constructed graph, but can also quickly converge to its local optima. Experimental results by comparing EEGapc with 8 different types of state-of-the-art clustering algorithms on 14 real-world EEG data sets clearly demonstrate the superiority of EEGapc, and its performances with respect to average *NMI* (normalized mutual information), *ARI* (adjusted rand index), *F-score* and *kappa* are at least improved by 86.88%, 58.01%, 6.29%, 61.17%, respectively.

Key words electroencephalogram clustering; pseudo label propagation; adjacency matrix optimization; pseudo label classifier; multi-objective optimization

摘 要 作为一种非侵入式分析载体,脑电信号目前被广泛应用于脑-机接口、医疗辅助诊断及康复领域,但这些应用通常依赖需要完整标签的有监督分析技术,如分类.随着无标签脑电信号的与日俱增,现有的有监督方法不能有效解决无标签脑电信号分析问题,也在一定程度上限制了无标签脑电信号这类新型数据的应用拓展.为了解决无标签脑电信号的无监督分析问题,提出了一种基于多项正则化约束的伪标签传播优化聚类模型.该模型通过同时优化学习伪标签传播矩阵、脑电信号相似度邻接矩阵、标签分类器的方式实现聚类.将提出的脑电信号聚类模型转化为一个多目标优化问题,并提出了一种基于梯度下降策略的聚类算法 EEGapc (electroencephalogram clustering with pseudo label propagation).该算法不仅充分考虑了

收稿日期: 2022-04-12; **修回日期**: 2023-01-16
基金项目: 国家自然科学基金项目 (62106087); 江苏省自然科学基金项目 (BK20210455); 中央高校基本科研业务费专项资金 (JUSRP122033)
This work was supported by the National Natural Science Foundation of China (62106087), the Natural Science Foundation of Jiangsu Province (BK20210455), and the Fundamental Research Funds for the Central Universities (JUSRP122033).

脑电信号之间的相关性及脑电信号间的信息传递，还能快速收敛到局部最优. 在 14 个真实脑电信号数据集上的实验结果表明，提出的 EEGapc 脑电信号聚类算法比现有的 8 种聚类算法性能更好，且在平均 *NMI* (normalized mutual information), *ARI* (adjusted rand index), *F-score*, *kappa* 这 4 个指标上，EEGapc 与现有的 8 种聚类算法相比，分别至少提升了 86.88%, 58.01%, 6.29%, 61.17%.

关键词 脑电信号聚类；伪标签传播；邻接矩阵优化；伪标签分类器；多目标优化

中图法分类号 TP391

人类进化进程中形成的最复杂、信息处理最完善的器官就是大脑，它有效控制着人的情绪、思想和行动，是人类从外界获取信息、处理信息的重要“工具”. 大脑的认知机制、工作机制以及与相关疾病的关联等都逐渐成为研究学者们所关注的焦点，也形成了脑科学研究的热潮. 国家中长期科学和技术发展规划纲要（2006—2020 年）针对脑科学进行了研究规划，其中涉及脑重大疾病的产生机理、脑信息的表达与处理以及人脑与计算机交互等内容. 近年来，国家则侧重以发展类脑人工智能的计算技术与器件以及研发脑重大疾病的诊断干预方法作为应用导向. 2013 年“973”重点资助研究领域的研究内容包括了脑-机一体化以及生机电融合的智能型医疗设备、康复装备的研究. 国家自然基金委重大研究计划“视听觉信息的认知计算”（2008—2017 年）包括了视听觉认知相关的脑-机接口（brain-computer interface, BCI）技术. 2016 年，脑科学与类脑研究被“十三五”规划纲要确定为重大科技创新项目和工程之一. 目前，“脑科学与类脑科学”作为“科技创新 2030 重大项目”已经启动指南意见征求稿. 另外，近几年的军委装备发展部的支持项目中也涉及基于脑电信号的脑控系统、人机交互接口的研究内容. 可见，基于脑电信号分析的脑科学技术在未来研究与应用拓展中具有重大意义.

作为一种能有效反映大脑工作状态的媒介，脑电信号（electroencephalogram, EEG）成为了脑重大疾病辅助诊断、人机交互、康复等领域的重要纽带. 脑电信号是由脑细胞受到刺激产生的生理信号，即生物电信号. 它不仅能有效反映出当前大脑功能状态，也能反馈人的身体机能状况[1]，因此被广泛应用于神经系统疾病分析[2-5] 与康复[6]、脑-机接口[7-8]. 其中，基于脑-机接口的应用尤为突出，该应用包含潜在的军事应用[9-10]. 随着脑电信号数据的日益增加，无标签脑电数据的比重也随之增大，这给目前普遍的有监督分析方法带来了巨大挑战. 因为有监督的分析方法，如分类，十分依赖数据的标签信息[5,8,11]. 同时，人工标记脑电信号既费时又费力，这也在很大程度上限制了无标签脑电信号这类新型数据在脑-机接口、疾病辅助诊断、康复等领域的应用. 在当前脑电信号的研究环境下，现有的研究主要存在 3 个问题或挑战：

1）现有的有监督分析方法，如分类，需要给定数据标签进行学习训练，因此无法有效迁移到无标签脑电信号的分析任务中. 而且，人工标记与日俱增的无标签脑电信号成本高、费时费力.

2）针对无标签脑电信号的研究相对较少，可借鉴的分析方法有限，需要提出新的方案、方法来解决该具有挑战性的任务.

3）与传统的时间序列数据不同，脑电信号是一种具有高震荡、高非线性、低信噪比等特点的弱生物信号，现有的大部分传统时间序列聚类方法可能对无标签脑电信号这类新型数据的处理效果不佳.

为此，本文提出了一种基于多项正则化约束的伪标签传播优化聚类方法，旨在充分利用无标签脑电信号自身特征及相关性，自主学习、优化得到相对最优的脑电信号聚类结果. 具体而言，本文的主要贡献及工作包括 3 个方面：

1）针对无标签脑电信号的分析问题，提出了一种基于多项正则化约束的伪标签信息传播优化聚类模型. 该模型充分利用脑电信号之间的相关性与信息传播，并通过同时学习伪标签传播矩阵、脑电信号相似度邻接矩阵、标签分类器的方式实现无标签脑电信号聚类.

2）为了解决由脑电信号聚类转化的多目标优化问题，提出了一种基于梯度下降策略的多目标优化算法EEGapc（electroencephalogram clustering with pseudo label propagation）. 该算法能在一个很小的学习率条件下很快收敛到局部最优，不仅确保了多目标函数的可解性，也有利于脑电信号的快速聚类.

3）通过在 14 个真实脑电信号数据集、4 个评价指标上的实验分析，展示了本文提出的 EEGapc 算法的性能优越性. 与现有的多种聚类算法相比，本文提出的 EEGapc 算法能得到更优的聚类结果.

1 相关工作

随着脑电信号规模在脑-机接口应用、疾病诊断、康复等领域的不断增大，无标签或误标记脑电信号的数量也随之增加，特别是针对患有脑部疾病的被试所采集的脑电信号，其标签缺失、误标尤为突出[12]. 同时，人工标记脑电信号是一件费时费力的事情，在一定程度上制约了脑电信号的应用. 因此，针对无标签脑电信号这类新型数据进行研究，既能拓展其应用范围，也能为其提供鲁棒性更高、性能更好的分析策略与方法.

作为一种有效的无监督分析方法，聚类可以用来处理无标签脑电信号. 目前专门针对脑电信号的聚类研究相对较少，主要是基于相似度衡量和中心搜寻相结合的脑电信号聚类方法，如 MTEEGC（multi-trialelectroencephalogram clustering）[13]，FCM（fuzzy C-means）[14]，以及基于图的脑电信号聚类方法，如 mwc-EEGc（maximum weight clique electroencephalogram clustering）[15]. 具体而言，MTEEGC 采用一种互相关相似度衡量方法评估脑电信号之间的相关性，并利用优化目标函数搜寻聚类中心，最后根据脑电信号与聚类中心的相似度实现脑电信号聚类；FCM 脑电信号聚类方法与 k-means 相似，通过判断脑电信号与聚类中心的相对距离实现脑电信号的聚类. 这类脑电信号聚类方法的性能在很大程度上受聚类中心初始化与中心更新策略的影响. mwcEEGc 同样先利用相似度衡量方法评估脑电信号之间的相关性权重，并将脑电信号及其相关性权重映射为一个无向加权图，最后结合相似度阈值搜寻图中最大加权团的方式实现脑电信号聚类. 该脑电信号聚类方法虽然具有较高的性能，但多次的最大加权团搜寻过程会消耗大量的时间，对于实时性要求较高的应用而言，如脑-机接口等，mwcEEGc 的应用会受到一定限制.

由于面向无监督分析的脑电信号研究方法有限，在一定程度上影响了日益增长的无标签脑电信号的分析和应用拓展. 幸运的是，作为一种具有时间序列特性的生物电信号，丰富的时间序列聚类方法可以为脑电信号聚类提供方法指导. 目前，涉及到脑电信号聚类的研究方法也通常是基于时间序列的聚类策略[13-14]，如 MTEEGC，FCM. 一般而言，基于时间序列的聚类方法主要通过自身的学习、优化策略，例如数据之间的相关性、数据特征之间的相关性等进行自动划分，最终实现数据的聚类[12,15]. 目前为止，基于时间序列的聚类方法可以大致分为 6 类：

1）经典的 k-means 型聚类算法[16]. 这类算法继承了经典 k-means 的聚类策略，即根据数据与初始化聚类中心之间的距离，迭代更新聚类中心并根据预先设定好的聚类个数将数据划分到与其最近的聚簇中，达到聚类的目的，如 k-means++[17]，k-variates++[18]，KMM（K-multiple-means）[19] 等. 不难看出，这类 k-means 型聚类方法的性能依赖于一个合适的初始化策略或一个中心（重心）搜寻更新方法. 换言之，这类聚类方法的最终效果对聚类中心（重心）的初始化要求较高，虽然效率高，但聚类精度通常相对较差.

2）基于相似度的聚类. 该类方法主要利用相似度（或距离）衡量方法评估数据之间的相关性，并结合设定的相似度阈值将满足条件的数据划分到同一聚簇中，如 DBA（dynamic time warping barycenter averaging）[20]，K-SC（K-spectral centroid）[21] 等. 这类聚类算法主要依赖于相似度（或距离）衡量方法，并在聚类过程中也涉及到聚类中心的搜寻，因此这类算法在计算数据相似度且根据相似度衡量方法不断更新聚类中心时会消耗大量的时间，所以它们的时间复杂度相对较高.

3）基于密度的聚类. 这类方法主要利用预先设定的密度评估参数衡量数据之间构成的稠密区域是否满足最低密度阈值，从而实现数据的聚类划分. 基于密度的方法有 DBSCAN（density-based spatial clustering of applications with noise）[22]，OPTICS（ordering points to identify the clustering structure）[23]，SNN（shared nearest neighbor）[24]，DP（density peak clustering）[25-27] 等. 这类方法的效率较高，但聚类性能也受到密度评估参数的影响. 而且，基于密度的方法通常情况下对“点”型数据聚类效果相对较好，而对于时间序列数据而言相对较差.

4）基于特征选择的聚类. 这类方法主要从数据降维的角度，通过特征识别与选择方法将原始数据映射到低维的特征空间，然后利用数据与特征之间的相关性进行聚类操作，如 NDFS（nonnegative discriminative feature selection）[28]，RUFS（robust unsupervised feature selection）[29]，RSFS（robust spectral learning framework for unsupervised feature selection）[30]，CGSSL（clustering-guided sparse structural learning）[31] 等. 虽然这类方法将降低原始数据的维度，但其聚类性能同样也受到特征识别与选择方法以及特征个数设定的影响. 而对于特征复杂的脑电信号而言，特征提取的时间消耗也比较大.

5）基于数据形态的聚类. 这类方法与基于特征

选择的聚类方法策略相似, 通过学习将原始数据用较短的数据段, 即数据形态, 降低原始数据的维度, 然后利用学习到的多个数据形态与原始数据之间的相互关系进行聚类操作, 如 u-shapelet（clustering using unsupervised shapelets）[32], USSL（unsupervised salient subsequence learning）[33], kShape（k-shape clustering）[34]等. 同理, 这类方法的聚类性能同样受到数据形态学习方法的影响. 而且, 想要学习到差异性较高、代表性较强的形态数据段也需要一个良好的学习目标函数和大量的学习时间.

6）谱聚类. 这类方法是基于图的聚类策略, 通过将数据映射为图中节点, 数据之间的相关性映射为图边上的权重, 然后进行图分割实现聚类, 例如 BSGP（bipartite spectral graph partition）[35], USPEC（ultra-scalable spectral clustering）[36]等. 谱聚类需要数据构成的相似度矩阵, 然后根据拉普拉斯矩阵进行相似度变换, 最后进行图分割. 这类方法对数据结构无需过多的假设要求, 也不需要对数据的概率分布做假设. 同时, 通过构造稀疏相似度图, 可以降低聚类的时间消耗. 但是, 谱聚类方法对图相似度比较敏感.

本文针对无标签脑电信号进行聚类分析, 提出了一种基于图的伪标签传播优化聚类方法. 在多项正则化的约束下, 建立多目标优化模型, 即将伪标签传播学习与脑电信号相似度邻接矩阵正则化约束和标签分类器正则化约束相结合, 从而学习到与真实标签接近的伪标签, 实现无标签脑电信号的聚类任务.

2 准备工作

聚类最本质的目的是将相似度高的数据划分到同一聚簇中, 并使得聚簇之间的差异最大化. 本文采用的基于图的聚类方式的主要目的也是将相似度较高的脑电信号通过图中的边权重将脑电信号划分到对应的子图中, 这个过程类似于谱聚类. 为了让文章更好地被理解, 在详细介绍本文方法之前, 先对方法涉及到的相关知识, 如拉普拉斯矩阵等进行简要介绍.

假设 $G=(V, E)$为脑电信号构成的无向加权图, 其中, $V=\{e_1, e_2, \cdots, e_n\}$为 n 条脑电信号构成的图节点, E 为图中节点之间的边. 对于边(e_i, e_j)上的权重 s_{ij}, s_{ij} 为邻接矩阵 $\boldsymbol{S}$ 的元素, 本文采用皮尔逊相关系数的变形方式来衡量, 即给定 2 条长度为 m 的脑电信号 e_x 与 e_y, 它们之间的皮尔逊相关性 r_{xy}[37] 定义为

$$r_{xy}=\frac{\sum_{i=1}^{m}(e_{x_i}-\bar{e}_x)(e_{y_i}-\bar{e}_y)}{\sqrt{\sum_{i=1}^{m}(e_{x_i}-\bar{e}_x)^2}\sqrt{\sum_{i=1}^{m}(e_{y_i}-\bar{e}_y)^2}}, \tag{1}$$

其中$\bar{e}_x, \bar{e}_y$分别为脑电信号 e_x 与 e_y 数据点的平均值, 且 $r_{xy}\in[-1,1]$. 因此, 为了获得非负的脑电信号相似度, 本文将皮尔逊相关性衡量进行变形来衡量脑电信号 e_x 与 e_y 之间的相似度 s_{xy}:

$$s_{xy}=1-\frac{1-r_{xy}}{2}. \tag{2}$$

最终由 s_{xy}构成脑电信号相似度邻接矩阵 $\boldsymbol{S}$, $\boldsymbol{S}=(s_{xy})$, $s_{xy}\in[0,1]$.

基于脑电信号相似度邻接矩阵 $\boldsymbol{S}$, 脑电信号构成的图结构中拉普拉斯矩阵 $\boldsymbol{\mathcal{L}}_S$ 定义为

$$\boldsymbol{\mathcal{L}}_S=\boldsymbol{D}-\boldsymbol{S}, \tag{3}$$

其中 $\mathrm{rank}(\boldsymbol{\mathcal{L}}_S)=n-c$, c 为 $\boldsymbol{\mathcal{L}}_S$ 特征值为 0 的个数, 也表示为图 G 中可划分的子图个数[38-39]; 同时, $\boldsymbol{D}=(\boldsymbol{d}_i)=(d_{ii})$ 为对角矩阵上的元素, 定义为

$$d_{ii}=\sum_{j=1}^{n}s_{ij}. \tag{4}$$

对于标准的基于图的聚类, 其目的是将相似度较高的数据划分到同一聚簇中, 因此一个好的相似度衡量策略有利于得到更好的聚类结果. 对于基于图的聚类而言, 其相似度矩阵可以作为数据间的传播信息, 有利于将相关性较高的数据划分到同一聚簇中, 提高数据的聚类效果. 换言之, 无标签数据的类标也可以通过数据间的信息传播, 实现标记的动态学习与优化, 从而实现无标签数据的标签化,即聚类[40]. 为了更好地基于相似度邻接矩阵对无标签脑电信号的标签进行传播学习, 可以将脑电信号相似度邻接矩阵 $\boldsymbol{S}$ 重定义为一个传播矩阵[41]:

$$\boldsymbol{Q}=\boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}. \tag{5}$$

本文中采用的伪标签传播学习则可以定义为

$$\boldsymbol{P}^t=\alpha\boldsymbol{P}^{t-1}\boldsymbol{Q}+(1-\alpha)\boldsymbol{L}, \tag{6}$$

其中$0<\alpha<1$为权重系数, $\boldsymbol{L}\in\mathbb{R}^{c\times n}$为伪标签指示矩阵. $\boldsymbol{L}=(L_{ij})$具体定义为

$$L_{ij}=\begin{cases}1, & l_i=j.\\ 0, & \text{其他}.\end{cases} \tag{7}$$

伪标签指示矩阵 $\boldsymbol{L}$ 表示: 对于一个 c 聚类问题, 当脑电信号 e_i 被划分到第 j 类聚簇中, 即 $l_i=j$, 则有 $L_{ij}=1$, 否则 $L_{ij}=0$. 当然, 该伪标签会根据优化策略不断地更新.

为了能更好地让标签信息（即脑电信号的相似度

邻接矩阵所包含的信息)在整个图结构中平稳传播,实现未标记节点的标签化,本文将标准的标签传播模型引入正则化约束,最后形成正则化标签传播模型[41]:

$$\min_{\boldsymbol{P}} \frac{1}{2}\sum_{k=1}^{c}\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}(P_{ik}-P_{jk})^2+\gamma\sum_{k=1}^{c}\sum_{i=1}^{n}(P_{ik}-L_{ik})^2, \tag{8}$$

其中 c 为聚类个数,γ为正则化系数且用来平衡模型中的约束权重.

3 本文提出的聚类算法 EEGapc

为了解决无标签脑电信号聚类问题,提出了一种多项正则化约束的伪标签传播学习脑电信号聚类模型. 该方法主要结合了邻接矩阵传播优化、伪标签传播学习及伪标签分类器优化等模块,充分利用脑电信号之间的相关性进行聚类.

3.1 邻接矩阵传播优化

根据文献 [38] 提出的理论: 与邻接矩阵相关的拉普拉斯矩阵中特征值为 0 的数量与图中相连子图的个数相同. 可用该理论对脑电信号构建的图进行划分,实现脑电信号聚类. 对于原始脑电信号相似度构成的初始邻接矩阵 $\boldsymbol{M}\in\mathbb{R}^{n\times n}$, 文献 [38] 中提出的聚类方法的目的是学习到一个与 $\boldsymbol{M}$ 相当接近的伪邻接矩$\boldsymbol{S}\in\mathbb{R}^{n\times n}$, 并使得与之相关的拉普拉斯矩阵满足以下约束条件$\mathrm{rank}(\boldsymbol{\mathcal{L}}_S)=n-c$. 为了得到与真实邻接矩阵 $\boldsymbol{M}$ 最接近的伪邻接矩阵 $\boldsymbol{S}$, 定义其优化目标函数为

$$\begin{aligned}&\min_{\boldsymbol{S}}\|\boldsymbol{S}-\boldsymbol{M}\|_{\mathrm{F}}^2,\\ &\text{s.t. }\boldsymbol{S}\boldsymbol{1}_n=\boldsymbol{1}_n,\mathrm{rank}(\boldsymbol{\mathcal{L}}_S)=n-c,\end{aligned} \tag{9}$$

其中$\boldsymbol{1}_n=(1,1,\cdots,1)^{\mathrm{T}}$.

3.2 伪标签传播学习

结合标签传播学习函数, 见式(8), 正则化约束能更好地利用脑电信号之间的相互关系, 即脑电信号之间的相似度信息, 为未标记脑电信号进行标签学习. 不难看出, 标准的标签传播学习函数可收敛到最优解, 如定理 1 所述.

定理 1. 当$0<\alpha<1$时, 标签传播函数(式(8))可收敛到最优解: $\boldsymbol{P}^*=(1-\alpha)\boldsymbol{L}(\boldsymbol{I}-\alpha\boldsymbol{Q})^{-1}$.

证明. 为了不失一般性, 首先定义 $\boldsymbol{P}_0=\boldsymbol{L}$, 式(6)的迭代更新过程则可等价为

$$\boldsymbol{P}^t=(\alpha\boldsymbol{L})^t+(1-\alpha)\boldsymbol{L}\sum_{i=0}^{t-1}(\alpha\boldsymbol{Q})^i. \tag{10}$$

结合在邻接矩阵优化中定义的约束条件: $\boldsymbol{S}\boldsymbol{1}_n=\boldsymbol{1}_n$, 所以$\boldsymbol{Q}=\boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}=\boldsymbol{S}$, 且有$0<s_{ij}<1$. 同时, 因为$0<\alpha<1$, 所以可知$\lim_{t\to\infty}(\alpha\boldsymbol{P})^t=0$以及$\lim_{t\to\infty}\sum_{i=0}^{t-1}(\alpha\boldsymbol{S})^i=(\boldsymbol{I}-\alpha\boldsymbol{S})^{-1}$, 其中 $\boldsymbol{I}$ 表示单位矩阵.

综上所述, 可以得到

$$\boldsymbol{P}^*=(1-\alpha)\boldsymbol{L}(\boldsymbol{I}-\alpha\boldsymbol{S})^{-1}. \tag{11}$$

证毕.

定理 1 中的结论将用于后文求解多目标函数. 在本文提出的方法中, 对于脑电信号伪标签传播学习, 则可以通过正则化约束来获得最优的伪标签, 具体表示为

$$\min_{\boldsymbol{P}}\mathrm{tr}(\boldsymbol{P}\boldsymbol{\mathcal{L}}_S\boldsymbol{P}^{\mathrm{T}})+\gamma\|\boldsymbol{D}^{1/2}\boldsymbol{P}-\boldsymbol{D}^{-1/2}\boldsymbol{L}\|_{\mathrm{F}}^2, \tag{12}$$

其中$\|\cdot\|_{\mathrm{F}}$为 Frobenius 范数, γ为正则化参数. 结合脑电信号相似度邻接矩阵的约束条件$\boldsymbol{S}\boldsymbol{1}_n=\boldsymbol{1}_n$, 可得 $\boldsymbol{D}=\boldsymbol{I}$. 因此, 式(12)最终变换为

$$\min_{\boldsymbol{P}}\mathrm{tr}(\boldsymbol{P}\boldsymbol{\mathcal{L}}_S\boldsymbol{P}^{\mathrm{T}})+\gamma\|\boldsymbol{P}-\boldsymbol{L}\|_{\mathrm{F}}^2. \tag{13}$$

3.3 伪标签分类器优化

本文方法旨在将未标记的脑电信号划分到正确的聚簇中, 除了利用伪标签传播学习与脑电信号相似度邻接矩阵优化策略外, 还采用了正则化标签分类器优化来约束脑电信号的伪标签学习, 即通过最小化伪标签正则化误差来学习最优分类器, 具体表示为

$$\begin{aligned}&\min_{\boldsymbol{W}}\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}-\boldsymbol{L}\|_{\mathrm{F}}^2+\gamma\|\boldsymbol{W}\|_{\mathrm{F}}^2,\\ &\text{s.t. }\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{1}_n=\boldsymbol{1}_n.\end{aligned} \tag{14}$$

该正则化约束函数不仅可以学习满足约束条件的脑电信号标签分类器 $\boldsymbol{W}$, 同时还能进一步约束优化脑电信号的相似度邻接矩阵 $\boldsymbol{S}$, 有利于得到满足约束条件的解, 从而提升脑电信号的聚类性能.

3.4 聚类模型建立

充分利用脑电信号之间的信息传播, 在图的基础上, 基于伪标签传播学习建立了多项正则化约束下的多目标优化模型, 以确保脑电信号之间的信息传播正确性与聚类准确性. 建立的多目标优化模型主要包含了伪标签传播学习、脑电信号相似度邻接矩阵优化和伪标签分类器优化等模块. 具体的模型公式化为

$$\begin{aligned}&\min_{\boldsymbol{L},\boldsymbol{P},\boldsymbol{W},\boldsymbol{S}}\mathcal{F}=\frac{1}{2}\mathrm{tr}(\boldsymbol{P}\boldsymbol{\mathcal{L}}_S\boldsymbol{P}^{\mathrm{T}})+\frac{\gamma_1}{2}\|\boldsymbol{P}-\boldsymbol{L}\|_{\mathrm{F}}^2+\\ &\frac{\gamma_2}{2}\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}-\boldsymbol{L}\|_{\mathrm{F}}^2+\frac{\gamma_3}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2+\\ &\frac{\gamma_4}{2}\|\boldsymbol{S}-\boldsymbol{M}\|_{\mathrm{F}}^2,\\ &\text{s.t. }\boldsymbol{S}\boldsymbol{1}_n=\boldsymbol{1}_n,\mathrm{rank}(\boldsymbol{\mathcal{L}}_S)=n-c,\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}\boldsymbol{1}_n=\boldsymbol{1}_n,\end{aligned} \tag{15}$$

其中$\gamma_1,\gamma_2,\gamma_3,\gamma_4$为正则化参数, 用于平衡约束项的权重. 式(15)中的前 2 项为伪标签传播学习模块, 用于

在脑电信号相似度邻接矩阵的优化约束下学习得到接近真实结果的伪标签；第 3, 4 项为标签分类器的优化学习模块，同样也是为了在脑电信号相似度邻接矩阵约束下学习得到最优的伪标签；最后 1 项是邻接矩阵优化模块，用于学习最接近原始脑电信号相似度邻接矩阵的伪邻接矩阵，同时也为优化伪标签传播矩阵与标签分类器提供更好的传播信息.

3.5 脑电信号聚类算法 EEGapc 介绍

由于多目标优化函数，即式(15)，严格意义上是非凸的，并不能得到全局最优解，因此利用梯度下降策略求解目标函数的局部最优解. 通过固定其他变量的方式得到各个变量的局部最优解.

3.5.1 更新伪标签传播矩阵 $\boldsymbol{P}$

伪标签传播矩阵 $\boldsymbol{P}$ 的更新可以通过固定伪标签 $\boldsymbol{L}$、脑电信号相似度邻接矩阵 $\boldsymbol{S}$ 和标签分类器 $\boldsymbol{W}$ 来实现. 此时，目标函数(式(15))关于 $\boldsymbol{P}$ 的弱化函数变换为

$$\min_{\boldsymbol{P}} \mathcal{F}(\boldsymbol{P}) = \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\mathcal{L}_{\boldsymbol{S}}\boldsymbol{P}^{\mathrm{T}}) + \frac{\gamma_1}{2}\|\boldsymbol{P}-\boldsymbol{L}\|_{\mathrm{F}}^2. \tag{16}$$

式(16)的偏导为

$$\frac{\partial \mathcal{F}(\boldsymbol{P})}{\partial \boldsymbol{P}} = \boldsymbol{P}\mathcal{L}_{\boldsymbol{S}} + \gamma_1(\boldsymbol{P}-\boldsymbol{L}). \tag{17}$$

令 $\dfrac{\partial \mathcal{F}(\boldsymbol{P})}{\partial \boldsymbol{P}} = 0$，得到伪标签传播矩阵的最优解 $\boldsymbol{P}^*$ 为

$$\boldsymbol{P}^* = \gamma_1\boldsymbol{L}(\mathcal{L}_{\boldsymbol{S}} + \gamma_1\boldsymbol{I})^{-1}. \tag{18}$$

式(18)与式(11)一致.

3.5.2 更新伪标签 $\boldsymbol{L}$

为了更新伪标签 $\boldsymbol{L}$，同样固定其他变量即可. 因此，目标函数(式(15))关于 $\boldsymbol{L}$ 则可以弱化为

$$\min_{\boldsymbol{L}} \mathcal{F}(\boldsymbol{L}) = \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\mathcal{L}_{\boldsymbol{S}}\boldsymbol{P}^{\mathrm{T}}) + \frac{\gamma_1}{2}\|\boldsymbol{P}-\boldsymbol{L}\|_{\mathrm{F}}^2 + \frac{\gamma_2}{2}\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}-\boldsymbol{L}\|_{\mathrm{F}}^2. \tag{19}$$

式(19)关于 $\boldsymbol{L}$ 的偏导为

$$\frac{\partial \mathcal{F}(\boldsymbol{L})}{\partial \boldsymbol{L}} = (\boldsymbol{P}\mathcal{L}_{\boldsymbol{S}} + \gamma_1\boldsymbol{P} - \gamma_1\boldsymbol{L})\frac{\partial \boldsymbol{P}}{\partial \boldsymbol{L}} - \gamma_1(\boldsymbol{P}-\boldsymbol{L}) - \gamma_2(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}-\boldsymbol{L}). \tag{20}$$

根据式(18)，$\dfrac{\partial \boldsymbol{P}}{\partial \boldsymbol{L}} = \gamma_1(\mathcal{L}_{\boldsymbol{S}} + \gamma_1\boldsymbol{I})^{-1}$，将其带入式(20)，得到最终的 $\boldsymbol{L}$ 偏导为

$$\frac{\partial \mathcal{F}(\boldsymbol{L})}{\partial \boldsymbol{L}} = \boldsymbol{L}\left((\gamma_1+\gamma_2)\boldsymbol{I} - \gamma_1^2(\mathcal{L}_{\boldsymbol{S}} + \gamma_1\boldsymbol{I})^{-1}\right) - \gamma_2\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}. \tag{21}$$

同理，$\dfrac{\partial \mathcal{F}(\boldsymbol{L})}{\partial \boldsymbol{L}} = 0$，得到伪标签的最优解 L^* 为

$$\boldsymbol{L}^* = \gamma_2\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}\left((\gamma_1+\gamma_2)\boldsymbol{I} - \gamma_1^2(\mathcal{L}_{\boldsymbol{S}} + \gamma_1\boldsymbol{I})^{-1}\right)^{-1}. \tag{22}$$

3.5.3 更新标签分类器 $\boldsymbol{W}$

同理，为了更新标签分类器 $\boldsymbol{W}$，固定其他变量. 目标函数(式(15))关于 $\boldsymbol{W}$ 的弱化函数变为

$$\min_{\boldsymbol{W}} \mathcal{F}(\boldsymbol{W}) = \frac{\gamma_2}{2}\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}-\boldsymbol{L}\|_{\mathrm{F}}^2 + \frac{\gamma_3}{2}\|\boldsymbol{W}\|_{\mathrm{F}}^2. \tag{23}$$

式(23)关于 $\boldsymbol{W}$ 的偏导可表示为

$$\frac{\partial \mathcal{F}(\boldsymbol{W})}{\partial \boldsymbol{W}} = (\gamma_2\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + \gamma_3\boldsymbol{I})\boldsymbol{W} - \gamma_2\boldsymbol{S}\boldsymbol{L}^{\mathrm{T}}. \tag{24}$$

令 $\dfrac{\partial \mathcal{F}(\boldsymbol{W})}{\partial \boldsymbol{W}} = 0$，标签分类器的最优解 $\boldsymbol{W}^*$ 为

$$\boldsymbol{W}^* = \gamma_2(\gamma_2\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + \gamma_3\boldsymbol{I})^{-1}\boldsymbol{S}\boldsymbol{L}^{\mathrm{T}}. \tag{25}$$

3.5.4 更新邻接矩阵 $\boldsymbol{S}$

固定伪标签 $\boldsymbol{L}$、伪标签传播矩阵 $\boldsymbol{P}$ 和标签分类器 $\boldsymbol{W}$，目标函数(式(15))关于脑电信号相似度邻接矩阵 $\boldsymbol{S}$ 可以弱化为

$$\min_{\boldsymbol{S}} \mathcal{F}(\boldsymbol{S}) = \frac{1}{2}\mathrm{tr}(\boldsymbol{P}\mathcal{L}_{\boldsymbol{S}}\boldsymbol{P}^{\mathrm{T}}) + \frac{\gamma_1}{2}\|\boldsymbol{P}-\boldsymbol{L}\|_{\mathrm{F}}^2 + \frac{\gamma_2}{2}\|\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}-\boldsymbol{L}\|_{\mathrm{F}}^2 + \frac{\gamma_4}{2}\|\boldsymbol{S}-\boldsymbol{M}\|_{\mathrm{F}}^2. \tag{26}$$

由于式(26)不是一个严格意义上的凸函数，为了得到其最优解，采用了梯度下降方法对其进行求解. 即，设定一个学习率 η，让 $\boldsymbol{S}$ 梯度下降，逼近最优解：$\boldsymbol{S}_{t+1} = \boldsymbol{S}_t - \eta\nabla\boldsymbol{S}_t$. 同理，式(26)的偏导可表示为

$$\frac{\partial \mathcal{F}(\boldsymbol{S})}{\partial \boldsymbol{S}} = (\boldsymbol{P}\mathcal{L}_{\boldsymbol{S}} + \gamma_1\boldsymbol{P} - \gamma_1\boldsymbol{L})^{\mathrm{T}}\frac{\partial \boldsymbol{P}}{\partial \boldsymbol{S}} + \frac{1}{2}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{P}\frac{\partial \mathcal{L}_{\boldsymbol{S}}}{\partial \boldsymbol{S}} + \gamma_2\boldsymbol{W}(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}-\boldsymbol{L}) + \gamma_4(\boldsymbol{S}-\boldsymbol{M}). \tag{27}$$

根据目标函数(式(15))中的约束条件 $\boldsymbol{S}\boldsymbol{1}_n = \boldsymbol{1}_n$，可以得到邻接矩阵的对角矩阵 $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{S}\boldsymbol{1}_n) = \boldsymbol{I}$，所以可知 $\mathcal{L}_{\boldsymbol{S}} = \boldsymbol{D} - \boldsymbol{S} = \boldsymbol{I} - \boldsymbol{S}$，最后可以得到

$$\frac{\partial \mathcal{L}_{\boldsymbol{S}}}{\partial \boldsymbol{S}} = \frac{\partial \boldsymbol{D}}{\partial \boldsymbol{S}} - \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{S}} = \frac{\partial \boldsymbol{I}}{\partial \boldsymbol{S}} - \boldsymbol{I} = \boldsymbol{0} - \boldsymbol{I}. \tag{28}$$

同时，根据式(18)，得到

$$\frac{\partial \boldsymbol{P}}{\partial \boldsymbol{S}} = \frac{\partial \boldsymbol{P}}{\partial \mathcal{L}_{\boldsymbol{S}}}\frac{\partial \mathcal{L}_{\boldsymbol{S}}}{\partial \boldsymbol{S}} = \gamma_1\boldsymbol{L}(\mathcal{L}_{\boldsymbol{S}} + \gamma_1\boldsymbol{I})^{-1}\left((\mathcal{L}_{\boldsymbol{S}} + \gamma_1\boldsymbol{I})^{-1}\right)^{\mathrm{T}}. \tag{29}$$

最后，根据式(27)~(29)，脑电信号相似度邻接矩阵可以更新为

$$\boldsymbol{S}_{t+1} = \boldsymbol{S}_t - \eta\nabla\boldsymbol{S}_t = \boldsymbol{S}_t - \eta\frac{\partial \mathcal{F}(\boldsymbol{S}_t)}{\partial \boldsymbol{S}_t}. \tag{30}$$

3.5.5 脑电信号聚类算法 EEGapc

根据多目标优化函数中各个变量的更新求解过程，提出的脑电信号聚类算法 EEGapc 的具体描述见算法 1.

算法 1. 脑电信号聚类算法 EEGapc.

输入：脑电信号 $\boldsymbol{E} \in \mathbb{R}^{n\times m}$，正则化参数 $\gamma_1, \gamma_2, \gamma_3, \gamma_4$，学习率 η，最大迭代次数 $i_{\max}$；

输出：$\boldsymbol{L}^*, \boldsymbol{P}^*, \boldsymbol{W}^*, \boldsymbol{S}^*$.

① 计算脑电信号相似度初始邻接矩阵 $\boldsymbol{M}$；

② 初始化 $\boldsymbol{L}_0, \boldsymbol{P}_0, \boldsymbol{W}_0, \boldsymbol{S}_0$；

③ while 目标函数未收敛 do

④ 计算脑电信号相似度初始邻接矩阵 $\boldsymbol{M}$;
⑤ 更新伪邻接矩阵 $\boldsymbol{S}_{t+1}$:
⑥ for $i=1:i_{\max}$ do
⑦ 根据式(27)~(29)计算 $\nabla \boldsymbol{S}_i=\dfrac{\partial \mathcal{F}(\boldsymbol{S}_i)}{\partial \boldsymbol{S}_i}$;
⑧ $\boldsymbol{S}_{i+1}=\boldsymbol{S}_i-\eta\nabla \boldsymbol{S}_i$;
⑨ end for
⑩ $\boldsymbol{S}_{t+1}=\boldsymbol{S}_{i_{\max}+1}$;
⑪ 更新伪标签分类器 $\boldsymbol{W}_{t+1}$:
⑫ $\boldsymbol{W}_{t+1}=\gamma_2(\gamma_2\boldsymbol{S}_{t+1}\boldsymbol{S}_{t+1}^{\mathrm{T}}+\gamma_3\boldsymbol{I})^{-1}\boldsymbol{S}_{t+1}\boldsymbol{L}_t^{\mathrm{T}}$;
⑬ 更新伪标签传播矩阵 $\boldsymbol{P}_{t+1}$:
⑭ $\boldsymbol{P}_{t+1}=\gamma_1\boldsymbol{L}_t(\boldsymbol{\mathcal{L}}_{S_{t+1}}+\gamma_1\boldsymbol{I})^{-1}$;
⑮ 更新伪标签 $\boldsymbol{L}_{t+1}$:
⑯ $\boldsymbol{L}_{t+1}=\gamma_2\boldsymbol{W}_{t+1}^{\mathrm{T}}\boldsymbol{S}_{t+1}\left((\gamma_1+\gamma_2)\boldsymbol{I}-\gamma_1^2(\boldsymbol{\mathcal{L}}_{S_{t+1}}+\gamma_1\boldsymbol{I})^{-1}\right)^{-1}$;
⑰ $t=t+1$;
⑱ end while
⑲ return $\boldsymbol{L}^*=\boldsymbol{L}_{t+1},\boldsymbol{P}^*=\boldsymbol{P}_{t+1},\boldsymbol{W}^*=\boldsymbol{W}_{t+1},\boldsymbol{S}^*=\boldsymbol{S}_{t+1}$.

3.6 收敛性分析

由于邻接矩阵更新函数并不是一个严格意义上的凸函数, 故不能有效得到其全局最优解. 因此, 采用了梯度下降策略来获得其局部最优解. EEGapc 可以在设定好的学习率条件下收敛到局部最优解.

引理 1. 对于函数 $f(x):\mathbb{R}^n\to\mathbb{R}$, 其梯度下降更新策略为 $x_{t+1}=x_t-\eta\nabla f(x)$. 在足够小的学习率 η 下, 该函数可收敛到最优解 x^*.

证明. 根据文献 [42], 对于函数 f, 其梯度下降 ∇f 在尺度系数 $\mu>0$ 条件下满足 Lipschitz 连续, 即

$$\forall x,y,\ 0\leqslant\|\nabla f(x)-\nabla f(y)\|^2\leqslant\mu\|x-y\|^2. \tag{31}$$

对 $f(y)$ 进行泰勒二项展开可得

$$f(y)=f(x)+\nabla f(x)^{\mathrm{T}}(y-x)+\frac{1}{2}\nabla^2 f(x)\|y-x\|^2. \tag{32}$$

对式(32)进行变换可得

$$f(x)-f(y)=\nabla f(x)^{\mathrm{T}}(x-y)-\frac{1}{2}\nabla^2 f(x)\|x-y\|^2. \tag{33}$$

结合梯度下降更新策略 $x_{t+1}=x_t-\eta\nabla f(x)$, 可得

$$\|x_{t+1}-x^*\|^2=\|x_t-\eta\nabla f(x)-x^*\|^2=\|x_t-x^*\|^2-2\eta\nabla f(x)^{\mathrm{T}}(x_t-x^*)+\eta^2\|\nabla f(x)\|^2. \tag{34}$$

结合式(33), 可以得到

$$f(x_t)-f(x^*)\leqslant\nabla f(x_t)^{\mathrm{T}}(x_t-x^*)-\frac{1}{2\mu}\|\nabla f(x_t)-\nabla f(x^*)\|^2. \tag{35}$$

又因为 $f(x_t)-f(x^*)\neq 0$, 以及 $\nabla f(x^*)=0$, 所以不等式(35)等价为

$$\frac{1}{2\mu}\|\nabla f(x_t)\|^2\leqslant\nabla f(x_t)^{\mathrm{T}}(x_t-x^*). \tag{36}$$

综上所述,

$$\|x_{t+1}-x^*\|^2\leqslant\|x_t-x^*\|^2-\eta\left(\frac{1}{\mu}-\eta\right)\|\nabla f(x_t)\|^2. \tag{37}$$

即, 当学习率为 $0<\eta<\dfrac{1}{u}$ 时, $f(x)$ 函数可以收敛到最优解 x^*. 证毕.

结合引理 1, EEGapc 在足够小的学习率 η 条件下可以收敛到局部最优解, 如定理 2 所述.

定理 2. 当学习率满足 $\eta<\dfrac{\tau}{\sum_{i=1}^{n}4w_i^2a_i^4}$ 且 τ 为 $\sum_{i=1}^{n}w_i^2a_i^2\|a_i(x+y)-2b_i\|^2$ 的下界时, EEGapc 将收敛到局部最优解 x^*.

证明. 针对式(15)(16)(19)(23)(26)中的目标函数及关于各参数弱化后的表达形式, 结合文献 [33], 我们先构建以下形式的表达式:

$$f(x)=\sum_{i=1}^{n}w_i(a_ix-b_i)^2, \tag{38}$$

其中 $w_i,a_i,b_i\in\mathbb{R}$.

根据引理 1, 容易得出式(38)可以收敛到最优解, 并且有

$$f(x)-f(y)=\sum_{i}^{n}w_ia_i(a_i(x+y)-2b_i)(x-y). \tag{39}$$

同时, 可以得到

$$\|f(x)-f(y)\|^2=\sum_{i}^{n}w_i^2a_i^2\|a_i(x+y)-2b_i\|^2\|x-y\|^2. \tag{40}$$

同理可得,

$$\nabla f(x)-\nabla f(y)=f(x)'-f(y)'=\sum_{i=1}^{n}2w_ia_i^2(x-y). \tag{41}$$

相应地,

$$\|\nabla f(x)-\nabla f(y)\|^2=\sum_{i=1}^{n}4w_i^2a_i^4\|x-y\|^2. \tag{42}$$

综上所述, 结合引理 1 中的不等式(37), 当学习率为 $\eta<\dfrac{\tau}{\sum_{i=1}^{n}4w_i^2a_i^4}$, 且 τ 为 $\sum_{i=1}^{n}w_i^2a_i^2\|a_i(x+y)-2b_i\|^2$ 的下界时, 构建的式(31)利用梯度下降算法可以收敛到最优解 x^*. 显然, 本文提出的聚类模型目标函数(式(15))及各变量弱化后的表达式均可以转换为式(38)的形式. 因此, 基于梯度下降策略的算法 EEGapc 在给定的合适学习率条件下可以收敛到局部最优解. 证毕.

3.7 模型初始化

由于本文提出的脑电信号聚类模型所变换的多

目标优化函数并不是一个严格意义上的凸函数，不能得到全局最优解. 因此，本文采用了梯度下降策略来求解其局部最优解. 对于 EEGapc 而言，合适的初始化有利于得到较好的聚类性能，因此，对 EEGapc 初始化也进行了相应介绍.

3.7.1 伪邻接矩阵$\boldsymbol{S}_0$初始化

EEGapc 其中一个优化目标是学习到相对最优的脑电信号相似度伪邻接矩阵$\boldsymbol{S}\in\mathbb{R}^{n\times n}$，使其接近于脑电信号的原始相似度矩阵$\boldsymbol{M}\in\mathbb{R}^{n\times n}$. 结合多目标优化函数中的约束条件$\boldsymbol{S}\boldsymbol{1}_n=\boldsymbol{1}_n$，$\boldsymbol{1}_n=(1,1,\cdots,1)^{\mathrm{T}}$，故本文将$\boldsymbol{S}_0$初始化为$\boldsymbol{S}_0=\boldsymbol{M}$，并进行归一化使其满足$\boldsymbol{S}\boldsymbol{1}_n=\boldsymbol{M}\boldsymbol{1}_n=\boldsymbol{1}_n$.

3.7.2 伪标签$\boldsymbol{L}_0$初始化

本文通过 k-means 算法先粗略为无标签脑电信号分配标签，即利用 k-means 将 n 条脑电信号划分到 c 个聚簇中，并对其标签赋予伪标签$\boldsymbol{L}_0\in\mathbb{R}^{c\times n}$；然后再根据多目标优化函数的更新策略对其进行优化，使其满足目标函数的最小化约束要求.

3.7.3 伪标签传播矩阵$\boldsymbol{P}_0$初始化

结合初始化的脑电信号相似度邻接矩阵$\boldsymbol{S}_0$和初始化伪标签$\boldsymbol{L}_0$，标签传播矩阵的初始化$\boldsymbol{P}_0\in\mathbb{R}^{c\times n}$可根据式(11)实现，即

$$\boldsymbol{P}_0=(1-\alpha)\boldsymbol{L}_0(\boldsymbol{I}-\alpha\boldsymbol{S}_0)^{-1}. \tag{43}$$

根据文献[40]，通常情况下$\alpha=0.99$.

3.7.4 伪标签分类器$\boldsymbol{W}_0$初始化

为了得到相对最优的脑电信号聚类结果，EEGapc 利用标签分类器的优化对伪标签传播学习进行了约束，该约束的目的是为了学习与最初的脑电信号相似度邻接矩阵接近的标签分类器，提升脑电信号之间的信息传播准确性，并最终将 n 条脑电信号更准确地划分到 c 个聚类中. 因此，本文将标签分类器$\boldsymbol{W}_0\in\mathbb{R}^{n\times c}$初始化为

$$\boldsymbol{W}_0=(\boldsymbol{L}_0\boldsymbol{M}^{-1})^{\mathrm{T}}. \tag{44}$$

利用 3.7.1~3.7.4 节中所述的 4 个模块(矩阵变量)的初始化，EEGapc 可以在预先设定好的学习率η条件下进行迭代优化，直至目标函数收敛到局部最优，从而最终得到脑电信号的聚类结果.

3.8 时间复杂度分析

如算法 1 所示，采用了梯度下降的策略求解多目标优化函数的局部最优解，该过程需要通过多次迭代更新、优化，也是算法 1 最主要的时间消耗. 从理论角度分析了算法 EEGapc 的时间复杂度.

在算法 1 中，首先计算原始脑电信号相似度矩阵 $\boldsymbol{M}$，如算法第①行，其时间复杂度为 $O(n^2)$，其中 n 表示脑电信号的条数；如算法第②行所示，初始化$\boldsymbol{L}_0$，$\boldsymbol{P}_0,\boldsymbol{W}_0,\boldsymbol{S}_0$，分别需要的时间为 $O(n^2)$，$O(n^3+cn^2+n^2)$，$O(n^3+n^2+cn)$，$O(n^2)$. 当各个变量初始化后，算法 EEGapc 进行迭代优化、更新，直到目标函数收敛到局部最优. 该过程需要多次迭代，如算法第③~⑱行. 对于每一次迭代，如算法第⑥~⑩行，利用梯度下降更新 $\boldsymbol{S}$ 的时间复杂度为 $O(i_{\max}(2n^3+6cn^2+3n^2+3cn))$；如算法第⑫行所示，更新 $\boldsymbol{W}$ 的时间复杂度为 $O(2n^3+2cn^2+cn)$；更新 $\boldsymbol{P}$(如算法第⑭行)的时间复杂度为 $O(n^3+cn^2+n^2)$；更新 $\boldsymbol{L}$ 的时间复杂度为 $O(2n^3+2cn^2+2n^2+cn)$，如算法第⑯行所示. 假设目标函数经过 $iter$ 次迭代优化后收敛到局部最优，则整个过程的时间复杂度为 $O(iter\times(5n^3+5cn^2+3n^2+2cn+i_{\max}(2n^3+6cn^2+3n^2+3cn)))$. 综上，包含初始化与迭代优化过程的整个算法时间复杂度为 $O(2n^3+2cn^2+4n^2+cn+iter(5n^3+5cn^2+3n^2+2cn+i_{\max}(2n^3+6cn^2+3n^2+3cn)))$. 实际上，采用的脑电信号数据集类别个数 c 远小于数据条数 n，所以算法 EEGapc 的实际时间复杂度可以表示为 $O(n^3+iter(n^3+i_{\max}n^3))$. 不难看出，一旦初始化得到脑电信号的原始相似度，EEGapc 在聚类过程中的时间消耗仅与脑电信号的条数有关，而与其长度无关. 而且通常情况下，脑电信号的条数也远远小于其长度，这也使得 EEGapc 在聚类过程中具有较高的效率.

4 实验与讨论

4.1 数据集

采用了目前应用广泛的公共脑电信号数据集对 EEGapc 进行性能分析，数据主要来源于国际 BCI 竞赛①，主要包括：1)慢皮质脑电信号(slow cortical potential, SCP)，其中一个数据集 II_Ia 采集于 1 个健康被试，另一个数据集 II_Ib 来源于 1 个渐冻症患者. 2) 四分类运动想象脑电信号，该数据集采集了 9 个健康被试的左手、右手、双脚和舌头的运动想象脑电信号，分别对应数据集 IV_2a_s1, IV_2a_s2, IV_2a_s3, IV_2a_s4, IV_2a_s5, IV_2a_s6, IV_2a_s7, IV_2a_s8, IV_2a_s9. 3) 简单二分类运动想象脑电信号，采集了 3 个健康被试的左手和右手运动想象脑电信号，分别对应数据集 IV_2b_s1, IV_2b_s2, IV_2b_s3. 14 个脑电信号数据集的具体描述如表 1 所示. 表 1 中，1 条脑电信号是由所有有效通道在同一时间段采集到的同一脑任务刺激所产生的脑电信号依次拼接而成的，

① 网站分别为 https://www.bbci.de/competition/ii/, https://www.bbci.de/competition/iv/

Table 1 EEG Data Sets

表 1 脑电信号数据集

数据集	数据描述	条数	数据长度	通道数	类别数
II_Ia	健康被试慢皮质脑电信号	268	5 377	6	2
II_Ib	渐冻症患者慢皮质脑电信号	>200	8 065	7	2
IV_2a_s1 IV_2a_s2 IV_2a_s3 IV_2a_s4 IV_2a_s5 IV_2a_s6 IV_2a_s7 IV_2a_s8 IV_2a_s9	来自 9 位健康被试的复杂四分类运动想象脑电信号，包括左手、右手、双脚和舌头的运动想象	288	16 501	22	4
IV_2b_s1 IV_2b_s2 IV_2b_s3	来自 3 位健康被试的简单二分类运动想象脑电信号，包括左手、右手的运动想象	120	940	3	2

该操作便于衡量脑电信号之间的相关性，即相似度. 同时，该处理方式也比较适用于其他基于时间序列的聚类算法进行脑电信号的聚类，便于与 EEGapc 进行聚类性能比较. 另外，采用的脑电信号已经人工剔除了无用采集通道及伪迹数据，如肌电信号、眼电信号等.

4.2 评价指标

采用 4 种聚类评价指标：*NMI*（normalized mutual information）[43]，*ARI*（adjusted rand index）[44]，*F-score*[45]，*kappa*[46] 对提出的脑电信号聚类算法性能进行评价.

1）*NMI*

NMI 衡量聚类结果与真实结果之间的相似程度，即可以有效衡量聚类算法的聚类准确性. 定义为

$$NMI=\frac{\sum_{i=1}^{v}\sum_{j=1}^{v}|C_i\cap C_j|\log\frac{N|C_i\cap C_j|}{|C_i||C_j|}}{\sqrt{\left(\sum_{i=1}^{v}|C_i|\log\frac{|C_i|}{N}\right)\left(\sum_{j=1}^{v}|C_j|\log\frac{|C_j|}{N}\right)}},\tag{45}$$

其中$|C_i|$与$|C_j|$分别表示聚类 C_i 与 C_j 中的元素个数即脑电信号条数，v 表示脑电信号的类别数，N 表示脑电信号的总条数.

2）*ARI*

ARI 是标准 rand index（*RI*）的一种改进，$RI=\frac{TP+TN}{TP+FP+TN+FN}$. *ARI* 可以衡量 2 个随机划分的匹配程度，即可以在类标不明确的情况下衡量聚类结果与真实结果的匹配情况. *ARI* 的具体定义为

$$ARI=\frac{\sum_{i,j}\binom{n_{ij}}{2}-\left[\sum_{i}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_{i}\binom{a_i}{2}+\sum_{j}\binom{b_j}{2}\right]-\left[\sum_{i}\binom{a_i}{2}\sum_{j}\binom{b_j}{2}\right]\Big/\binom{n}{2}},\tag{46}$$

其中 n_{ij} 表示类别μ_i与类别v_j交集中的元素个数，a_i 表示类别μ_i包含的元素个数，b_i 表示类别v_i包含的元素个数，n 为所有数据元素的总个数.

3）*F-score*

F-score 采用了一种非平等加权的方式来衡量聚类算法的性能，即对召回率增加了尺度权重$\beta\geqslant 0$. 其具体定义为

$$F\text{-}score=\frac{(1+\beta^2)pr}{\beta^2 p+r},\tag{47}$$

其中精确度 $p=\frac{TP}{TP+FP}$，召回率 $r=\frac{TP}{TP+FN}$. *TP*，*FP*，*FN* 分别表示真阳性数据个数、假阳性数据个数以及假阴性数据个数.

4）*kappa*

kappa 用以估计出现一致性的概率，即可用于衡量聚类决策（结果）的一致性，其具体定义为

$$kappa=\frac{\overline{P}-\overline{P}_{\mathrm{e}}}{1-\overline{P}_{\mathrm{e}}},\tag{48}$$

其中

$$\overline{P}=\frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k}n_{ij}^2-Nn\right),\overline{P}_{\mathrm{e}}=\sum_{j=1}^{k}\left(\frac{1}{Nn}\sum_{i=1}^{N}n_{ij}\right)^2.$$

4.3 对比算法

为了评估 EEGapc 的脑电信号聚类性能，将其与其他时间序列聚类算法进行比较. 主要对比了 8 种聚类算法，包括经典的 k-means 型聚类算法 KMM[19]、基

于距离(相似度)的聚类算法 K-SC[21]、基于密度的聚类算法 DP[26]、基于特征选择的聚类算法 RUFS[29]、基于形态的聚类算法 kShape[34]、谱聚类 USPEC[36]、基于邻接矩阵传播优化的聚类算法 AP[40] 以及基于最大加权团的脑电信号聚类算法 mwcEEGc[15].

1)KMM. 该算法是一种 k-means 型聚类算法, 其首先设定多个聚类子中心, 然后通过优化策略更新子聚类中心及子聚类结果, 经过多次优化、更新及融合, 将数据划分到 k 个聚类中.

2)K-SC. 该算法首先采用一种尺度可伸缩的相似度衡量方法迭代搜寻聚类中心, 然后再结合数据与聚类中心的相似度进行数据划分与聚类中心更新, 直到聚类中心不再变化, 且所有数据被划分到相应的聚类中.

3)DP. 该算法根据数据之间构成的团体密度进行聚类, 即根据各数据与其周围数据的密度高低确定聚类中心, 然后再根据密度阈值将数据划分到对应的高密度数据团体中, 即聚类中.

4)RUFS. 该算法利用 $l_{2,1}$ 范式最小化与局部正则化后的正交矩阵实现数据的特征选择与降维, 然后将具有相似特征的数据划分到相应的聚类中.

5)kShape. 该算法利用标准化互相关衡量方法评估数据间的相关性, 并利用该方法得到数据差异性较大的数据形态, 最终结合学习到的形态特征将相似的数据划分到同一聚类.

6)USPEC. 该算法采用混合特征搜寻策略和 kNN 相结合构建稀疏邻接矩阵, 并利用转切(transfer cut)策略实现数据的聚类任务.

7)AP. 该算法于 2014 年发表在 Science 上, 其主要利用数据相似度邻接矩阵作为数据之间的传播信息, 该传播过程为无标签数据的标签化提供有效信息, 最终实现数据的聚类.

8)mwcEEGc. 该算法利用改进的 Fréchet 距离衡量脑电信号之间的相似度, 并将其映射为脑电信号无向加权图的边权重, 然后结合相似度阈值, 以搜寻图中最大加权团的方式实现脑电信号聚类.

以上 8 种对比方法的参数设定与相应文献的原文设定一致; 聚类数量根据脑电信号数据原本的类别数确定; 提出的 EEGapc 算法的参数 $\gamma_1,\gamma_2,\gamma_3,\gamma_4$ 采用网格搜寻的方式自动寻优, 即它们的取值范围设定为 $\{10^{-2},10^{-1},1,10^{1},10^{2}\}$; 学习率 η 设定为 0.01; 算法的最大迭代次数为 50 次. 本文将参数自动寻优后取得的最佳结果作为 EEGapc 的最终聚类结果. 另外, EEGapc 由 Matlab R2021b 实现, 且在 MacOS 12.1 (M1 Pro、8 核处理器、16 GB 内存)上运行. 对比算法均运行 10 次, 并取最优结果作为最终聚类结果.

4.4 实验及分析

4.4.1 聚类结果比较

为了评价 EEGapc 的脑电信号聚类性能, 比较了 EEGapc 与上述 8 种聚类算法在 *NMI*, *ARI*, *F-score*, *kappa* 这 4 个指标下的聚类结果, 分别如表 2~5 所示. 结果表明, 与 8 种不同的聚类算法相比, EEGapc 脑电信号聚类算法在 14 个真实脑电信号数据集上的聚类性能最好. 表 2~5 中的 *NMI* 均值、*ARI* 均值、*F-score* 均值和 *kappa* 均值表明了 EEGapc 在不同类型的无标签脑电信号上的聚类性能稳定性与优越性. 而且, EEGapc 脑电信号聚类算法与其他 8 种聚类算法中性能最好的 mwcEEGc 相比, 分别在 *NMI* 均值、*ARI* 均值、*F-score* 均值和 *kappa* 均值上分别提升了 86.88%, 58.01%, 6.29%, 61.17%. 另外, 表 2~5 中的显著性分析结果也表明, EEGapc 脑电信号聚类算法性能与 KMM, K-SC, DP, RUFS, kShape, USPEC, AP 的相比具有显著性优势. 潜在原因主要是因为 EEGapc 聚类算法利用多项正则化对伪标签传播矩阵优化进行了约束, 并且充分利用了以脑电信号相似度邻接矩阵为载体的传播信息, 为得到相对最优的脑电信号标签提供了更符合实际情况的优化方向.

另外, 为了更直观地展示各聚类算法在 14 个脑电信号数据集上的整体聚类性能(也可以理解为聚类算法的通用性), 对表 2~5 中聚类算法的 *NMI*, *ARI*, *F-score*, *kappa* 评价指标结果进行了排序: 评价指标越高, 排名数值越小(即排名越靠前); 然后, 计算出各聚类算法的平均评价指标排名, 平均排名越低, 表明聚类性能越好, 具体结果如图 1 所示. 图 1 中的平均排名结果同样表明 EEGapc 算法的整体聚类性能最佳, 且具有在不同类型脑电信号数据集上保持性能稳定的能力, 这也从侧面反映出 EEGapc 聚类算法可以应用到不同规模、类型的脑电信号数据集上.

4.4.2 聚类效率比较

本节主要比较 EEGapc 与其他聚类算法在 14 个脑电信号数据集上的聚类效率. 需要说明的是, 本文仅将固定正则化参数下的运行时间作为 EEGapc 的聚类时间消耗, 并没有将自动参数寻优的总体时间作为 EEGapc 的最终运行时间. 与不同聚类算法的比较结果如图 2 所示. 该结果展示了 EEGapc 聚类算法在所有的脑电信号数据集上都具有较好的运行效率, 聚类的时间消耗比大部分对比算法小. 正如 3.8 节的阐述, 在聚类过程中, EEGapc 的聚类时间消耗与脑电

Table 2 Performance Evaluation with *NMI* for Clustering Algorithms

表 2 聚类算法的 *NMI* 性能指标

数据集	KMM	K-SC	DP	RUFS	kShape	USPEC	AP	mwcEEGc	EEGapc（本文）
II_Ia	0.045 6	0.021 9	0.081 3	0.023 3	0.021 4	0.000 1	0.042 5	0.101 8	**0.147 5**
II_Ib	0.010 2	0.001 2	0.016 9	0.008 7	2.8872E-4	0.000 1	0.001 2	0.020 4	**0.038 6**
IV_2a_s1	0.028 2	0.02	0.014 2	0.009 4	0.017 3	0.017 6	0.003 2	0.031 1	**0.039 6**
IV_2a_s2	0.054 6	0.015 8	0.016 9	0.013	0.014 7	0.008 5	0.000 5	0.275 1	**0.642 5**
IV_2a_s3	0.038 1	0.031 1	0.034 1	0.017 7	0.016 3	0.023 1	0.010 3	0.290 5	**0.475 6**
IV_2a_s4	0.049 9	0.019	0.037 1	0.012	0.017 1	0.022 1	0.010 2	0.253 6	**0.586 3**
IV_2a_s5	0.045 2	0.021 1	0.014 3	0.006 4	0.008 5	0.027 3	0.001 4	0.052 3	**0.106**
IV_2a_s6	0.037 7	0.039 3	0.021	0.015 6	0.023 1	0.016 3	0.006 5	0.201 4	**0.365 2**
IV_2a_s7	0.032 6	0.007 3	0.022 7	0.017 5	0.019 7	0.018 1	0.014 9	0.033 8	**0.036 4**
IV_2a_s8	0.024 3	0.010 2	0.017 9	0.011 3	0.020 2	0.006 2	0.001 3	0.025 2	**0.028 3**
IV_2a_s9	0.026 9	0.017	0.014 4	0.005 3	0.020 5	0.009 5	0.005 6	0.028 1	**0.031 3**
IV_2b_s1	0.018 2	0.013	0.020 9	0.002 2	0.020 4	0.005 4	0.010 2	0.026 8	**0.045 6**
IV_2b_s2	0.062	0.007 4	0.015 6	0.01	0.009 9	0.000 2	0.009 6	0.065 8	**0.109 9**
IV_2b_s3	0.031 8	0.005 6	0.021 3	0.000 3	0.005	0.031 8	0.003 3	0.034 1	**0.039 9**
NMI 均值	0.036 1	0.016 4	0.024 9	0.010 9	0.015 3	0.013 3	0.008 6	0.102 9	**0.192 3**
显著性	$p<0.05$（*）	$p<0.01$（**）	$p<0.01$（**）	$p<0.01$（**）	$p<0.01$（**）	$p<0.01$（**）	$p<0.01$（**）	$p=0.19$（−）	−−

注: *表示显著, **表示十分显著, −表示不可比较；黑体数值表示最优值.

Table 3 Performance Evaluation with *ARI* for Clustering Algorithms

表 3 聚类算法的 *ARI* 性能指标

数据集	KMM	K-SC	DP	RUFS	kShape	USPEC	AP	mwcEEGc	EEGapc（本文）
II_Ia	0.001 1	0.000 8	0.068 7	0.028 4	0.025 8	−0.003 5	0.050 1	0.096 7	**0.119 7**
II_Ib	0.000 6	−0.003 9	0.008 8	0.007 1	−0.004 6	−0.004 9	−0.003 2	0.014 4	**0.048 2**
IV_2a_s1	0.000 2	0.002 4	0.004 7	0.000 6	0.004 7	0.003 3	−0.002 1	0.019 1	**0.024 1**
IV_2a_s2	0.000 2	−1.6723E-4	0.005 3	0.004 6	0.003 8	−0.001 6	−0.004 7	0.283 6	**0.464**
IV_2a_s3	0.000 1	−3.4709E-4	0.021 7	0.001 4	0.005 5	0.006 2	0.004 8	0.294 2	**0.351 9**
IV_2a_s4	0.000 2	0.003 3	0.022 9	0.000 4	0.005 4	0.010 1	0.004 7	0.305 7	**0.443 5**
IV_2a_s5	0.000 6	0.001 2	0.002 3	0.001	0.002 6	0.002 4	−0.003 8	0.028 8	**0.092 1**
IV_2a_s6	0.000 5	0.001 6	0.011 6	0.002 2	0.009 6	0.002 6	0.001 2	0.121 8	**0.272 5**
IV_2a_s7	0.000 6	−0.002	0.007 1	0.003 2	0.004 7	0.001 4	0.009	0.013 2	**0.019 4**
IV_2a_s8	0.000 2	−0.001 2	0.008 6	−0.001 2	**0.018**	−0.001 4	−0.004	0.011 6	0.015 5
IV_2a_s9	0.000 4	−1.2489E-4	0.005 1	−0.002 5	0.018 8	−0.000 4	0.000 3	0.014 2	**0.020 3**
IV_2b_s1	0.000 6	0.009 5	0.032 4	0.005	0.008 1	−0.001 1	0.005 4	0.036 7	**0.054 6**
IV_2b_s2	0.001 7	0.001 8	0.003 2	0.005 4	0.005 3	−0.008 2	0.003 2	0.075 6	**0.139 8**
IV_2b_s3	0	−0.000 8	0.010 3	−0.006 9	−0.001 5	0	−0.003 8	0.021 4	**0.046 5**
ARI 均值	0.000 5	0.000 9	0.015 2	0.003 5	0.007 6	0.000 4	0.004 1	0.095 5	**0.150 9**
显著性	$p<0.01$（**）	$p<0.01$（**）	$p<0.01$（**）	$p<0.01$（**）	$p<0.01$（**）	$p<0.01$（**）	$p<0.01$（**）	$p=0.30$（−）	−−

注: **表示十分显著, −表示不可比较；黑体数值表示最优值.

信号的条数有关, 与脑电信号的长度无关. 通常情况下, 脑电信号的条数远小于长度. 一旦脑电信号的相似度初始化, EEGapc 便能快速地进行迭代学习和收敛, 并最终实现脑电信号聚类. 与 EEGapc 相似, USPEC, DP, KMM, AP 的聚类时间消耗也主要与脑电信号的条数有关, 而不是其长度, 因此它们的聚类时间消耗也相对较小. USPEC, DP, KMM 的聚类时间之所以小于 EEGapc, 是因为它们在聚类过程中采用了相对简单的聚

Table 4 Performance Evaluation with *F-score* for Clustering Algorithms

表 4 聚类算法的 *F-score* 性能指标

数据集	KMM	K-SC	DP	RUFS	kShape	USPEC	AP	mwcEEGc	EEGapc（本文）
II_Ia	0.367 5	0.339	0.335 8	0.566 8	0.413 2	0.496 1	0.376	0.575	**0.629 4**
II_Ib	0.365 8	0.479 8	0.34	0.552 4	0.489 9	0.504 9	0.473 6	0.581 5	**0.614 2**
IV_2a_s1	0.246 5	0.178 7	0.255 7	0.182 3	0.22	0.201 6	0.229 2	**0.475 1**	0.319 8
IV_2a_s2	0.260 4	0.236 1	0.249 2	0.208	0.268 8	0.271 9	0.246 5	0.465 3	**0.5**
IV_2a_s3	0.243 1	0.232 6	0.245 3	0.242 7	0.256 3	0.299 7	0.291 7	**0.463 5**	0.323
IV_2a_s4	0.128 1	0.182	0.295 7	0.204 9	0.294 1	0.292 6	0.246 5	0.347 8	**0.493 1**
IV_2a_s5	0.145 3	0.215 3	0.219 8	0.205	0.242 8	0.291 2	0.256 9	0.271 9	**0.323 6**
IV_2a_s6	0.137	0.222 2	0.221 9	0.238 2	0.207 9	0.277 4	0.225 7	0.291 7	**0.318 9**
IV_2a_s7	0.130 6	0.168 6	0.206 9	0.241 2	0.218 5	0.207 9	0.225 7	0.265 4	**0.331 2**
IV_2a_s8	0.128 2	0.208 3	0.250 2	0.213 9	0.270 7	0.249 7	0.239 6	0.273 8	**0.341 5**
IV_2a_s9	0.121 5	0.150 2	0.231 8	0.177 3	0.313 4	0.237 6	0.243 1	0.295 9	**0.316 5**
IV_2b_s1	0.369 3	0.424	0.375	0.455	0.517 3	0.451 9	0.555 8	0.583 4	**0.625**
IV_2b_s2	0.386 6	0.473 2	0.375	0.517 5	0.440 7	0.491 6	0.527 8	0.568 4	**0.633 3**
IV_2b_s3	0.351 6	0.447 2	0.425	0.464 6	0.458 3	0.508 3	0.521 2	0.548 6	**0.616 2**
F-score 均值	0.241 5	0.282 7	0.287 7	0.319 3	0.329 4	0.341 6	0.332 8	0.429 1	**0.456 1**
显著性	$p<0.001$（***）	$p<0.01$（**）	$p<0.001$（***）	$p<0.05$（*）	$p<0.05$（*）	$p<0.05$（*）	$p<0.05$（*）	$p=0.61$（–）	––

注: *表示显著, **表示十分显著, ***表示极其显著, –表示不可比较；黑体数值表示最优值.

Table 5 Performance Evaluation with *kappa* for Clustering Algorithms

表 5 聚类算法的 *kappa* 性能指标

数据集	KMM	K-SC	DP	RUFS	kShape	USPEC	AP	mwcEEGc	EEGapc（本文）
II_Ia	0.030 3	0.036 6	0.234 6	0.179 1	0.171 1	0.009 3	0.232 6	0.238 8	**0.350 4**
II_Ib	0.02	0.03	0.088 7	0.12	0.02	0.01	0.04	0.12	**0.23**
IV_2a_s1	0.023 1	0.060 2	0.088	0.050 9	0.074 1	0.074 1	0.032 4	**0.202 7**	0.148 1
IV_2a_s2	0.032 4	0.013 9	0.092 6	0.060 2	0.074 1	0.064 8	0.013 9	0.211 3	**0.423 1**
IV_2a_s3	0.023 1	0.050 9	0.106 5	0.050 9	0.064 8	0.092 6	0.055 6	0.286 6	**0.328 7**
IV_2a_s4	0.027 8	0.069 4	0.078 7	0.046 3	0.064 8	0.078 7	0.055 6	0.275 3	**0.384 3**
IV_2a_s5	0.027 8	0.050 9	0.074 1	0.032 4	0.050 9	0.083 3	0.023 1	0.082 1	**0.189 8**
IV_2a_s6	0.023 1	0.050 9	0.083 3	0.060 2	0.111 1	0.069 4	0.041 7	0.131 2	**0.310 2**
IV_2a_s7	0.027 8	0.041 7	0.069 4	0.074 1	0.083 3	0.078 7	0.078 7	0.092 5	**0.115 7**
IV_2a_s8	0.027 8	0.050 9	0.083 3	0.050 9	0.088	0.041 7	0.023 1	0.101 9	**0.125**
IV_2a_s9	0.018 5	0.050 9	0.060 2	0.027 8	0.106 5	0.037	0.046 3	0.099 6	**0.125**
IV_2b_s1	0.033 3	0.166 7	0.149 7	0.05	0.116 7	0.083 3	0.116 7	0.140 6	**0.25**
IV_2b_s2	0.05	0.066 7	0.090 9	0.05	0.116 7	0.016 7	0.1	0.166	**0.383 3**
IV_2b_s3	0.016 7	0.016 7	0.111 1	0.016 7	0.083 3	0.016 7	0.066 7	0.082 4	**0.233 3**
kappa 均值	0.027 3	0.054	0.100 8	0.062 1	0.087 5	0.054	0.066 2	0.159 4	**0.256 9**
显著性	$p<0.001$（***）	$p<0.001$（***）	$p<0.001$（***）	$p<0.001$（***）	$p<0.001$（***）	$p<0.001$（***）	$p<0.001$（***）	$p<0.01$（**）	––

注: **表示十分显著, ***表示极其显著, –表示不可比较；黑体数值表示最优值.

类中心确定策略. 相反, RUFS, kShape, K-SC, mwcEEGc 等聚类算法的时间消耗不仅与脑电信号的条数有关, 更与脑电信号的长度有关, 因为它们都需要多次处理整条脑电信号, 从较长的脑电信号中识别、提取和搜寻特征、形态和聚类中心或多次迭代搜寻最大加权图, 直到所有脑电信号被划分到相应的聚类为止, 这大大增加了它们的时间成本. 特别地, kShape, K-SC, mwcEEGc 在聚类过程中用来计算脑电信号相似度的

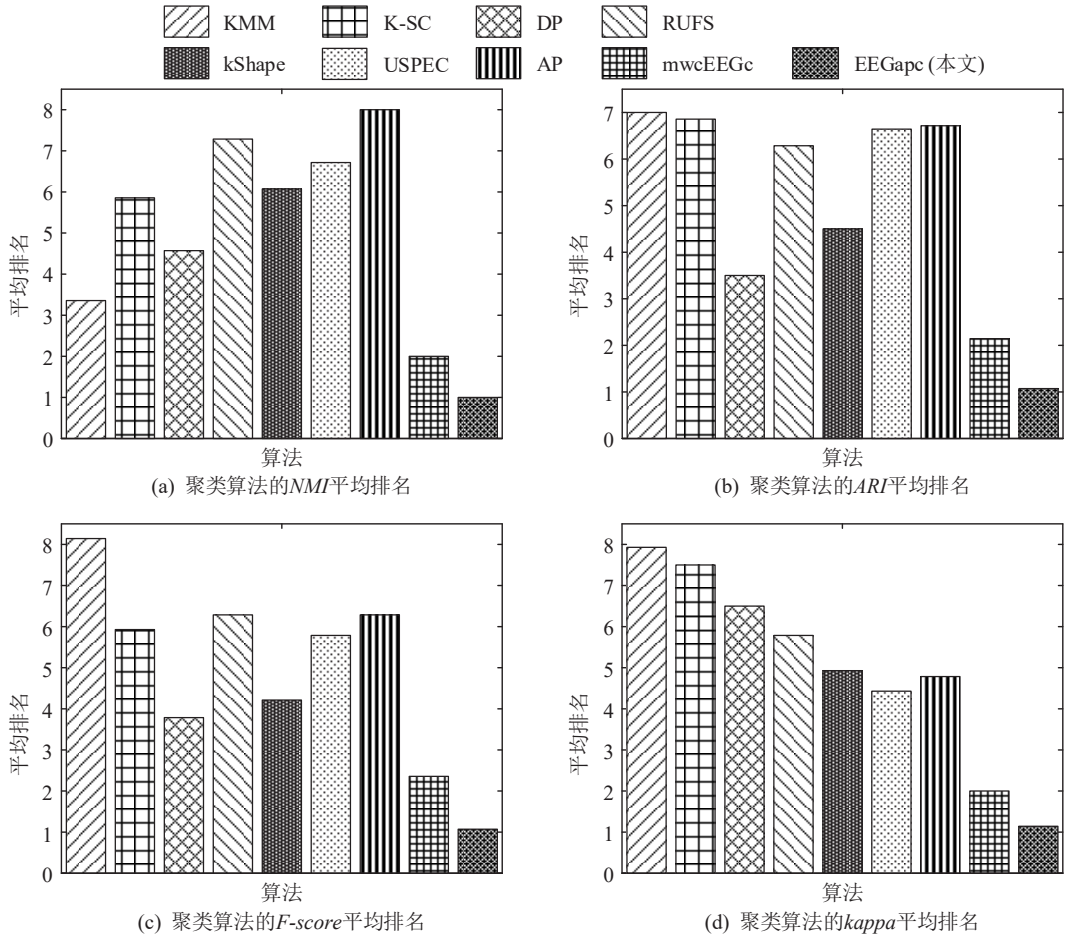


Fig. 1 Performance average rank of clustering algorithms

图 1 聚类算法的性能平均排名

方法复杂度较高, 这也是导致它们聚类效率较低的原因之一.

不难看出: 根据 14 个脑电信号数据集的规模, EEGapc 算法的聚类时间消耗并没有随着脑电信号数据集的规模增大而明显增加, 这也表明 EEGapc 对于规模更大的脑电信号在理论上也可以在较短时间内完成数据标签化. 同时, EEGapc 聚类算法的高效率也在一定程度上归功于它的快速收敛特性, 如图 3 所示.

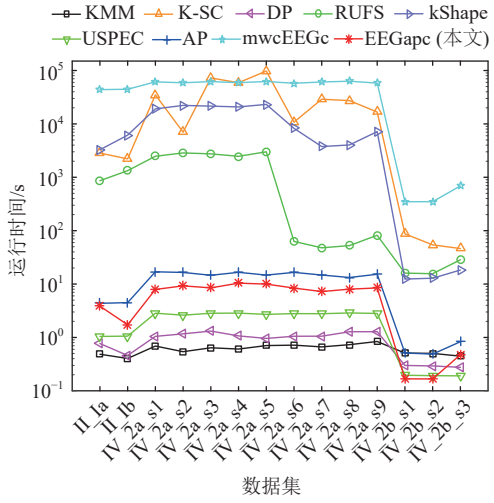


Fig. 2 Runtime comparison of clustering algorithms

图 2 聚类算法的运行时间比较

4.4.3 收敛性分析

图 3 展示了 EEGapc 聚类算法在 6 种不同规模、类型的脑电信号数据集 II_Ia, II_Ib, IV_2a_s1, IV_2a_s5, IV_2a_s9, IV_2b_s1 上的收敛情况, 其他数据集的收敛情况与图 3 相似, 故未在文中全部展示. 从图 3 可以看出, EEGapc 在不同类型脑电信号数据集上都能在较少的迭代后收敛, 这也表明 EEGapc 能很快速地得到多目标优化函数的局部最优解和脑电信号聚类结果. 同样, 图 3 中的结果也表明即使在规模更大的脑电信号数据集上, EEGapc 也能在较短的时间内收敛并得到聚类结果, 这有利于其在现实场景中的应用. 同理, 图 3 的收敛结果也照应了图 2 中 EEGapc 的聚类高效性.

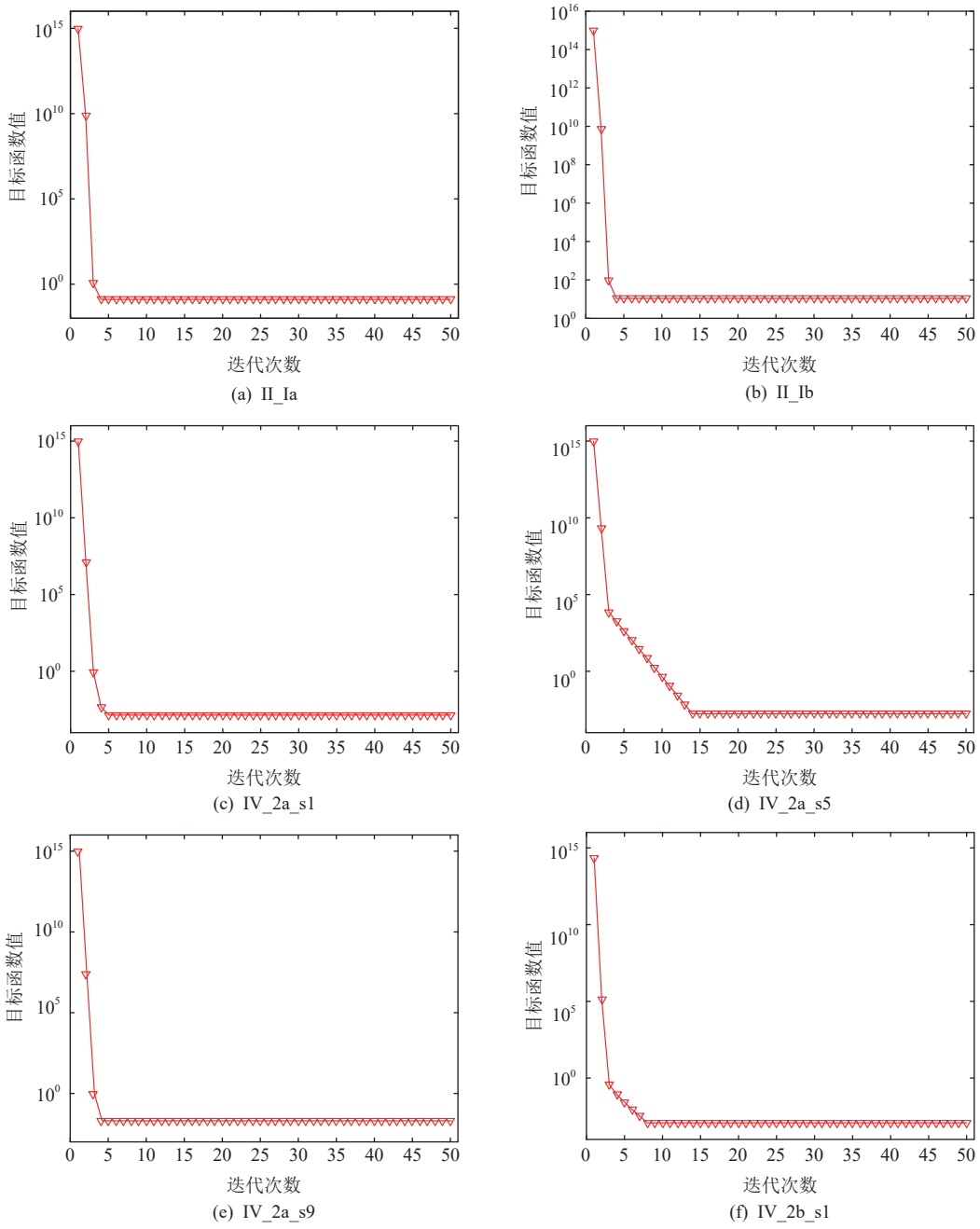


Fig. 3 Convergence for EEGapc on six types of EEG data sets

图 3 EEGapc 在 6 种脑电信号数据集上的收敛性

5 总结与展望

随着脑电信号数据中无标签数据的日益增加, 给基于脑电信号的应用, 如脑-机接口、疾病辅助诊断、康复, 甚至军事领域等, 都带来了巨大挑战. 由于现有的大部分针对脑电信号的研究及应用都集中在有监督分类方面, 而有监督的分类方法十分依赖数据中的标签信息, 所以其并不适用于处理无标签脑电信号这一新型数据. 为了解决该问题并拓展无标签脑电信号的应用领域, 提出了一种多项正则化约束的伪标签传播优化脑电信号聚类方法. 该方法主要采用了伪标签传播、脑电信号相似度邻接矩阵及标签分类器联合优化的方式实现无标签脑电信号的标签化与聚类, 并将该聚类模型转化为一个多目标优化问题, 随之提出了一种基于梯度下降策略的聚类算法 EEGapc. EEGapc 聚类算法能在较短时间内收敛到局部最优, 并实现无标签脑电信号的聚类. 结合 14 个真实的脑电信号数据集, 从 4 个评价指标比较了 8 种不同类型的时间序列聚类算法, 结果都表明了本

文提出的EEGapc算法具有最好的脑电信号聚类性能.

在未来的工作中,可以充分利用原始脑电信号数据中有限的先验信息,如少量的标签信息,对提出的模型进行约束,实现脑电信号的半监督分析,从而进一步提高无标签脑电信号的聚类性能.由于脑电信号是一种具有时间序列特性的生物电信号,因此,提出的聚类方法理论上也可以应用到其他基于时间序列的无监督分析领域中,能为该领域提供一定的方法指导.诚然,这也可以作为未来的一个研究内容和应用方向.另外,根据现有的研究,针对脑电信号数据的预处理,如时频域分析、时空域分析有助于提升脑电信号的分析性能,因此,其也可以作为今后的一个技术手段,辅助脑电信号聚类算法实现性能的进一步提升.

作者贡献声明: 代成龙提出了模型理论、算法思路、实验方案并撰写论文;李光辉提出指导意见,参与论文修改;李栋完成实验数据处理与实验分析;申佳华完成数据获取与实验分析;皮德常提出指导意见,参与论文修改.

参 考 文 献

[1] Glasser M F, Coalson T S, Robinson E C, et al. A multi-modal parcellation of human cerebral cortex[J]. Nature, 2016, 536(7615): 171−178

[2] Furui A, Onishi R, Takeuchi A, et al. Non-gaussianity detection of EEG signals based on a multivariate scale mixture model for diagnosis of epileptic seizures[J]. IEEE Transactions on Biomedical Engineering, 2021, 68(2): 515−525

[3] Yu Haitao, Lei Xinyu, Song Zhenxi, et al. Supervised network-based fuzzy learning of EEG signals for Alzheimer's disease identification[J]. IEEE Transactions on Fuzzy Systems, 2020, 28(1): 60−71

[4] Jayaram V, Widmann N, Forster C, et al. Brain-computer interfacing in amyotrophic lateral sclerosis: Implications of a resting-state EEG analysis [C]// Proc of the 37th Annual Int Conf of the IEEE Engineering in Medicine and Biology Society. Piscataway, NJ: IEEE, 2015: 6979−6982

[5] Ali Shah S A, Zhang Lei, Bais A. Dynamical system based compact deep hybrid network for classification of Parkinson disease related EEG signals[J]. Neural Networks, 2020, 130: 75−84

[6] Liang Xu, Wang Weiqun, Hou Zengguang, et al. Interactive control methods for rehabilitation robot [J]. SCIENTIA SINICA Informationis, 2018, 48(1): 24−46 (in Chinese)
(梁旭, 王卫群, 侯增广, 等. 康复机器人的人机交互控制方法[J]. 中国科学: 信息科学, 2018, 48(1): 24−46)

[7] Willett F R, Avansino D T, Hochberg L R, et al. High-performance brain-to-text communication via handwriting[J]. Nature, 2021, 593(7858): 249−254

[8] Gao Zhongke, Dang Weidong, Liu Mingxu, et al. Classification of EEG signals on VEP-based BCI systems with broad learning[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(11): 7143−7151

[9] Miranda R A, Casebeer W D, Hein A M, et al. DARPA-funded efforts in the development of novel brain-computer interface technologies[J]. Journal of Neuroscience Methods, 2015, 244: 52−67

[10] Binnendijk A, Marler T, Bartels E M. Brain-Computer Interfaces: US Military Applications and Implications, an Initial Assessment [M]. Santa Monica, CA: RAND Corporation, 2020

[11] Li Ren, Johansen J S, Ahmed H, et al. The perils and pitfalls of block design for EEG classification experiments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 316−333

[12] Dai Chenglong, Wu Jia, Pi Dechang, et al. Electroencephalogram signal clustering with convex cooperative games[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(12): 5755−5769

[13] Dai Chenglong, Pi Dechang, Cui Lin, et al. MTEEGC: A novel approach for multi-trial EEG clustering[J]. Applied Soft Computing, 2018, 71: 255−267

[14] Wahlberg P, Lantz G. Methods for robust clustering of Epileptic EEG spikes[J]. IEEE Transactions on Biomedical Engineering, 2000, 47(7): 857−868

[15] Dai Chenglong, Wu Jia, Pi Dechang, et al. Brain EEG time-series clustering using maximum-weight clique[J]. IEEE Transactions on Cybernetics, 2022, 52(1): 357−371

[16] Selim S Z, Ismail M A. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1984, 6(1): 81−87

[17] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding [C]// Proc of the 18th Annual ACM-SIAM Symp on Discrete Algorithms. Philadelphia, PA: SIAM, 2007: 1027−1035

[18] Nock R, Canyasse R, Boreli R, et al. k-variates++: More pluses in the k-means++ [C]// Proc of the 33rd Int Conf on Machine Learning. Cambridge, MA: MIT, 2016: 145−154

[19] Nie Feiping, Wang Chenglong, Li Xuelong. K-multiple-means: A multiple-means clustering method with specified k clusters [C]// Proc of the 25th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2019: 959−967

[20] Petitjean F, Ketterlin A, Gancarski P. A global averaging method for dynamic time warping, with applications to clustering[J]. Pattern Recognition, 2011, 44(3): 678−693

[21] Yang J, Leskovec J. Patterns of temporal variation in online media [C]// Proc of the 4th Int Conf on Web Search and Web Data Mining. New York: ACM, 2011: 177−186

[22] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [C]// Proc of the 2nd Int Conf on Knowledge Discovery and Data Mining. Palo Alto, CA: AAAI, 1996: 226−231

[23] Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: Ordering points to identify the clustering structure [C]// Proc of the 1999 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 1999: 49−60

[24] Ertőz L, Steinbach M, Kumar V, et al. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data [C]// Proc of the 3rd SIAM Int Conf on Data Mining. Philadelphia, PA: SIAM, 2003: 47−58

[25] Chen Yewang, Shen Lianlian, Zhong Caiming, et al. Survey on density peak clustering algorithm[J]. Journal of Computer Research and Development, 2020, 57(2): 378−394 (in Chinese)
(陈叶旺, 申莲莲, 钟才明, 等. 密度峰值聚类算法综述[J]. 计算机研究与发展, 2020, 57(2): 378−394)

[26] Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492−1496

[27] Ren Yazhou, Hu Xiaohui, Shi Ke, et al. Semi-supervised DenPeak clustering with pairwise constraints [C]// Proc of the 15th Pacific Rim Int Conf on Artificial Intelligence. Berlin: Springer, 2018: 837−850

[28] Li Zechao, Yang Yi, Liu Jing, et al. Unsupervised feature selection using nonnegative spectral analysis [C]// Proc of the 26th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2012: 1016−1032

[29] Qian Mingjie, Zhai Chengxiang. Robust unsupervised feature selection [C]// Proc of the 23rd Int Joint Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2013: 1621−1627

[30] Shi Lei, Du Liang, Shen Yidong. Robust spectral learning for unsupervised feature selection [C]// Proc of the 2014 IEEE Int Conf on Data Mining. Los Alamitos, CA: IEEE Computer Society, 2014: 977−982

[31] Li Zechao, Liu Jing, Yang Yi, et al. Clustering-guided sparse structural learning for unsupervised feature selection[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(9): 2138−2150

[32] Zakaria J, Mueen A, Keogh E. Clustering time series using unsupervised-shapelets [C]// Proc of the 12th IEEE Int Conf on Data Mining. Los Alamitos, CA: IEEE Computer Society, 2012: 785−794

[33] Zhang Qin, Wu Jia, Zhang Peng, et al. Salient subsequence learning for time series clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2193−2207

[34] Paparrizos J, Gravano L. k-shape: Efficient and accurate clustering of time series [C]// Proc of the 2015 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2015: 1855−1870

[35] Song Kun, Yao Xiwen, Nie Feiping, et al. Weighted bilateral K-means algorithm for fast co-clustering and fast spectral clustering[J]. Pattern Recognition, 2021, 109: 107560

[36] Huang Dong, Wang Changdong, Wu Jiansheng, et al. Ultra-scalable spectral clustering and ensemble clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 32(6): 1212−1226

[37] Rodriguez-Lujan I, Huerta R, Elkan C, et al. Quadratic programming feature selection[J]. Journal of Machine Learning Research, 2010, 11(2): 1491−1516

[38] Das K C. The Laplacian spectrum of a graph[J]. Computers and Mathematics with Applications, 2004, 48(5-6): 715−724

[39] von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395−416

[40] Frey B J, Dueck S. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972−976

[41] Zhou Dengyong, Bousquet O, Navin Lal T, et al. Learning with local and global consistency [C]// Proc of the 16th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2003: 321−328

[42] Boyd S, Vandenberghe L. Convex Optimization [M]. Cambridge, MA: Cambridge University Press, 2004

[43] Zhang Hui, Ho Tubao, Zhang Yang, et al. Unsupervised feature extraction for time series clustering using orthogonal wavelet transform[J]. Informatica, 2006, 30(3): 305−319

[44] Hubert L, Arabie P. Comparing partitions[J]. Journal of Classification, 1985, 2(1): 193−218

[45] van Rijsbergen C J. Information Retrieval [M]. London: Butterworth-Heinemann, 1979

[46] Fleiss J L. Measuring nominal scale agreement among many raters[J]. Psychological Bulletin, 1971, 76(5): 378−382



Dai Chenglong, born in 1988. PhD, lecturer. Member of CCF. His main research interests include brain-computer interface, EEG data mining, and machine learning.

代成龙, 1988 年生. 博士, 讲师. CCF 会员. 主要研究方向为脑-机接口、脑电信号数据挖掘、机器学习.



Li Guanghui, born in 1970. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include Internet of things, edge computing, fault tolerant computing, and nondestructive testing and evaluation.

李光辉, 1970 年生. 博士, 教授, 博士生导师. CCF 高级会员. 主要研究方向为物联网、边缘计算、容错计算、无损检测与分析.



Li Dong, born in 1998. Master candidate. His main research interests include brain-computer interface and EEG analysis.

李　栋, 1998 年生. 硕士研究生. 主要研究方向为脑-机接口、脑电信号分析.



Shen Jiahua, born in 1998. Master candidate. His main research interests include brain-computer interface and EEG analysis.

申佳华, 1998 年生. 硕士研究生. 主要研究方向为脑-机接口、脑电信号分析.



Pi Dechang, born in 1971. PhD, professor, PhD supervisor. Senior member of CCF. His main research interests include big data analysis, data mining, and system reliability.

皮德常, 1971 年生. 博士, 教授, 博士生导师. CCF 高级会员. 主要研究方向为大数据分析、数据挖掘、系统可靠性.